Predictive Data Analytics for E-Learning in Higher Education.

By

Adam James Wilden

BInfTech(Hons), BBus(Mgmt)

Thesis
Submitted to Flinders University
for the degree of

DOCTOR OF PHILOSOPHY

College of Science and Engineering, Flinders University, Adelaide

Date of Submission: February 2024

Abstract

This study delves into the significant potential of Learning Management System (LMS) usage to influence student performance across diverse range of academic disciplines, utilising predictive data analytics to deepen our understanding and enhance educational outcomes. As E-Learning becomes increasingly central to educational delivery, understanding the complex effects of varied LMS interactions on student success is critical. Traditional analytics methods, focusing on customising learning materials based on student interactions within LMSs, often overlook the differential impacts of E-Learning across disciplines where engagement and effectiveness can vary significantly.

Structured around two pivotal questions, the research seeks to uncover:

RQ1: "How does Learning Management System (LMS) use across disciplines impact student performance?" This is explored through inquiries into how LMS usage varies across disciplines and its relation to student performance metrics, identification of specific LMS features as significant academic performance predictors, and the role of predictive data analytics models in pinpointing at-risk students across colleges. The question of dimensionality reduction's necessity in capturing essential LMS use aspects and its impact on predictive model performance is also addressed.

RQ2: "Do colleges differ significantly in approach and consistency?" This question investigates the variability of student individual differences, such as learning styles and engagement patterns, and their academic repercussions. It explores the implications of student behaviour variations captured through LMS data, for instructional design and student support services, and examines distinctive pedagogical approaches as evidenced by LMS data in relation to student engagement and performance.

Employing a broad suite of machine learning algorithms, including tree-based classifiers, probabilistic models, ensemble methods, and hybrid models, the study offers an in-depth analysis of discipline-specific engagement patterns and material consumption within the LMS.

This methodological innovation facilitates a more detailed examination of how E-Learning implementations impact student performance across disciplines.

Findings reveal that distinct student profiles can be accurately identified for each college, enabling the prediction of course enrolment and the tailoring of machine learning approaches to predict performance within specific domains. Theoretically, the research advances a predictive data analytical model for E-Learning across disciplines, integrating insights from machine learning and E-Learning research to not only accurately predict student college affiliation through LMS data but also to leverage critical topic structure features for highlighting effective pedagogical strategies across disciplines.

Practically, this study equips educators with the insights needed to select and implement suitable E-Learning approaches, optimising teaching methods, material selection, and topic construction to meet disciplinary needs. Incorporating the instructional design methodology knows as ADDIE (Analysis, Design, Development, Implementation, and Evaluation) model, it provides a structured yet adaptable framework for designing and implementing E-Learning experiences, moving beyond generic solutions to offer tailored methodologies suited to distinct educational domains.

Moreover, by identifying effective algorithms for E-Learning tasks and elucidating effective pedagogical strategies in each discipline, the study enhances our understanding of E-Learning dynamics. It significantly advances predictive data analytics in E-Learning, employing innovative machine learning techniques and integrating established instructional design models to deliver actionable insights for educators and researchers, thus improving student performance across diverse educational settings.

Contents

List of figures and tables

List of figures

List of tables

List of abbreviations, and company/software names

General abbreviations

1S – One semester enrolment into topic.

2S – two (or more) semesters enrolment into topic.

ANN - Artificial Neural network.

AUD – Australian Dollars.

C3MS – Community, Content, & Collaboration Management Systems.

CAGR – Compound Annual Growth Rate.

CART – Classification and Regression Trees.

CMS – Course Management Systems.

CPU – Computer Processing Unit.

CSBA – Computer Supported Behavioural Analytics.

CSE – Computer Self-Efficacy.

CSLA – Computer Supported Learning Analytics.

CSPA – Computer Supported Predictive Analytics.

CSVA – Computer Supported Visualisation Analytics.

E1 – Exam period one.

E2 – Exam period two.

ELP – E-Learning Personalisation.

FLO – Flinders Learning Online.

IBM – International Business Machines.

IT – Information Technology.

KMS – Knowledge Management Systems

KNN – K-Nearest Neighbour.

LMS – Learning Management Systems.

LSTM – Long short-term Memory.

MB1 – Mid-Semester Break one.

MB2 – Mid-Semester Break two.

MIT – Massachusetts Institute of Technology.

MITx – Massachusetts Institute of Technology's MOOC program.

MLP – Multi-Layer Perceptron

MOOC – Massive Open Online Course.

NS1 – Non-Semester one.

NS2 – Non-Semester two.

NVMe – Non-Volatile Memory Express.

OS – Operating System.

PCIe – Peripheral Component Interconnect Express.

PLATO – Programmed Logic for Automated Teaching Operations.

PLE - Personal Learning Environments.

RAM – Random-access memory.

REPTree – Reduced Error Pruning Tree.

ROC – Receiver Operator Curve.

ROCCH – Receiver Operator Curve Convex Hull.

S1 – Semester one.

S2 – Semester two.

sCART – Simple Classification and Regression Trees.

SQL – Structured Query Language.

SSD – Solid State Drive.

SVM – Support Vector Machine.

T-SQL – Transactional Structured Query Language (Microsoft version).

USD – United States Dollars.

VLE – Virtual Learning Environments.

WEKA – Waikato Environment for Knowledge Analysis.

Flinders university grade abbreviations

CO – Continuing.

Cr – Credit (65-74).

DN – Distinction (75-84).

F – Fail.

F/A – Fail: Academic Assessment.

F/M – Fail: deferred Medical Assessment.

FAS – Fail: deferred Academic Assessment.

FCP – Failed Compulsory Part of assessment.

HD – High Distinction (85+).

I – Incomplete.

I/M – incomplete deferred medical assessment.

NGP – Non-Graded Pass.

NoGrade – No Grade recorded in results.

P – Pass (50-64).

PAS – Pass after supplementary assessment.

WF – Withdraw with Fail.

WN – Withdraw No Fail record.

Flinders university college abbreviations

BGL – College of Business, Government, and Law.

EPS – College of Education, Psychology, and Social Work.

HAS – College of Humanities, Arts, and Social Sciences.

MPH – College of Medicine, and Public Health.

NHS – College of Nursing, and Health Sciences.

S&E – College of Science and Engineering.

Glossary

Term	Definition
Data analytics	The application of computer systems to large datasets for decision making (Runkler 2020, p. 2).
E-Learning	The provision of learning using technology, accounting for differences in physical location and time. Providing additional levels of interactivity compared to in-person delivery.
Learning Technologies	An alternative term used for E-Learning, where the focus is of the educational technologies. Note, use of the term Learning Technologies becomes an issue when combined with other subjects containing the word Learning (Machine Learning, E-Learning), and can cause confusion when naming and searching for research material.
Machine Learning	A field of scientific research that resulted from attempts to replicate human biological processes to simulate learning (Rosenblatt 1957). The field includes a wide variety of methodologies, and algorithms to simulate learning, and is commonly used in Predictive Analytics.
Predictive Analytics	Forecasting future events through statistical techniques and machine learning models (Eckerson 2007, pp. 4-8).
Predictive Data Analytics	The combination of Data Analytics, and Predictive Analytics. Utilising past data to identify patterns that can be used for models to predict future performance (Kelleher, Mac Namee & D'arcy 2015, p. 1).

Statement of declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university.
2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed 23/02/24.

_____

Adam James Wilden

Publications

Publication from this thesis

Wilden, A.J., Shillabeer, A. and de Vries, D., 2017. **Predicting Success In E-Learning Courses**. In *e-Proceeding of the 5th Global Summit on Education: Trends and Challenges in Education, Kuala Lumpur, Malaysia 2017 (pp 274-281).*

Wilden, A.J., Shillabeer, A. and de Vries, D., **Predictors of successful learning outcomes in e-learning courses**. STARS Conference Stamford Grand Glenelg, Adelaide 2nd July 2017 (pp.74-75).

Mirzaei, S., Wilden, A. and Shillabeer, A., 2018, November. **A Preliminary Analysis of Australia's Readiness for E-Learning**. In *5th WorldCALL Conference 2018*.

Other publications

Wilden, A.J., Nasim, M., Williams, P., Legrand, T., Turnbull, B.P. and Williams, P.A., 2023, June. **On Benchmarking and Validation in Wargames**. In *22nd European Conference on Cyber Warfare and Security, ECCWS 2023* (pp. 533-543). Curran Associates Inc.

Acknowledgements

I would like to thank everyone who played a part in my journey to complete my PhD candidature, there are a great number of people who have helped, but I would like to highlight the most important people. First, I would like to thank my wonderful wife Mandy, for being there for me, and providing support throughout. The same would apply for both of my parents, who remained patient and supportive.

I would also like to thank my friends who were always there for me and helped provide a way to unplug from the intensity of my candidature and unwind. Andrew, Ryan, Adam, Eliza, and Witold for being there to bounce subjects off, and for going out and having some normality, even if only once or twice a week.

Next, I would like to thank my primary supervisor Giselle Rampersad, who believed in my ability to complete my thesis, and offered a tremendous amount of support. In addition, I would like to thank my secondary supervisors Greg Falzon, and David Powers, who both provided a large amount of support into the area of machine learning, which initially I was very new to; both providing years of experience publishing and researching into machine learning.

I would also like to thank my previous supervisors, Denise deVries, Anna Shillabeer and John Roddick. Anna and Denise helped me to begin my PhD journey, as well as providing insight into what direction I needed to take in the crucial first few years. John was also very supportive, coming on to the supervision team at a critical point, where there was massive shake-up both at the university, as well as the rest of the world, during the pandemic. John was very supportive (along with David), with assisting through that period, and letting me find a balance of work and research. I'm glad I was able to have Denise, Anna and John be part of my candidature.

Finally, I would also like to thank Flinders University as a whole, for providing me the opportunity to take on this goal to achieve, and providing the tools, and financial support to complete my goal. I would also like to acknowledge the Australian Government Research Training Program Scholarship I received, as well as the Flinders University Research Scholarship (FURS) I received during my candidature; without these I would not have likely been able to complete my candidature.

1. Introduction

1.1. Chapter overview

This chapter introduces the pivotal role of E-Learning tools in enriching educational outcomes through the analysis of student and Learning Management System (LMS) data. It emphasises the necessity of tailoring educational content and LMS features to meet the specific needs of learners, based on comprehensive data analysis, aligning with the broader goal of personalising learning experiences.

The chapter explores the influence of LMS usage across different disciplines on student performance and examines the unique pedagogical approaches of various educational institutions. It utilises machine learning and data mining techniques, especially decision-tree based algorithms, to discern the relationship between E-Learning methodologies and student performance across diverse disciplines.

Addressing the notable scarcity of discipline-specific E-Learning strategy research in existing literature, the chapter proposes a novel research model to refine educational practices by integrating predictive analytics, akin to adapting teaching methods to accommodate individual learning styles and preferences.

Furthermore, the chapter delves into the historical context of E-Learning, charting its development and the significant impact of recent global events like the COVID-19 pandemic on the educational sector. It acknowledges the accelerated transition to E-Learning, highlighting the resultant challenges and opportunities for educational institutions, educators, and students.

The chapter sets the foundation for a thorough exploration into the potential of data-driven personalisation in E-Learning, reflecting a detailed understanding of individual learning preferences and material types as crucial elements in enhancing the quality of educational delivery and student performance.

1.2. Research topic

This section will explore the key research questions pertaining to this research as well and the key educational objectives in providing implications for teaching practices.

There are two important questions to consider when researching E-Learning and predictive data analytics; what is the benefit to students and educators? And to what areas can they be applied to be beneficial? The primary focus of this research is to explore the benefits of predictive analytics, and data mining to benefit students. Additionally, it uncovers how educators and institutions can better utilise these tools to improve their own processes.

When considering the benefits of predictive data analytics, or data mining in the E-Learning context, or educational context, it is important to identify what sort of research has been performed previously on the subject, and what areas of teaching or learning can it be used to help improve. A systematic literature review of educational data mining by Rodrigues, Isotani & Zárate (2018), identified four major themes of research: the evaluation of E-Learning in traditional classroom, the evaluation of pedagogical actions, the evaluation of administrative management, and the evaluation of Multimedia resources. The review itself consisted of 72 articles or conference papers published from 1994 to 2016, each directly relating to researching the improvement of teaching and learning processes. Of the four themes identified, the first two directly relate to the interactions between the student, the educator, and the learning environment, whereas, the last two themes relate to the administrative processes of the institution, and the evaluation of multimedia resources.

1.2.1. Research questions

This research focuses on analysing the impact of pedagogical E-Learning approaches on student performance. This is well aligned to the second major theme identified by Rodrigues, Isotani & Zárate (2018) pertaining to the evaluation of pedagogical actions (analysis of educator's actions, and analysis of student's actions).

These themes are echoed in other research into the evaluation of teaching and learning actions, such as the evaluation of overall E-Learning pedagogy (Beetham & Sharpe 2007, p. 14; Mallillin et al. 2020; Patwari, Dubey & Jagdale 2023; Srinivasa, Kurni & Saritha 2022, p. 300), the evaluation of student behaviours (Beatty, Merchant & Albert 2017; Bertholdo et al. 2018; Qiu, Zhang, et al. 2022; Wang 2017), and student individual differences such as student learning styles (Assiry & Muniasamy 2022; El-Sabagh 2021; Essa, Celik & Human-Hendricks 2023; Huang, T-C, Chen & Hsu 2019; Kika et al. 2019; Rasheed & Wahid 2021; Sheeba & Krishnan 2019; Vaidya & Joshi 2018; Wijaya, Setiawan & Shapiai 2023), student cognitive load (Altinpulluk et al. 2019; Huang, CL et al. 2019; Kruger & Doherty 2016; Lange 2023), and student self-efficacy (Baherimoghadam et al. 2021; Bai 2017; Ithriah, Ridwandono & Suryanto 2020; Latip et al. 2022).

While prior research focused on the general impact of the Learning Management System (LMS) on student outcomes, it did not have a discipline specific view and therefore further research is critical in creating targeted and more effective recommendations for educators within specific disciplines.

Therefore, after this consideration, the following research questions were devised, to best identify the most important factors to consider when researching E-Learning and predictive data analytics:

RQ1. How does Learning Management System (LMS) use across discipline impact student performance?

RQ1.1. How does LMS usage differ across disciplines, and how are these differences associated with student performance metrics?

RQ1.2. Which specific features of LMS usage are significant predictors of student academic performance?

RQ1.3. How can predictive analytics models, incorporating LMS usage data, enhance the identification of at-risk students across different colleges?

RQ1.4. Is dimensionality reduction necessary to accurately capture the essential aspects of LMS use, and what impact does this reduction have on the performance of predictive models?

RQ2. Do colleges differ significantly in approach and consistency?

RQ2.1. In what ways do colleges differ in terms of student engagement patterns, and how are these differences reflected in academic outcomes?

RQ2.2. How do student behaviours, as captured through LMS data, vary across colleges, and what implications do these variations have for instructional design and student support services?

RQ2.3. What are the distinctive pedagogical approaches adopted by different colleges as evident from the LMS data, and how do these approaches correlate with student engagement and performance?

Answering these questions will allow educators to potentially use discipline-specific information from data mining to improve student performance, as well as identify common trends, which will enable colleges in learning lessons from other colleges, with regards to student usage of LMS materials.

Further research relating to RQ2, especially RQ2.1, and RQ2.2 has been identified, which involves factors not captured by the data from this research, such as demographic data, enrolment type, and specifically captured learning and teaching styles.

1.3. Scope of the study

In terms of methodology, the overall scope of this research focuses on quantitative research using machine learning and data mining techniques. In particular, it uses decision-tree based algorithms, which will allow for easier understanding of the output (Perner 2011) and is not at a significant disadvantage (Aytekin 2022; Grinsztajn, Oyallon & Varoquaux 2022) to the more popular neural network approaches. These features make decision-trees a suitable option to inform educational practice. The quantitative approach will also allow for a larger dataset of student user data, that can be processed through machine learning techniques.

This research will primarily concentrate on well-established machine learning and data mining techniques, to identify rules, and best-practice from a wealth of data from decision-tree based algorithms. This approach is useful in predicting the impact on different E-Learning approaches on performance among students in different disciplines. The study is novel in the application of ML to E-Learning to compare disciplines to uncover trends, similarities, and differences in the impact of E-Learning pedagogical approaches on student performance.

In terms of past theoretical foundations, this research will not focus on areas such as Technology Acceptance Models (TAM), Artificial Intelligence (AI), or qualitative research. It focuses on predictive analytics in E-Learning and aims to make a valuable contribution on advancing understanding across disciplines. This selected approach is useful as existing research has primarily focused on self-reported, educator-focused data (Becher 2001, p. 90; Biglan 1973a, p. 196; 1973b, p. 205; Gaff & Wilson 1971, p. 187; Kolb 1981, p. 237), and not the analysis of large student usage data. Therefore, it will add robustness and advance the literature through the integration of predictive data analytics.

Finally, this research will focus on higher education use of E-Learning, as opposed to corporate use, or primary/secondary education. The literature review will still account for the existence of these different segments; however, deeper examination will only be made on higher education uses. This focus is useful because of the large amount of research available in the higher educational setting (Baek & Doleck 2022; Djeki et al. 2022; Gao et al. 2022; Irwanto et al. 2023; Jia, K et al. 2022; Khan, FM & Gupta 2022; Prioteasa et al. 2023; Vaicondam et al. 2022; Wijaya, Setiawan & Shapiai 2023).

1.4. Significance and contribution of the research

The purpose of the research is to develop a predictive data analytics model for E-Learning across disciplines. This research is valuable due to the current lack of research involving prediction of performance, and discipline-based characteristics, as opposed to a general focus on individual performance (Anitha et al. 2022; Fahd, Miah & Ahmed 2021; Gajwani & Chakraborty 2020; Hamadneh et al. 2022; Qiu, Zhang, et al. 2022; Qiu, Zhu, et al. 2022; Sathe & Adamuthe 2021) and individual differences (Bandura 1977; Barbara & Donna 2005; Fariani, Junus & Santoso 2022; Lange 2023; Mayer, Richard E & Moreno 2003; Mikić et al. 2022; Morris, Finnegan & Wu 2005; Picciano 2002).

At present there is a focus on E-Learning personalisation based on learner characteristics and learning pedagogies (Mikić et al. 2022), rather than differentiating student characteristics to a specific discipline.

1.4.1. Gap in the literature

A Scopus literature search was conducted to identify related research. This search was not limited to a specific date range, or research discipline. The Scopus advanced search option was used to identify any E-Learning research that involves research into performance and/or prediction, regarding the discipline-based approach mentioned in the research question. The search string included variations of discipline (such as college, faculty, or college), as well as including performance or prediction/predict.

The results of the Scopus search included 105 documents published between 2002 and 2023, and from a manual review of the documents, revealing a dearth of research on differences across disciplines in E-Learning.

This lack of research is confirmed by multiple recent bibliometric surveys (Baek & Doleck 2022; Djeki et al. 2022; Fauzi 2022; Gao et al. 2022; Irwanto et al. 2023; Jia, K et al. 2022; Khan, FM & Gupta 2022; Prioteasa et al. 2023; Vaicondam et al. 2022; Wijaya, Setiawan & Shapiai 2023), of documents published within a series of date ranges, with the earliest by Gao et al. (2022) containing documents from 1998 to 2020, and Irwanto et al. (2023) containing the most recent documents from 2012 to 2022. Specific research areas of E-Learning were also considered, with the survey by Khan & Gupta (2022) including documents relating to M-Learning, and Irwanto et al. (2023) including documents relating to MOOCs. The surveys identified were limited in their exploration of discipline differences in their thematic analysis or identified keywords.

1.4.2. High-level view of research model

At present, there is a large amount of research into customisation of both teaching practices and accommodating student individual differences, to better improve learning outcomes. This research is positioned above those two factors and would allow for a better level of control in managing teaching practices, and student individual differences, at a domain (or discipline) level.

This type of research is important, not only due to the lack of documented research, but also due to the way that it affects both teaching practices, and student personalisation. As shown in Figure 1, the teaching practices influence the types of student personalisation that would be chosen. For example, a more constructivist approach may involve allowing a student to have enough time to adequately digest the information presented, and the individual difference of the student such as the preferred learning style would then affect what type of information is presented. Having a method of limiting the number of possible combinations of teaching practices, and known individual differences, that relate to a specific domain, would save educators considerable time.

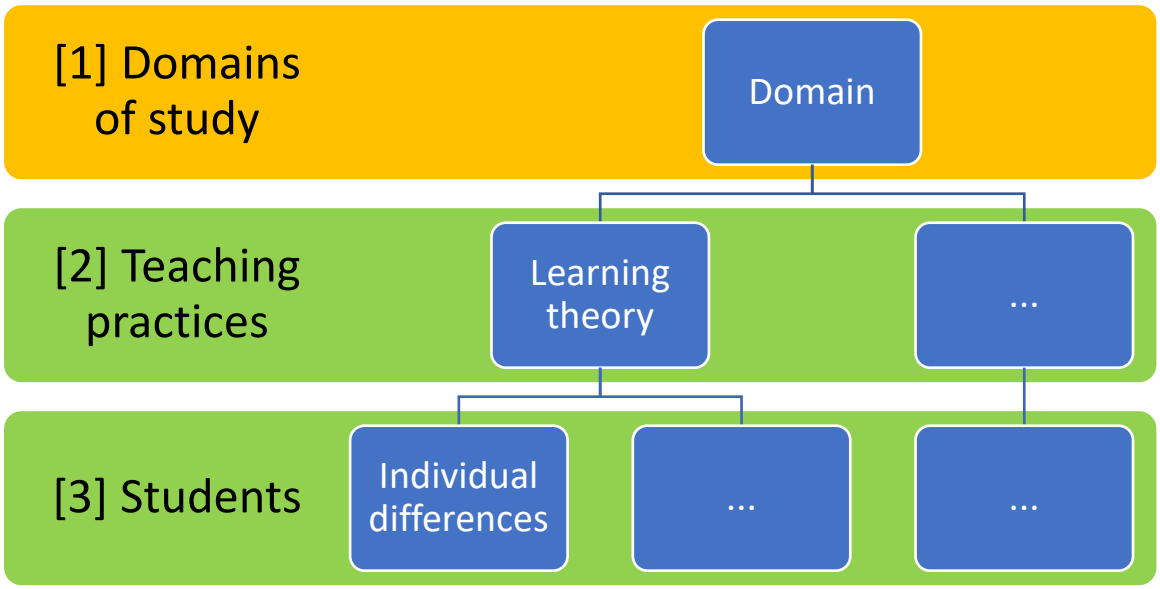

Figure 1 - Levels of research into student E-Learning performance

1.5. E-Learning background and importance

The E-Learning research field was in a very different place pre-COVID, with many of the current industry leaders either just emerging or had different purposes (Hill 2022). Additionally, the outlook of the industry, while positive, was not as much of a juggernaut that it is now. The increased interest into E-Learning requires the participation of world governments, to implement policies on adapting or enforce the usage of E-Learning.

The direction and overall context of this thesis was also affected due to the COVID-19 pandemic, with great delays imposed on the overall candidature timeline, and further emphasis on the incorporation of E-Learning into higher education (as well as all other facets of education. An heavier initial focus was placed on MOOCs, as well as a pilot study by the author (Wilden, Shillabeer & deVries 2017). However, the research was directed away from the open learning arena, towards the integration of E-Learning into traditional educational infrastructure, due to the rapid growth of research into the area post-COVID-19.

1.5.1. MOOCs

One of the first major differences with the E-Learning environment pre and post COVID-19, was that of the larger focus on MOOCs (Massively Online Open Courses), and their promise and potential to change the entire E-Learning landscape.

Back in 2016 when research for this thesis began, the E-Learning environment was in a completely different situation to what it is at present; MOOCs were an emerging use of technology, that at the time was being suggested to improve access to, and performance of online education (Yousef et al. 2014). Taneja and Goel (2014) provide a simple definition of what a MOOC is; namely an online course attended by a large number of students, for a defined duration, and who are generally expected to dedicate around 2 to 6 hours per week of work towards. However this definition does not delve into the economics of a MOOC, which can range from completely free, pay for certificate, or sub-licencing out to institutions (Jia, Y et al. 2017).

Research into E-Learning commonly involved some focus on MOOCs or the potential use of MOOC like environments. Valverde-Berrocoso et al. (2020) found that the most common keywords for research studies into E-Learning involved MOOC, higher education, teaching-learning strategies, and interactive learning environments.

Sustainability of MOOCs and other such environments was also a concern, given the large number of users, and the cost of maintaining such large systems. This large user base can be shown with such MOOC providers as Coursera, EdX and Udacity (Taneja & Goel 2014), providing access to 10.5M, 3M, and 1.5M respectively (Shah 2014). Suggestions that fully open and free MOOCs would not be sustainable into the future were warranted given the increased service offering and complexity of such systems (Porter 2015). Some of the more popular MOOCs included: EdX, FutureLearn, Coursera, Udemy, Udacity, and Iversity (Spyropoulou, Pierrakeas & Kameas 2014).

MOOCs were also a focus for global research due to its arguably hyped nature of allowing people of lower income (or developing countries), to access quality education, without needing to travel large distances, and to free up the faculty of traditional learning environments (Das, Das & Das 2015).

There were also concerns that the quality of such a learning environment may be not up to standard with regards to pedagogical support, however the main providers appeared to be quite sufficient in these regards (Lebron & Shahriar 2015). There also were suggestions that MOOCs as a separate category would entirely disappear, and will be developed in conjunction with other offerings, becoming a norm amongst institutions (Sandeen 2015).

1.5.2. Growth and financial projections

Early predictions of E-Learning success and growth were significant but not out of the ordinary, additionally they did not anticipate the global pandemic and the subsequent need for global focus on distance learning. Comparing financial projections from 2014 to more recent projections show a slight difference with what took place, under predicting the overall growth of the industry. It was projected that by 2016 the E-Learning market would reach 51.5B USD

(up from 35.6B USD in 2011), with a five year compound annual growth (CAGR) rate of 7.6% (Docebo 2014). However, more recent projections estimate a CAGR of 13.16% by 2027, (Setiawan et al. 2023). Therefore, there is a much larger and economically profitable industry than was initially predicted pre COVID-19.

This surge in growth can be attributed to several key factors. Firstly, the pandemic has drastically altered the landscape of education and professional development, necessitating rapid adoption of E-Learning solutions worldwide. As discussed by Rogers (2023), companies with large numbers of students grew from 22 percent in 2019 to 29 percent in 2020, reflecting the shift to remote learning during the COVID-19 pandemic in particular.

1.5.3. Integration into existing infrastructure

While there was initial scepticism regarding the introduction of E-Learning and its potential to replace in-person teaching, the focus has gradually shifted towards enhancing and integrating online content into existing courses to improve performance. This shift, however, brought to light significant challenges in technological adoption and system compatibility. Hodges et al. (2020) differentiate between emergency remote teaching and planned online learning, highlighting the necessity for educational institutions to not only adapt quickly in response to immediate needs, such as those presented by the COVID-19 pandemic, but also to thoughtfully integrate E-Learning technologies into their existing infrastructures for long-term sustainability and effectiveness.

One of the more important challenges faced for effective integration, has been ensuring that the existing infrastructures can support new E-Learning platforms and tools. This includes not only upgrading hardware and ensuring reliable internet access but also integrating E-Learning platforms with existing LMS to create a seamless learning experience for learners. The distinction made by Hodges et al. (2020) highlighting the importance of strategic planning in the adoption of E-Learning technologies. Njenga & Fourie (2010) suggested that there needed to be more focus on what innovations E-Learning could bring, and that ‘technopositivism’ (a term that they used of the belief that technology in and of itself is a good thing) within the E-Learning industry could lead to the adoption of inferior innovations in education.

They also questioned whether the call for more human interaction in teaching would be a better use of technology that would bring more learners and teachers together and treat the E-Learning technology as just a medium for teaching, and not the end goal.

The post-pandemic landscape has further emphasised the lack of focus on the quality of E-Learning, as noted by Abdul Razzak (2022). It has become evident that to maximise the benefits of E-Learning, materials and pedagogies must be specifically designed for online delivery, rather than retrofitting in-person materials for LMS use. This approach ensures that E-Learning is not just an afterthought but a well-integrated component of the educational offering, providing value beyond what can be achieved in traditional classroom settings. Considering that pre-pandemic there was a stronger focus on what will improve student performance, and an overall focus on quality through E-Learning, it is therefore important that, post-pandemic the focus should remain as such, and not simply move towards a focus on quantity and whatever can be integrated at short notice. As mentioned by Njenga & Fourie (2010) who suggest that to best exploit the technology in E-Learning; learning materials needed to be geared towards delivery through E-Learning platforms, rather than being an afterthought; not providing any extra benefit than the in-person materials they were based on.

Moreover, the integration of E-Learning, M-learning (Mobile learning), and MOOCs presents an opportunity to not only promote the benefits of these technologies to students but also to address their overall acceptance within the education domain. Research suggests that the success of E-Learning implementations depend on a variety of factors, including system and educational quality, content, service quality, user satisfaction, and the positive intention of students (Hassanzadeh, Kanaani & Elahi 2012; Sun et al. 2008). This is in addition to student and instructor attitude towards E-Learning overall, with perceived usefulness of the system, and perceived ease of use also being a good predictor.

Finally, there was a stronger focus on making the content more learner focused and using the technology to provide features to the students that they would find valuable. Zhang et al. (2004), questioned if E-Learning would replace in-person teaching, suggesting that E-Learning would be an indispensable tool for academics, but would not be a complete replacement.

Rather, it is more of a complement than a replacement, allowing students to re-experience lectures at will, and not necessarily in the order presented in a live lecture. While this is true, it may not have been imagined that there would be a pressing need to make it the essential mode of delivery due to the pandemic. Barbara & Donna's (2005) thoughts on a learner centred approach involved utilising the technology, much like Zhang et al. (2004), but to use it to expand and transform notions of education, to provide life-long learning to students.

By addressing these technological and pedagogical challenges, educational institutions can better integrate E-Learning implementations into their existing infrastructures, enhancing the quality of education and ensuring that technology serves as a medium for bringing learners and teachers together, rather than as an end goal.

1.5.4. Impact of COVID-19 in E-Learning

One of the most unexpected, but important events to shape E-Learning research and E-Learning as an industry, was that of the COVID-19 pandemic; the effect of the pandemic itself, as well as the world-wide response to the pandemic. In this section, we will show that not only did the pandemic force an accelerated adoption of E-Learning, but it also encouraged a large surge of research into E-Learning.

While officially beginning around January 25th 2020 in Australia, with the first recorded case, the first major milestone occurred in mid to late March, when all Australian states were put into partial lockdown (Stobart & Duckett 2022). This lockdown forced many non-essential businesses to close and forcing schools to start to implement online activities. Australians were forced to stay home 'where possible' and utilise a variety of online services to maintain everyday life activities, such as shopping, working, and education.

During 2020, there was a massive surge of research into E-Learning, Mseleku (2020) performed a comprehensive literature search on 16 databases for E-Learning and COVID-19 related publications in the year 2020. The search terms used by Mseleku (2020) were Covid-19, coronavirus, online learning, E-Learning, Eteaching and higher education. The number of initial results found were 960 papers, 920 of those being peer-reviewed papers.

The research areas that Mseleku (2020) identified included; 'Higher education institutions response to COVID-19 and lockdown', 'Online versus offline', 'Challenges for online teaching and learning', 'Academics' difficulties to adjust', 'Students' difficulties to adjust', 'Connectivity, network and internet issues', 'Unconducive physical space and environment', 'Mental health related issues', 'Lack of basic needs', 'Lack of teaching and learning resources', 'COVID-19 and academic outcomes', and 'COVID-19 induced opportunities'.

One of the biggest criticisms of the response to the pandemic by education institutions has been the type of approach to having students go online for topics. The large number of challenges reported by institutions after instituting a remote teaching approach suggest that a blended approach may benefit students better and provided a smoother transition to fully-online for students and teaching staff (Aboagye, Yawson & Appiah 2021). While it is difficult to know for sure what would have been best, we now have a large amount of data, and research resulting from a wide variety of implementations across the globe. What is suggested however, is that the pandemic has caused a great deal of damage to the educations of students across the world (Bryant et al. 2022).

1.5.5. E-Learning in 2024+

After the events of the pandemic, there was a significant change in the E-Learning environment. Industry leaders were no longer in their esteemed positions, and while most of the early platforms remained, some have disappeared entirely. Currently there are many E-Learning providers in the industry, with a wide variety of different LMS (Learning Management System) offerings. Many of those different LMS, are more commonly deployed (or only deployed) within a higher educational environment. According to market analysis reports by Phil Hill & Associates (2022), the most common LMS platforms for higher education use are as follows; Blackboard, D2L Brightspace, Canvas, Moodle, and Sakai. Currently Canvas is mentioned as being the most common in the North American market, with 34% of US and Canadian higher education institutions adopting it as their learning management system (LMS), while Moodle, Blackboard, and D2L were 21%, 20%, and 14% respectively (Hill 2022).

Research into E-Learning also exploded during the two years of the pandemic, with significant increases in academic publications. A bibliometric analysis of research into E-Learning by Fauzi (2022), found 1,496 related publications related to E-Learning (excluding review articles, book chapters, and conference proceedings). Another bibliometric analysis by Prahani et al. (2022) showed that during the period of 1991 to 2021, research into LMSs were contributed by 116 countries with the top 10 countries being the United States of America (391 documents), China (191 documents) and Malaysia (146 documents).

Funding sources for research into E-Learning has also expanded significantly across the globe. Prahani et al. (2022) showed the top 5 funding sources for LMS research was the National Natural Science Foundation of China (28 documents), the National Science Foundation (26 documents), the European Commission (24 documents), the National Research Foundation of Korea (21 documents), and the Japan Society for the Promotion of Science (16 documents).

The interest into E-Learning has only increased since the COVID-19 pandemic began and is likely to continue to increase. Fauzi (2022) suggests that since the beginning of the pandemic researcher interest in E-Learning has only increased and would be predicted to continue to rise well after the pandemic, this sentiment is echoed by Gao et al. (2022) who also found in their bibliometric analysis (E-Learning of publications from 1998 to 2020), that not only has scientific documents on E-Learning increased exponentially, but the diversity of researcher fields has increased of those contributing.

In summary, the landscape of E-Learning in 2024 and going forward, reflects a period of significant transformation and growth, caused by the global COVID-19 pandemic. The reshuffling of industry leaders and the evolution of LMS highlights the continually changing nature of the E-Learning environment. With Canvas emerging as a frontrunner in North America, the diversity in LMS platforms highlights a customised approach to digital education, catering to the specific needs of higher education institutions. The surge in academic research on E-Learning, as evidenced by the work of Fauzi (2022) and Prahani et al. (2022), illustrates an expanded global interest and investment in understanding and optimisation of online learning environments.

The commitment from leading funding bodies across the world, as mentioned by Prahani et al. (2022), further supports the burgeoning field of E-Learning research, indicating a robust future trajectory for digital education. As the field continues to attract a wide variety of academic disciplines and global contributions, the post-pandemic era promises an era of innovation, inclusivity, and enhanced learning experiences. The enduring impact of the pandemic on E-Learning has not only expanded the horizon of digital education but has also set the stage for continued exploration, improvement, and adoption of E-Learning platforms worldwide.

1.5.6. Accelerated transitioning to E-Learning

With greater focus on quick adoption and student acceptance due to the pandemic, there has been less focus on providing quality education and improving technology, and instead a larger focus on dealing with the associated negatives, and the possibility to have a full transfer from in-person to online where possible. Post-pandemic, E-Learning continues to be highly used following exposure to benefits and shortfalls due to rapid adoption through the pandemic.

A review by Turnbull, Chugh & Luck (2021) on the challenges for higher education institutions in transitioning to E-Learning found that some of the most common issues were the integration of existing systems with regards to synchronous and asynchronous tools, to form a single seamless online delivery, barriers to technology access, dealing with any inadequacies in online competencies for both learners and teaching staff, and finally dealing with academic dishonesty and privacy and confidentiality.

As previously mentioned, research into E-Learning had been primarily about prediction of student performance, improving student outcomes, and utilising the technology available. With the increase in use of E-Learning world-wide (enforced or otherwise), it has become a more important field of research. In addition, issues that had not been as prolific in research prior, such as academic dishonesty, have become more and more present. Since the COVID-19 pandemic, additional issues have been identified, such as the preparedness of students for fully online education, and social isolation. Aboagye, Yawson & Appiah (2021) documented the challenges for students in tertiary institutions, with the most important challenge being that of accessibility, and not being prepared for a complete online experience.

With the transition towards E-Learning, also comes the option to integrate other technological aspects into the learning environment. For example, the integration of more recent technologies such as Artificial Intelligence (AI) and cloud computing has been essential in enhancing the E-Learning experience, facilitating personalised learning and scalable infrastructure to support increased demand (Zhou et al. 2020). Additionally, there has been an evolution of pedagogical strategies to include blended and flipped classroom models. Strategies such as these, have been suggested by Abeysekera & Dawson (2015) to show promise in maintaining engagement and effectiveness in online settings.

In terms of global access to the benefits of E-Learning, there is also the need to examine, what is referred to by Ragnedda & Muschert (2013, pp. 1-4), as the 'Digital Divide', or how inequality of access to technology may affect social, economic, and political participation of individuals. Therefore, addressing this digital divide with regards to E-Learning in higher education is of critical importance, as disparities in availability and access to computing technology and internet connectivity have been suggested to affect students' ability and intention to participate in online education (Maheshwari 2021). Since the advent of COVID-19, the focus on the digital divide has become even more important. Research by Aissaoui (2022) suggest that since the pandemic, the digital divide has been worsened, and additionally, it has been shown that information on the digital divide is insufficient, not taking into consideration various metrics of digital inequalities mentioned in their research.

The lack of preparedness as well as social and instructor issues was a large factor with students' intention and willingness to study fully online. Feeling that there is a wider community of learners and teachers available to assist, as well as the availability for informal spaces for discussion, was found to increase student engagement with online topics by Kahu, Thomas & Heinrich (2022), in their research into LMS support through external tools such as Discord.

In conclusion, the accelerated transition to E-Learning during the global COVID-19 pandemic has identified both opportunities and challenges within higher education. This rapid shift has not only necessitated the adoption of online learning platforms but has also highlighted the need to address quality education, technology improvement, and the digital divide. As institutions attempt to navigate the complexities of integrating synchronous and asynchronous tools for seamless online delivery, they must also attempt to break through barriers of technology access, and online competencies. The more recent emergence of AI, cloud computing, and various innovative pedagogical strategies offer possible ways to enhance the overall E-Learning experience for learners, and to improve levels of engagement. However, the threat of the digital divide in a post-COVID-19 world highlights the need to ensure access to these advancements. Moving forward, it will be important to refine E-Learning methodologies and bridge the digital divide to foster an inclusive, effective, and resilient educational landscape. The lessons learned during this period of rapid change will shape the future of education, and will likely emphasise the significance of community support, and the strategic use of technology (such as E-Learning) to enhance learning outcomes and student engagement in an increasingly digital post-pandemic world.

1.6. Structure of the Thesis

The thesis is organised into eight chapters:

Chapter 1 Introduction: presents the pivotal role of E-Learning tools in enhancing education, defining the research scope, and delving into the significance and contributions of the study. Offering a background on E-Learning's evolution and its transformation post-COVID-19, concluding with a concise summary that encapsulates the core aims and expectations of the research.

Chapter 2 Literature review: Traces E-Learning's historical roots and examining the integration of machine learning and data mining. Each section within this chapter wraps up with a summary, synthesising the key findings and insights.

Chapter 3 Methods: Outlines the project methodology, including data preparation, exploratory data analysis, dimensionality reduction analysis, and machine learning analysis. It describes the systematic approach to predicting student grades, student college membership, and tailored approach to predicting student grade via college affiliation. The chapter also discusses expected integrity of the results, and the methodologies to help ensure integrity.

Chapter 4 Results: Provides detailed results of all tests explored within this research. Detailed results are accompanied by helpful visualisations, as well as analysis of the results.

Chapter 5 Discussion: begins with an overview of the results from previous chapters, as well as discussing the importance of findings. The chapter then discussed the commonalities of results found between colleges, subsequently investigating the specific findings on a college-by-college basis. Ending with a summary that underscores the primary insights and their implications for the broader academic context.

Chapter 6 Conclusions: The chapter addresses the central research questions, encapsulating the study's contributions such as discussing the broader educational implications for teaching practice, LMS design, and instructional design practices.

It discusses the study’s key contributions to the field of E-Learning, as well as Education, while highlighting the study's limitations and points towards potential future research avenues, summarising the chapter's key takeaways.

Chapter 7 Appendices: This chapter contains supplementary materials not suitable for the main body due to size, format, or redundancy of purpose. This includes items such as tables, figures, and MATLAB/R/Python scripts used in the production of visualisations and for exploratory statistical tests.

Chapter 8 References: Finally, this chapter offers a comprehensive compilation of references.

1.7. Chapter summary

This chapter serves as a foundational overview of the thesis and the project. Discussing the potential of E-Learning tools to enhance educational outcomes through the analysis of student and Learning Management System (LMS) data. Identifying the necessity of tailoring educational content and LMS features to individual learner needs, based on comprehensive data analysis. This approach aligns with the broader educational objective of personalising learning experiences, which will be discussed further in the individual differences in learning styles discussed in subsequent chapters.

The research questions proposed aim at investigating the influence of LMS usage across disciplines on student performance and examining the distinct pedagogical approaches of various colleges. The choice to utilise machine learning and data mining techniques (particularly decision-tree based algorithms) assists in the pursuit of identifying the relationship between E-Learning approaches and student performance across diverse disciplines through a unique data-driven approach.

Highlighting the research's significance, the chapter identifies a notable gap in existing literature, particularly the scarcity of studies focusing on discipline-specific E-Learning strategies. Proposing a novel research model, aiming to refine educational practices by integrating predictive analytics, much like the adaptation of teaching methodologies to suit individual learning styles and preferences.

Finally, the historical context of E-Learning was explored, charting its evolution and the impact of recent global events like the COVID-19 pandemic on the educational landscape. Acknowledging the accelerated transition to E-Learning, underscoring the resultant challenges and opportunities for educational institutions, educators, and students alike.

The chapter lays the groundwork for a comprehensive exploration into the potential of data-driven personalisation in E-Learning, mirroring the detailed understanding of individual learning preferences and material types as pivotal factors in enhancing educational delivery and student performance.

2. Literature review

2.1. Chapter overview

This chapter provides an in-depth review of E-Learning literature, tracing its historical evolution and examining the current state and future prospects of digital learning platforms. It delves into the pedagogical underpinnings, technological advancements, and the individual learner differences that shape the E-Learning landscape.

The chapter categorises E-Learning into systems like Course Management Systems (CMS), LMS, and Knowledge Management Systems (KMS), each serving distinct educational functions. The progression in technology has led to sophisticated platforms with modular design architecture, facilitating extensive functionalities like reporting, tracking, and analysis capabilities.

An in-depth discussion on various pedagogical frameworks and instructional design models, is provided. The chapter emphasises the need to align teaching methods with the diverse learning styles and preferences of students, advocating for a shift towards more personalised and adaptable E-Learning environments. Due to the scope of this research, and the overall focus, only the most identified, and used pedagogical frameworks, and design models are outlined, as they provide enough differentiation and use for most educational settings. The inclusion of additional frameworks and/or design models would be more beneficial in a solely education research focus.

The chapter stresses the significance of recognising and accommodating individual learner differences, such as learning styles, chronotypes, self-efficacy, and cognitive load, for the effective delivery of E-Learning. This section underscores the importance of personalising E-Learning experiences to enhance student engagement and learning outcomes.

The chapter reviews various machine learning algorithms and their efficacy in predicting E-Learning success. It highlights the need to choose algorithms that balance accuracy with usability and interpretability for educators and data scientists, leading to the preference for decision tree methodologies in this research. Again, like the inclusion of commonly used pedagogical frameworks, and design models, the algorithms chosen were due to both necessity (for those that are user-interpretable, and usable for the research), as well as those most common to be used in E-Learning research.

In conclusion, the chapter provides a detailed analysis of the multifaceted nature of E-Learning, encapsulating its historical roots, current trends, pedagogical approaches, technological advancements, and the need to consider individual learner differences. It pinpoints the necessity for ongoing research, especially in leveraging predictive analytics to enhance E-Learning platforms and methodologies, thereby enriching the educational experience and outcomes for learners.

2.2. History of E-Learning

To understand the present and future of E-Learning it is important to understand its origins, the initial intentions of its early innovations, and decisions made. The idea of E-Learning (learning over distances through technology) has existed since the early 1980s (Moore, Dickson-Deane & Galyen 2011), with correspondence courses becoming a common delivery medium towards the turn of the century (Imel 1998, p. 3; Sherry 1996; Valentine 2002).

Well before any modern incarnations of online based learning (or even computer based), the original genesis of E-Learning has been suggested to have its roots in the industrial revolution during the mid-19th century in the form of correspondence courses (Peters 1973 in Keegan, 1995:5). While not requiring any form of computer technology, a correspondence course still provided students with a way of engaging in learning over distance through means other than direct person to person interaction.

This remained the standard for geographic and time separated learners and teachers until the middle of the 20th century, where great advancements in technology had started to become more available. New technologies being introduced to classrooms gradually over the course of the 20th century such as film in the 1910s, Radio in 1920s, and television in the 1950s (Cuban 1986, pp. 11-27). With the introduction of radio and television, instructional courses increasingly made use of the new communication technologies to enhance and broaden the accessibility of their offerings (Sherry 1996). While still not having the functionality or accessibility of a modern E-Learning platform, it did push the boundaries of using cutting edge technologies to promote better communication between students and teachers.

2.2.1. First generation E-Learning platforms

The 1960s represented a transformative era in education and technology, setting the stage for the development of the first E-Learning platforms. As Cuban (1986, p. 34) notes that this period marked a shift from traditional instructional methods towards more interactive and technologically mediated learning experiences (initially through the use of classroom based telecasts). The societal push towards digitalisation, combined with significant advancements in computer technology, facilitated the creation of early E-Learning systems.

These systems aimed to avoid geographical barriers that had traditionally limited the audience and reach of educational activities and make educational content more accessible to a wider global audience. This significant technological leap marked the beginning of the first generation of E-Learning, utilising computer systems as opposed to traditional mail, radio or television. One of the first reported E-Learning platforms was PLATO (Programmed Logic for Automated Teaching Operations), initially developed at the University of Illinois in the early 1960s to assist with developing and delivering student literacy programs (Chaubey & Bhattacharya 2015). The system was a timeshared learning management tool which was commercialised and became the direct ancestor of modern learning management systems such as Blackboard and WebCT (Nicholson 2007, pp. 4-5; Woolley 1994, pp. 1-4). While still not quite what we would consider an E-Learning platform, it was a significant step towards E-Learning.

The next major technological leap was in the mid-1980s, with computing technology becoming cheaper, and the development of Project Athena, which was developed to take advantage of new distributed computing capacity. The project was the result of collaboration between MIT, Digital Equipment Corporation, and IBM, three of the largest technology companies in the world at that time (Chaubey & Bhattacharya 2015).

A decade after Project Athena, several new platforms were beginning to emerge but these were usually implemented proprietary formats developed for a specific course and were not highly extensible (Dagger et al. 2007). This era also saw the development of systems for other platforms such as FirstClass which was designed by SoftArc for the Macintosh platform (Chaubey & Bhattacharya 2015).

These platforms served as the foundational prototypes for the contemporary E-Learning systems and infrastructures we observe today, predominantly spearheaded by US-centric institutions in higher education and technological sectors. Although these initial systems were often characterised by their limited functionality and course-specific nature, they laid down fundamental components that continue to be integral to the architecture of modern E-Learning environments.

Despite the innovative leap, first-generation E-Learning platforms faced numerous challenges. Kirkwood and Price (2013) outline that early E-Learning platforms were often constrained by the technological limitations of the time. These included limited bandwidth, rudimentary graphical interfaces, and the high cost of computer hardware, which restricted access for many potential learners. Moreover, there was a lack of pedagogical frameworks guiding the use of these technologies, leading to varied effectiveness in their implementation. User acceptance was another significant hurdle, as both educators and students had to adapt to new modes of teaching and learning that differed markedly from traditional classroom environments.

The evolution of first-generation E-Learning platforms was closely tied to concurrent technological advances. As highlighted by Khan, BH & Ally (2015, pp. 51-8), the period saw rapid developments in computing power, data storage, and networking capabilities. These advances allowed E-Learning platforms to offer more sophisticated and interactive content, moving beyond simple text-based instruction to include multimedia elements such as images, audio, and video.

The introduction of distributed computing and timesharing systems, exemplified by Project Athena in the mid-1980s, further revolutionised E-Learning by enabling more scalable and accessible learning environments. These technological strides paved the way for the next generation of E-Learning platforms, which would leverage the Internet to offer unprecedented access to educational resources worldwide.

2.2.2. Second generation E-Learning platforms

Since the turn of the century, innovation and development complexity in E-Learning platforms have increased significantly, leading to a diverse range of systems available today. Modern platforms such as WebCT, Blackboard, Moodle, and Sakai are distinguished by their modular design architecture, facilitating semantic functionality, and supporting a variety of learning and teaching methodologies.

A crucial advancement in these systems has been the move away from the previous 'black box' approach towards a model of open service accessibility (Dagger et al. 2007). This provides a more sustainable and extensible platform and provides separation of content from administration whilst enhancing reporting, tracking and analysis capability. Open source platforms which provided free access to teachers and training program developers were introduced early in the new century with Moodle being the most popular and long lived for this period of time and generation of E-Learning technologies (Dougiamas & Taylor 2003). This removed the financial barrier for many providers and triggered a substantial growth in the development of E-Learning courses. As Siemens & Long (2011) discuss, the incorporation of analytics in education has enabled educators to customise learning experiences to the individual needs of learners, enhancing the effectiveness of online learning environments.

Furthermore, an increased emphasis on user experience (UX) has become an important aspect of the second generation of platforms. The focus on designing intuitive and accessible interfaces has been critical for fostering user engagement and satisfaction. Margaryan, Bianco & Littlejohn (2015) highlight the importance of instructional quality in Massive Open Online Courses (MOOCs), which extends to E-Learning platforms at large, underscoring the necessity of UX considerations in the development of educational technologies. This emphasis on UX has led to platforms that are not only more user-friendly but also more capable of supporting wider range of learning styles and individual preferences.

The introduction of open-source platforms, notably Moodle, has removed financial barriers for many educators and institutions, allowing for the greater use of E-Learning technologies in the field of education. The shift towards cloud technology further exemplifies the technological evolution in this domain, allowing for the seamless integration of a full suite of web-based tools without the need for installing or maintaining additional software (Chaubey & Bhattacharya 2015).

2.2.3. Third generation E-Learning platforms

Advancements made throughout the second generation of E-Learning not only standardised platforms to deliver quality educational content but to also leveraged the online digital nature of E-learning. What characterises the third generation of E-Learning platforms is the integration of cutting-edge technologies that enable tasks and learning experiences previously impossible in physical educational environments.

Cloud computing

One of the most significant technological advancements powering the third-generation platforms is cloud computing. The ability to remotely outsource computing resources to third-party providers offers substantial benefits to educational institutions, including reduced start-up and maintenance costs, and increased flexibility during upgrades (Khan, MA & Salah 2020).

In a study by Eljak et al. (2023), which examined 154 scholarly articles from 2010 to 2020, cloud computing was identified as the critical factor in the overall effectiveness of E-Learning platforms. The study identified aspects such as system architecture, software, performance, and the option to use more appropriate service models.

Wu & Plakhtii (2021) further highlight these benefits, in their examination of cloud based E-Learning platforms, specifically focusing on the 'Blackboard Learn' LMS. Their research showed a significant impact of cloud computing on the performance of learning environments, particularly through the dynamic scalability and resource efficiency that the cloud infrastructure provides.

Learning analytics

The second defining characteristic of the third generation of E-Learning is the integration of learning analytics (LA); utilising big data to enhance various aspects of education. According to Avella et al. (2016), who performed a systematic literature review of research from 2000 to 2016, LA benefits include personalised learning, curriculum refinement, and improved outcomes for students, instructors, and institutions. The study also highlighted challenges such as ethical and privacy concerns, as well as issues with data tracking.

Not only was there increased access to student data, but the focus of research into LA also changed focus, from student outcomes to a wider focus on the overall student learning experience. An analysis of 252 papers on LA in higher education from 2012 to 2018 by Viberg et al. (2018), suggested this shift in focus of the field beginning around that time period, where the field was Initially dominated by predictive methods aimed solely at predicting outcomes such as student retention and grade outcomes. However, since then research has since moved toward understanding the student learning experience, with an increasing emphasis on relationship mining and the collection of student data for human judgment rather than purely predictive modelling.

Fischer et al. (2020), in a review on big data in education, identified its application into three levels of student data:

- Microlevel - Involving detailed data, such as clickstream interactions. This level of data is captured during real-time learner interactions with platforms such as LMSs, MOOCs, and intelligent tutoring systems, supporting the analysis of individual behaviours and learning paths.
- Mesolevel - Involving textual data, including student writing and discussion posts. The greater amount of context enabling greater insights into learners' cognitive, emotional, and social development. With this greater amount of data, tools such as Natural Language Processing (NLP) are used to identify trends in understanding and affective states.
- Macrolevel - Institutional data, such as demographics, admissions, and course records. This type of data is typically collected over longer timescales. Applications of macrolevel data include early-warning systems, course guidance platforms, and administrative decision-making tools.

Fischer et al. (2020) also emphasised the actionable knowledge derived from big data, such as tailoring interventions to specific student subgroups and assessing the effectiveness of educational strategies.

Finally, it must be mentioned that there is a distinction between LM and Educational Data Mining (EDM). This distinction is outlined by Cerezo et al. (2024), in a systematic literature review of 129 papers published between 2012 and 2021. Finding that while both fields aim to improve educational processes, LA is more practically focused and has experienced faster growth, partly due to its broader appeal. In contrast, EDM is more technically oriented, with a specialised community. However, despite their differences, the fields have converged in methodologies and applications over time while maintaining distinct identities in terms of journals and conferences.

Regarding the tools used in applications of both LA and EDM, Paz & Cazella (2019) identify a range of tools in their systematic review of 10 articles published from 2008 to 2019. Computational tools include Google's MotionChart, QlikView, Tableau, and Analytics Dashboards, alongside datamining algorithms like Apriori, decision trees, and clustering techniques. These tools support academic analytics and enable advanced data processing for educational management.

Gamification

Another defining aspect of the third generation of E-Learning platforms is the increased use of gamification to enhance engagement and learning outcomes of students. Gamification incorporates game design elements into educational environments to motivate and immerse learners. In their systematic review of 90 papers, Denden et al. (2024) the prevalence of various gamification techniques and theories in digital higher education. The study found that points, badges, leaderboards, levels, feedback, and challenges were the most implemented game elements. However, the researchers noted a significant gap between theoretical frameworks and their application in gamified learning systems. Most studies lacked grounding in gamification theory, underscoring the need for a stronger connection between research and practice. Additionally, they observed a growing trend towards data-driven adaptive gamification, supported by machine learning techniques to personalise learning experiences.

A tailored approach to gamification has emerged as a focus of recent research, highlighting its potential to enhance learning outcomes by addressing individual learner characteristics. Oliveira et al. (2023), in their systematic review of 19 studies published between 2014 and 2020, examined the role of personalisation in gamification.

Their findings revealed that most studies centred on tailoring based on gamer types while often neglecting other critical human aspects, such as learning styles, personality traits, motivational stages, and demographic factors. The study proposed a two-level tailoring framework:

- Content Tailoring: Involving customising educational content to align with individual learner profiles.
- Game Element Tailoring: Involving adapting gamification components such as point systems or badge criteria, based on user behaviour and interactions captured through interfaces like cameras or sensors.

Despite the potential benefits, the research highlighted a lack of empirical evidence to generalise the positive effects of tailored gamification. For instance, Oliveira et al. (2023) suggested that learning outcomes differed between tailored and non-tailored gamified environments but acknowledged that this claim was supported by limited data (in this case a single study from the review).

Summary

The third generation of E-Learning platforms represents a significant evolution in educational technology, characterised by the integration of advanced technologies that redefine how learning experiences are designed and delivered. Cloud computing has emerged as a cornerstone of this generation, offering scalability, cost efficiency, and enhanced system performance, allowing educational institutions to leverage flexible infrastructures while improving the performance of platforms like LMSs.

Learning analytics further defines this generation by harnessing big data to personalise learning and educational content, with the goal of improving educational outcomes. Additionally, research has shown a shift from outcome-focused predictive models to a broader understanding of the student learning experience (Avella et al. 2016; Fischer et al. 2020; Viberg et al. 2018).

This evolution has been supported by tools like natural language processing and data mining techniques, which enable actionable insights at micro-, meso-, and macro-levels.

Gamification has also risen as a defining aspect of third-generation E-Learning platforms. The inclusion of game elements such as points, badges, and leaderboards has shown promise in enhancing engagement and learning outcomes. However, as Denden et al. (2024) and Oliveira et al. (2023) point out, the gap between theoretical frameworks and practical implementation remains a challenge. The move towards tailored gamification, focusing on individual learner characteristics, presents an opportunity for further development, despite limited empirical evidence supporting its effectiveness.

Looking ahead, the future of E-Learning platforms will likely be shaped by a combination of these technologies, with a strong emphasis on personalisation and adaptive learning. This focus will ensure that E-Learning continues to evolve in response to the diverse and dynamic needs of a global learner population.

2.3. What is E-Learning?

A standardised and commonly used definition of E-Learning remains elusive in the literature. This lack of definition in some part stems from a lack of consensus regarding terminology. Some authors use E-Learning and Online Learning interchangeably (Dringus & Cohen 2005), and others suggest that they are in fact very different and should only be used in certain circumstances (Nichols 2003). This variability in terminology, while allowing for innovation and diversity in educational technology, often leads to confusion among educators, learners, and researchers alike.

The overarching question when examining the current body of work in this field is, are they referring to the same thing? While using multiple terms for the same general concept can present innovative opportunities and be an easy way for commercial entities and education providers to enter the market and define their engagement, it presents potential confusion and avoidance by participants who do not know what to expect.

The field needs to have improved understanding, both from researchers intending to advance the literature, and those wishing to utilise its offerings. If each definition of E-Learning has a completely different meaning to another and represents a sub specialisation E-Learning, that will lead to incoherency in advancing the literature. Moving towards a more unanimous understanding of E-Learning will greatly benefit theory development.

2.3.1. Definitions of terms

This aim of this section is to discuss the varied definitions and understandings of E-Learning, highlighting the evolution of this concept over time and its implications for educational practice.

The exploration into the use of terms related to E-Learning reveals a diversity of descriptors. Moore et al. (2011) described three commonly used descriptors for E-Learning, namely; distance learning, online learning, and E-Learning. With these three categorisations, there are slight differences in technology, time, and location differences, as well as a good amount of overlap in usage of the terms. This is supported by a systematic literature review by Singh & Thurman (2019), which identifies differences in terminology depending on technology, time, interactivity, physical distance, and context.

Distance Learning

Historically, distance learning, as referenced by Keegan (1995) and King et al. (2001) denotes the separation of teacher and student through technology, requiring only geographic or temporal separation. This broad definition has evolved, from using lectures on videotape or correspondence courses to the sophisticated online platforms of today. The connection between distance learning and online learning remains nuanced, with the primary distinction often relating to the medium of delivery. Interestingly 'technology' does not necessarily equate to any form of online access or even the use of computers. With learning tools such as lectures on video-tape, audio-tape and telelecture, being referenced in distance learning studies in the 1980's (Beare 1989), and physical letter correspondence courses in the mid-19th century (Peters 1973 in Keegan, 1995:5).

Online Learning

The connection between distance learning and online learning itself is particularly vague, and primarily based on the type of methodology utilised. Seminal work by Moore et al. (2011) suggests that there is no real consensus in how online learning relates (or does not relate) to distance education or E-Learning. However, Singh & Thurman (2019) suggest that this comparison between distance and online learning is primarily about the distance between the learner and the source of education, and that this issue was primarily discussed up until the early 2000's.

There is limited published work to contextualise the connections between online learning and distance education using online mechanisms as its medium for example. It is therefore unclear whether the general understanding of online education is the same, similar or something altogether different, with only a minor relationship to distance education. For example, Nichols (2003) suggests that online learning describes a form of education that only occurs when delivered through online means, and therefore is not related to any form of physical materials or face-to-face. This suggests that E-Learning (referred to as 'eLearning' in that paper), is the use of technological tools that are based in an online context.

E-Learning

The term E-Learning itself exhibits several inconsistencies in spelling and general meaning, complicating the effort to achieve a consensus on its definition. Moore et al. (2011) noted that there were several variations in formatting (such as capitalisation of 'e', and the use of dashes or spaces after the 'e'). Research to date suggests that there are no regional or discipline consistencies in terminology, further increasing the potential for confusion and ambiguity. However, for the purposes of this research, the 'E-Learning' spelling variation was chosen, and used throughout for consistency, as many of the recent research in the discipline uses it (Prioteasa et al. 2023; Setiawan et al. 2023; Singh, P et al. 2023; Wairooy et al. 2023)

Given the rapid evolution of educational technologies, especially highlighted during the COVID-19 pandemic, necessitates a revaluation of the definitions surrounding E-Learning. Hodges et al. (2020) differentiate between the cases of emergency remote teaching, and that of actually planned online learning, providing a crucial context for understanding E-Learning. Emergency remote teaching is defined by Hodges et al. (2020) as a temporary shift in delivery to an alternate delivery mode, that is commonly due to crisis like circumstances (such as the COVID-19 pandemic). This contrasts with E-Learning that involves thoughtful design and pedagogical practices intended for online environments (Hodges et al. 2020). This distinction underscores the importance of deliberate pedagogical design and the integration of technology in defining E-Learning, beyond the mere use of digital tools for content delivery.

Furthermore, the scope and definition of E-Learning are influenced by the introduction of new technologies and teaching pedagogies. The 'Innovating Pedagogy 2019' report by Ferguson et al. (2019) identifies several emerging pedagogies that have the potential to transform educational practices, including the use of artificial intelligence for personalised learning, and social media as a tool for engagement and collaboration. These innovations highlight the expanding boundaries of E-Learning, which now encompasses a variety of teaching and learning practices supported by digital technologies (Ferguson et al., 2019). As such, E-Learning is not merely an electronic counterpart to traditional learning but a dynamic field that continuously adapts and evolves in response to technological advancements and pedagogical insights.

These perspectives suggest that E-Learning should be defined not only by the technological tools employed but also by the pedagogical strategies that underpin the educational experiences. The integration of thoughtful design, engagement, interactivity, and personalised learning opportunities are all hallmarks of effective E-Learning. Moving towards a more unified understanding of E-Learning will facilitate clarity in academic discourse, guide the development of E-Learning solutions, and ensure that educational practices keep pace with technological and pedagogical advancements.

Learning Technologies

An additional definition of the concept of utilising technology for the presentation of learning materials is that of 'Learning Technologies'. The term is used in a systematic literature review into technology integration in education by Laila Mohebi (2021), specifically in the context of the integration of the technologies into education. However, despite any growing adoption of the term in practice, scholarly literature predominantly references 'E-Learning' and 'Online Learning' as the primary descriptors of digital education. Singh & Thurman (2019) found that these terms dominated educational technology research between 1988 and 2018, with limited usage of 'Learning Technologies' in systematic reviews or theoretical discussions. This discrepancy suggests that while 'Learning Technologies' provides a broader, more inclusive framework, it has yet to gain the same level of academic recognition.

One challenge with the term ‘Learning Technologies’ is its ambiguity. It overlaps with other domains, such as ‘Machine Learning’ or ‘Educational Technology’, potentially conflating research areas and creating difficulties in defining clear boundaries. For example, searching for ‘Machine Learning Technologies’ often yields results unrelated to pedagogical contexts, further complicating its application in academic literature.

2.3.2. E-Learning categories

There are three commonly used terms to denote E-Learning management or delivery technologies; Course Management Systems (CMS), Learning Management Systems (LMS), and Knowledge Management Systems (KMS).

Course management systems

The first of the E-Learning categories is that of the CMS, which is one of less complex implementations of E-Learning. In its simplest form, a CMS is used to support the creation of a Blended Learning Course that involves a blend of both in-person, and online components (Watson & Watson 2007). This form of CMS is an administration support system used to track student performance, manage enrolments, associate students with courses, and facilitate communication. All these features provided, allow for the minimal amount of content required to perform E-Learning tasks, while providing enough resources to administrative staff.

While not as feature rich as an LMS, the CMS provides enough functionality to be useful for its intended purpose. CMSs are often confused with LMSs because they share a lot of the same functionality, however while a LMS commonly incorporates the functionality of a CMS, a LMS is more focussed on the participant than the manager (Watson & Watson 2007).

Learning management systems

The second category of E-Learning platforms is the LMS. This has more features and is generally what is thought of when an E-Learning platform is mentioned. An LMS is most commonly a web or cloud based system that directly facilitates the learning and teaching process and allows effective delivery of content beyond time and place restrictions (Chaubey & Bhattacharya 2015). Nichols (2003) also defines an LMS as a platform where online courses are assembled, consisting of a collection of E-Learning tools. Having the functionality of a CMS as well as additional learning tools, and features, the LMS is extensively used in higher education. The literature suggests that there are several ways to categorise an LMS. The first is by usage and accessibility. Chaubey and Bhattacharya (2015) propose three different categories;

1. Open-source
2. Software as a Service (SaaS)
3. Proprietary

These categories facilitate understanding of the different ways in which an organisation would acquire and implement an LMS. With open-source implementations being free to use (with restrictions), but requiring the organisation to provide its own troubleshooting, whereas proprietary or SaaS providing support, but at a cost.

According to multiple industry experts in LMS (Better Buys Staff 2023; Chang 2024; Ferriman 2017; Pappas 2015), there are generally four commonly agreed upon payment models (with some slight variations):

1. Pay per learner (or active learner).
2. Pay per use.
3. Licensing (limited or perpetual).
4. Free or Freemium.

Of the 25 active LMSs described by Ingwersen (2016), 44% were freemium, and of those mentioned as being freemium 36.4% utilised student number restrictions, 63% locked away features behind a fee/subscription, and one LMS utilised a limit on number of courses. Freemium models of LMS generally involve locking away some of the features of the full system or limit the number of users depending on the implementation for the specific system (Ingwersen 2016).

2.3.3. E-Learning environments

In addition to the functionality and implementation costs of an E-Learning platform, there is the different focuses of the platform, and how the participants are expected to interact with the system. According to Ouadoud, Rida & Chafiq (2021), there are four main types of online learning environments: Massive Open Online Courses (MOOC), Personal Learning Environments (PLE), Virtual Learning Environments (VLE), and Community, Content, & Collaboration Management Systems (C3MS).

The first type of learning environment is the MOOC, which is defined primarily by its ability to accommodate large numbers of participants, and that it is open to all regardless of institutional affiliation. With MOOCs being open to the public they generally must rely on either selling certifications or sub-licencing content to institutional users. MOOC platforms therefore tend to focus on a freemium strategy, with basic materials open and free to all users (Jia, Y et al. 2017).

Research by Ouadoud, Rida & Chafiq (2021) into MOOCs suggest two separate types of educational purposes; an informal learning network based on the education theory of connectivism, and the more traditional type of MOOC based on standard teaching materials and presented in a more cohesive manner.

The more traditional MOOC varieties are being integrated into universities' existing LMSs. Ouadoud, Rida & Chafiq (2021) suggest better integration would come in the form of allowing administrators to assign brands and credits to the student for using the additional connectivism style content, which students may or may not use, but would be available to use if they wish.

The last three types of learning environments are primarily differentiated by their purpose and intended audience as opposed to their capabilities and features, as with MOOCs. The first is Personal Learning Environment (PLE) which brings the focus to the individual learner, as its name suggests. Rather than being focused on connectivity amongst participants, its focus is on providing a suitable environment for the participant to structure their own learning resources (Ouadoud, Rida & Chafiq 2021). The opposite of that is the C3MS, which has a focus on community collaboration and management (Ouadoud, Rida & Chafiq 2021). In-between these two E-Learning environments is the VLE, which allows for groups and communities to be implemented, but also has a focus on the participant, taking a more measured approach compared to the PLE and the C3MS (Ouadoud, Rida & Chafiq 2021). Of these types of learning environments, the VLE is commonly confused with the concept of the LMS. For example Moodle is a LMS, however it can be utilised as a VLE if a more constructivist approach is followed, or used purely as a LMS if a more behaviourist approach is used (Pinner 2014).

2.3.4. E-Learning best practice

The introduction of E-Learning platforms into the educational landscape offered a considerable amount of flexibility regarding flexibility of use, and of access to an ever-increasing number of learning resources. Identifying a universal best practice in E-Learning given this increasing landscape, and diverse implementations remains a difficult task. However, through research and empirical evidence, several core elements appear to emerge as being critical in helping to assist educators in identifying what practices to follow.

These common elements that provide the most benefit to all those involved, are elements such as the students, the teaching staff, and the institution. Some of the factors identified by Castro & Tumibay (2019), that help to improve the efficacy of E-Learning platforms used in higher education institutions include; providing value to students through flexibility and personalisation, and to educators through the quality of technological infrastructure and organisational support.

Personalised learning

At the core of E-Learning's potential is the capacity for personalised learning. Unlike traditional educational models, E-Learning platforms offer an adaptable environment that can be customised to suit the individual needs of the learner. The significance of this personalisation is identified in the comprehensive review by Brusilovsky and Millán (2007, pp. 10-1), the review highlights the adaptability of E-Learning systems with regards to personalisation of content, interface, and feedback, depending on the learner's preferences and performance. Suggesting that this approach is critical for not only optimising the overall learning process, but also for enhancing learner engagement with the LMS.

Additionally, Xie et al. (2019) outline a systematic analysis of the overall evolution of personalised learning through technology enhancement. This personalisation was found to significantly contribute to the improvement of overall student outcomes, through a variety of personalisation techniques including sequencing of learning resources, automation of feedback, interface customisation, and adaptive content delivery. These techniques were suggested to help provide a more engaging and effective learning experience. These factor recommendations are echoed by Mikić et al. (2022) in their literature review of personalisation methods in e-learning.

Factors such as acceptance of E-Learning platforms and overall student satisfaction are important as well as overall importance. A systematic literature review by Fariani, Junus & Santoso (2022), of E-Learning literature from 2017 to 2021, identified that the students benefited from personalised learning not only in better learning outcomes, but also improved satisfaction of the LMS/content, as well as better acceptance and engagement with the system.

Organisational support

While it is indeed crucial to have a well-conceived Learning Management System (LMS) that boasts a wide array of features capable of personalising content and delivering value to both students and institutions, it's equally vital for these institutions to extend robust support to educators in content delivery. Maatuk et al. (2022) identify some of the issues that must be addressed, such as technical and financial support, training, and professional development for educators. Garrison, D. Randy, Anderson & Archer's (1999) introduces the Community of Inquiry framework, that highlights the importance of organisations supporting online learning environments through the inclusion of expanded social, cognitive, and teaching presence. This highlights the importance for institutions to provide a comprehensive level of support to educators, to better allow them to provide rich, interactive online learning communities that better support learners.

This does also suggest the importance of regular support for educators especially if they do not have backgrounds or skills in technology, as well as a focus on maintaining the IT infrastructure that supports the system. Turnbull, Chugh & Luck (2021) identified best practice involving four strategies for E-Learning organisational support. First, there should be transparent and multifaceted support provided to both students and staff involving learning materials. Second, a blended form of E-Learning (both in-person and online) should be included rather than online only. Third, like the suggestion by Maatuk et al. (2022) previously, there needs to be available training for both educators as well as students. Finally, there needs to be a focus on online connectedness, also referred to as 'virtual intimacy', to allow students to form learning communities. Maintaining a sense of connectedness, which will help to mitigate any negatives if the need to turn online-only should occur again.

Finally, support in the form of encouraging educators to engage with, and utilise LMS resources is an important factor to consider as well. A study by Diamond & Gonzalez (2016, pp. 401-8) on the use of digital badges in professional development, suggest that this approach to educator engagement, through recognition and incentivising skill development, was critical in encouraging a culture of continuous professional growth. This form of organisational support acknowledges the ever-increasing levels of technological competencies required for online teaching.

The efficacy of the use of E-Learning in higher education institutions depends on a balance of advanced technological features to promote a level of personalised learning, and a sufficient level of organisational support for educators that are required to implement said technological features. It becomes evident that successful E-Learning implementations are those that have sufficient organisational support, in addition to a wide variety of learning technologies. This support empowers educators through quality technological infrastructure and organisational backing, allowing them to maximise the benefits of E-Learning for students by providing a flexible, personalised learning experience.

2.3.5. E-Learning summary

This section on E-Learning delves into the complexities and evolving nature of digital education, highlighting the lack of a unified definition across the academic and educational fields. This ambiguity stems from the interchangeable use of terms like E-Learning, learning technologies, online learning, and distance learning, each carrying slight nuances in meaning and application. This would suggest a pressing need for a consensus to advance both theoretical understanding and practical application in the field.

E-Learning is characterised not merely by the technological infrastructure that supports it but more importantly, by the pedagogical strategies it employs. The distinction between emergency remote teaching, as necessitated by crises like the COVID-19 pandemic, and thoughtfully designed LMS environments highlight the importance of intentional pedagogical design in E-Learning. Innovations such as the use of AI for personalised learning and social media for engagement represent the expanding scope of E-Learning, pushing its boundaries beyond the traditional definitions.

The exploration into E-Learning terminology reveals a landscape marked by diversity yet plagued by inconsistency, suggesting the need for a more standardised approach to its conceptualisation. Historical perspectives on distance learning highlight the evolution from correspondence courses to the sophisticated online platforms of today, with a continued emphasis on the separation of teacher and student through technology. The relationship between distance and online learning, and the emergence of E-Learning as a distinct entity that leverages online technologies for educational purposes, further complicates the dialogue.

E-Learning management technologies, such as Course Management Systems (CMS), Learning Management Systems (LMS), and Knowledge Management Systems (KMS), play pivotal roles in the delivery of digital education. The LMS emerges as a comprehensive platform that facilitates a wide array of learning and teaching processes, adaptable to various educational settings and pedagogical approaches.

This section also addresses E-Learning environments, identifying key types such as MOOCs, PLEs, VLEs, and C3MS, each with distinct focuses and purposes within the broader context of digital learning. This diversity underscores the multifaceted nature of E-Learning, capable of accommodating a range of learning preferences and objectives.

Best practices in E-Learning are identified as critical for maximising the effectiveness of digital education platforms. These include the importance of personalised learning, which allows for the customisation of the learning experience to fit individual learner needs, and the necessity of robust organisational support for educators, ensuring they have the resources and training needed to effectively use E-Learning technologies.

In summary, E-Learning represents a dynamic and evolving field that transcends traditional educational boundaries through the effective integration of technology and educational pedagogy. Despite the challenges posed by varying definitions and terminologies, the core objective remains consistent: to enhance learning experiences through the thoughtful application of digital tools and strategies. The emphasis on personalisation and organisational support highlights the potential of E-Learning to offer flexible, engaging, and effective educational experiences, tailored to the needs of both students and educators.

2.4. E-Learning pedagogies and educational frameworks

To provide a more personalised learning experience, it is important to be able to identify the different approaches to both teaching and learning. There are many different teaching pedagogies (the method of teaching, and overall practice), and there is significand debate as to which is best. However, the focus of this section will be on current best practice in enabling an effective E-Learning environment. Additionally, it will identify suitable approaches for specific domains of study. Identifying pedagogies that are used in E-Learning is fairly simple, as Mayes and de Freitas (2007, pp. 14-23) noted that models for E-Learning have been repurposed from existing learning models. In addition, suggesting that most implementations will involve a blend of different learning theories; 'learning as behaviour, learning as the construction of knowledge and meaning, and learning as social practice' (Mayes & de Freitas 2007, p. 20).

To better understand E-Learning with regards to teaching pedagogies, we must first understand the pedagogies that E-Learning are commonly associated with, and how they emerged by investigating their backgrounds and differences. Then, once there is an understanding of the underlying pedagogical framework, there must also be an understanding of how these approaches are brought together to form a model of instructional design.

This section will provide analysis of both the pedagogical framework and common instructional design methodologies.

2.4.1. Behaviourist & cognitivist pedagogies

One of the earliest forms of learning theories is the behaviourist, which relies heavily on observable facts and outcomes. Early forms of Behaviourist research include the well-known research by Pavlov (1902) demonstrating classical conditioning in the form of learnt behaviours from external stimuli, and Skinner's (1965) Programmed Instruction research, demonstrating Operant Conditioning through positive and negative reinforcement to promote learning outcomes.

As the name implies, this approach focused on influencing behaviours in the participants, rather than engaging them in the learning process themselves.

The behaviourist approach is generally focussed on the learning materials, and how to best present them. Ally (2004) outline four strategies for a behaviourist approach to online learning. The first strategy is to properly explain the intended learning outcomes to students so they can judge for themselves if they have achieved the goals of the lesson. The second involves explicit testing of students to judge students' achievement levels. Third, proper sequencing of learning materials in level of difficulty is important, from known and unknown concepts. The final strategy focuses on providing adequate feedback so students can monitor their progress, and to take any corrective actions. As mentioned, none of these strategies involve personalisation or involving the participant other than to better judge if they have completed the task or not.

The behaviourist approach is also the simplest approach and requires it to be reasonably transparent to the participants. Harasim (2017, p. 12) suggests that for a behaviourist approach to be at all successful, learning objectives must be unambiguous and the performance/outcomes of the learning activity must be able to be judged and measured by a commonly agreed upon set of criteria. While this is not always possible, for example, an E-Learning task may be complete and a high score may be given, however the actual outcome (of teaching the participant), may not be as easily measured.

While limited in being student-centric, the behaviourist approach does have some benefits for simple E-Learning implementations. Research by Krouska et al. (2018) investigates the use of behaviourist based conditioning, in the performance of online tutoring of programming languages. The study implemented a methodology quite similar to Skinner's Programmed Instruction, which involved providing immediate feedback to participants, and a positive or negative reinforcement based upon performance (with correct answers receiving the positive reinforcement, and negative reinforcements after successive incorrect answers). Krouska et al. (2018) suggest that this approach is indeed both successful and popular with students, with the immediate feedback being found to be useful, and the positive and negative reinforcement being helpful for motivation to perform better.

While this is not surprising that this sort of approach is successful given the interactive nature of E-Learning, it is useful to understand that well researched educational theories dating back to the early 1900's are still relevant when it comes to the application of a modern E-Learning implementation.

In contrast to the behaviourist approach is the cognitivist approach, which rather than purely focusing on inputs and outputs, and positive and negative reinforcement; focuses on the internal aspects of the participants. Harasim (2017, p. 13) described the cognitivist learning theory as being focused more on the cognitive aspects of the participants as opposed to the 'black-box' approach of behaviourism. Suggesting that the behaviourist approach focuses almost purely on treating the participant as a machine to apply input stimuli to and expect well defined outputs to occur. The cognitivist approach instead focuses into the cognitive processes that intervened on the expected input/output processes of the behaviourist model of stimulus and response (Harasim 2017, pp. 13-4).

Given the increased focus on the participant, this approach will take considerably longer to implement properly but would potentially benefit from having a more personal learning experience. Ally (2004) suggests a cognitivist approach to E-Learning would require a more extensive and thoughtful strategy, in which students are given time to perceive and attend to the materials, so that it can be better transferred to working memory. That study provided recommendations such as paying more attention to delivery methods, to better facilitate transferal of meaning from the visual and audial sensations, as well as the pacing and medium (audio/text/video) of delivery.

Of the two approaches, the behaviourist approach is generally the least desired, as it is inherently dehumanising, and focusing on predominantly delivering learning materials, and measuring the output of said materials. However, it does have some benefit in a restricted environment, where only inputs and outputs are visible, such as a simple E-Learning environment and rudimentary assessment materials. However, the ideal would be the cognitivist approach, which would allow participants to have more agency in their learning, and to better accommodate any individual differences of the participants.

2.4.2. Instructivist & constructivist pedagogies

When discussing the uses of both Instructivist and Constructivist learning styles, the main issue is commonly; what is the purpose of instruction? And what is the role of teacher and instructor? Porcaro (2011) summarised the comparison between Instructivist and constructivist as both being two poles at opposite ends of the educational practice continuum; where the role of student is, either the recipient of the teacher's instruction (instructivist), or an active participant of knowledge creation (constructivist), and the role of teacher is either, centre of instruction (instructivist), or merely a facilitator (constructivist).

In addition to the dichotomy of instructivism and constructivism, there is also a differentiation between various constructivist theories; Mattar (2018) describes a selection of learning theories under the larger umbrella of Constructivism; 'situated cognition, activity theory, experiential learning, anchored instruction, and authentic learning' (Mattar 2018, p. 205).

The earliest forms of Constructivist theory were by Jean Piaget and Lev Vygotsky. Piaget's theory suggested that (in children), knowledge must be constructed through an incrementally more complicated sequence of situations (Piaget & Inhelder 1967), while Vygotsky suggests that the social activities of the participant were more important than the biological ones such as those suggested by Piaget (Vygotski 1929).

The key idea is that knowledge should not be just given (or flashed on-screen), but must be given incrementally, and allowing the participant to build up their own understanding in a manner that they are comfortable with. While Piaget was concerned with the development of children to adults, and the stages of mental development in-between, for the context of E-Learning, we would look at the constructivist approaches specifically applying to adults.

Similarly, Vygotsky's approach was a more developed investigation into the social functioning of participants, whereas in the context of an E-Learning environment, this social interaction would be (mostly) purely web-based interactions such as forum posting, direct messaging or emails.

After discussing the dichotomy between Instructivist and Constructivist pedagogies, it's crucial to understand the historical roots of this debate. The tension between Nativist and Constructivist views, particularly in early childhood learning and language acquisition, laid the groundwork for these pedagogical theories as discussed in a review of the debate between Jean Piaget and Noam Chomsky by Marras (1983). The debate was notably characterised by the contrasting perspectives of Piaget and Chomsky. While Piaget emphasised the role of environmental interactions in cognitive development and advocated for a child-centred approach in education, Chomsky argued for the existence of innate cognitive structures, postulating a Language Acquisition Device and emphasising the biological basis of language learning and cognition.

The constructivist approach, with its focus on staged cognitive development, creativity, and the importance of scaffolding instruction, not only influenced traditional educational practices but also played a significant role in shaping early E-Learning systems. These systems often incorporated principles of constructivist learning, fostering an environment where learners could actively construct knowledge through interaction and exploration.

In the context of E-Learning, the integration of Instructivist and Constructivist materials, as outlined by Moule's (2007) E-Learning ladder (Figure 2), reflects the potential synergy between these approaches. By providing a range of materials that cater to both instructivist and constructivist learning preferences, E-Learning platforms can offer flexible pedagogies that accommodate individual learner needs, fostering a more inclusive and adaptive learning environment.

This image has been removed due to copyright restriction. Available online from https://doi.org/10.1080/09687760601129588

Figure 2 - A conceptual model of online learning: the E-Learning ladder (Moule 2007)

Situated cognition and Activity theory

One of the mentioned sub-theories of constructivism that relates particularly well to E-Learning is that of situated cognition. According to Lave & Wenger (1991, pp. 33-5), the concept of situated learning is more than that of 'earning by doing', but rather about participation by the 'whole student', not as just a receiver of information from an instructor. This is further elaborated by Mattar (2018) describing situated cognition as learning in which an emphasis is placed upon the context and interaction involved in the activity. With regards to E-Learning, context involves the method of content delivery such as a browser or an LMS. This could also extend to the physical location in which students perform their learning activities (ranging from lecture theatres, workshop rooms, or their own bedroom on a smart phone).

An important aspect of making situated cognition transferable across different situations, is to accurately classify contextual information such as context-specific knowledge, which may potentially depend on domain-specific situations. Abu-Rasheed, Weber & Fathi (2023) found that approaches to classifying contextual factors were based on considerations such as the pedagogical setting, implementation infrastructures, and domain-specific requirements.

While the exact environment in which a student performs their learning activities will not be available for the purpose of this research, however, we can accurately assume that the activities will be conducted on either a traditional computer environment (desktop or laptop, and using an internet browser), or on a smart phone (with similar browsing environments in general). This situational context is obviously different to that of a student watching a lecture for the first time live in a lecture theatre or participating in a physical workshop environment.

The second sub-theory of constructivism that relates to the technological nature of E-Learning is that of Activity Theory, developed by Kaptelinin, Kuutti & Bannon (1995). This theory suggests a difference between internal and external activities, with an emphasis that both external activities must be internalised to be understood, and consequentially internal activities must also be externalised.

2.4.3. Connectivist & collaborativist pedagogies

Both connectivist and collaborativist learning pedagogies focus primarily on community and networking with regards to learning. Instead of focusing on the psychological aspects of learning, or the purpose of instructor and learner, it investigates how learning is achieved through external factors (external to the materials presented) affect learning, and how networking amongst learners can be beneficial. While both situated cognition and connectivism recognise the context of the educational interaction, situated cognition is primarily focused on the aspects of physical and social environments and not that of digital learning environments.

Outlined by Siemens (2004), connectivism is based upon the circular nature of networking, where an individual participant's knowledge is fed into the network and in turn spread throughout an organisation or institution. The cycle is then ultimately fed back to the participant, increasing their knowledge through the input of others in the network.

Harasim (2017) defines connectivist learning theory as being several learning processes that lead from divergent thinking to convergent thinking; where divergent thinking is defined as a process to generate a multitude of ideas, and convergent thinking is the process of eliminating the weak ideas generated from the divergent process. Siemens (2004) also suggests that this networking approach would allow learners to be kept up to date within their field of learning, which makes sense if the network is able to feed in up to date information.

More recent applications of this connectivist approach can be seen in the form of discipline-focused communities in messaging platforms such as Discord. Heinrich & Carvalho (2022) discuss two such communities in their research paper; Computer Science, Information Technology, Mathematics and Statistics (CSIT), and Veterinary (VET). Both Discord communities allowed for an external (to the LMS) environment for students to connect and engage in professional networking/development, as well as ask topic related questions in a non-formal environment.

Goldie (2016) argues that a pure connectivist approach to learning could be perceived as having a lack of direction, and teacher control, while additionally being difficult to assess. However, he also notes that there would be a significant amount of data generated from the online connectivity and collaboration to be able to evaluate learning outcomes better (Goldie 2016). This is an important aspect with regards to E-Learning, and in-particular this research. The large amount of data generated has the potential for deeper analysis of interactions of not only user to course, but also user to user, and user to facilitator.

Similar in nature to a connectivist approach, collaborativist learning theory is based upon the collaborative nature of group work, and the benefits associated. While group work is not the most common form of activity represented by E-Learning, it is more than possible in a virtual space, using videoconferencing, email, and instant messaging.

While typically not a large focus of E-Learning research, when compared to improving methods of teaching and LMS optimisation, connectivist and collaborativist styles of learning processes are still worth investigating. This is especially the case looking at social interactions between E-Learning participants, and between participants and facilitators, and how these interactions affect overall performance of the LMS, and of the learning materials.

2.4.4. Instructional design models

This section emphasises the necessity of a well-structured design and implementation process in creating E-Learning materials or courses. Instructional design processes, as Branch and Dousay (2015, p. 31) discuss, are both prescriptive, guiding through optimal procedures and general strategies, and descriptive, highlighting the interplay between different processes and actors.

These models are not only vital for educators who are unfamiliar with or new to the design process but also help structure an optimal flow for creating effective learning materials. This section will discuss several popular instructional design models suitable for E-Learning applications and then identify an optimal model to work with and optimise, drawing insights gained from this research.

Popular instructional design models

A diverse range of models is available for instructional design, each with unique characteristics and applications. Branch & Dousay (2015, pp. 35-92), provide a comprehensive taxonomy and descriptions of various models. This section will select models that are particularly applicable to E-Learning, based on their intended purpose and suitability, to discuss their overall design and the flow of their processes.

One of the foundational and relatively straightforward models to consider is the ADDIE model, as detailed by Molenda (2003). Molenda (2003) that the origins of the model may trace back to its development by the U.S. armed forces in the mid-1970s. It is widely adopted as a framework for creating customised instructional design models. The generic steps of the ADDIE model, depicted in Figure 3, include Analysis, Design, Development, Implementation, and Evaluation.

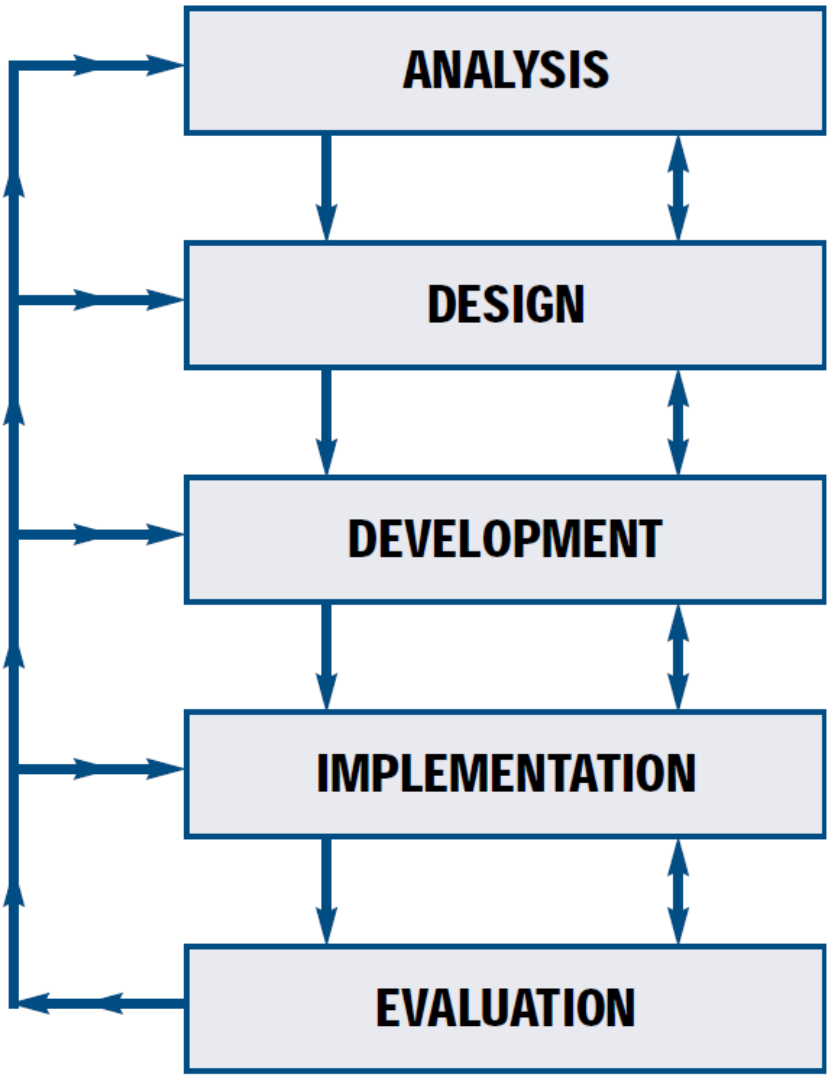

Figure 3 - Generic ADDIE model (Grafinger 1988)

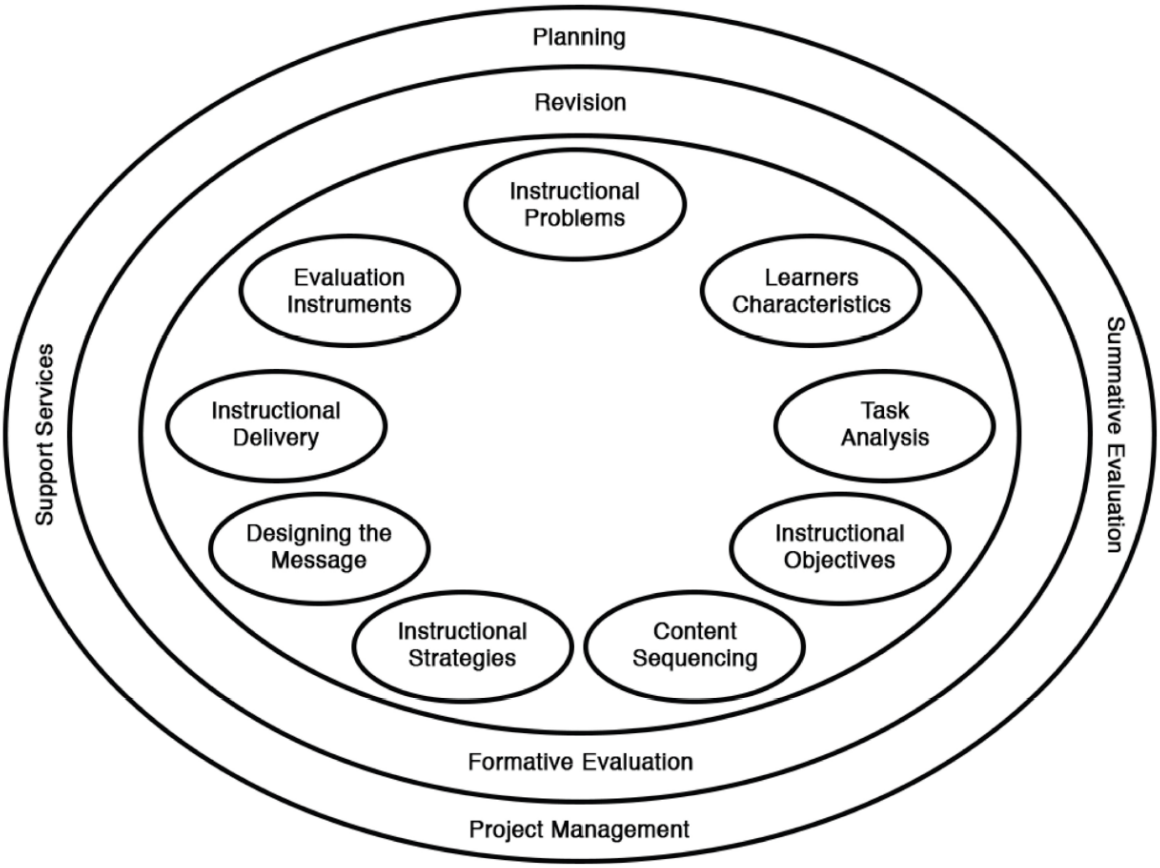

Figure 4 - Morrison, Ross, Kalman, and Kemp model (Morrison et al. 2019)

The next model that will be discussed is the Morrison, Ross, Kalman, and Kemp model outlined by Morrison et al. (2019). This model adopts a more holistic approach, considering various factors like learner characteristics, subject matter, and learning context. Branch & Dousay (2015, pp. 48-50) suggest that this approach is more learner-centric than traditional design practices and includes provision for technological approaches. Figure 4 outlines the overall processes involved in the model.

The subsequent model engages the design process in a fundamentally different way; by outlining the overall educational goal and working backwards. Wiggins and McTighe's (2005) Backward Design model guides educators in designing outcome-focused courses, starting with the end goals and designing the instructional path accordingly.

Following the process described by Branch & Dousay (2015, p. 51), the process begins with the end goal in mind (Step 1 - identifying the educational goals required for the learning materials). It then moves towards Step 2 - identifying the metrics required to measure the transfer of knowledge. Finally, Step 3 involves identifying activities and resources that will enable the success of the previous steps. This input and output-focused model may be considered more behaviourist, focusing on the transfer of knowledge rather than its acquisition, and adopting a more student-centric approach.

Next is the Four-Component Instructional Design (4C/ID) by Van Merriënboer & Kirschner (2017), another holistic model considering the learners and their cognitive requirements. The model, outlined in Figure 5, demonstrates the four main areas of learning.

Figure 5 - Van Merriënboer & Kirschner 4C/ID model (Van Merriënboer & Kirschner 2017)

The 4C/ID process involves 10 steps, as outlined by Van Merriënboer & Kirschner (2017):

Learning Tasks

1. Design learning tasks.
2. Sequence task classes.
3. Set performance objectives.

Supportive Information

4. Design supportive information.
5. Analyse cognitive strategies.
6. Analyse mental models.

Procedural Information

7. Design procedural information.
8. Analyse cognitive rules.
9. Analyse prerequisite knowledge.

Part-task Practice

10. Design part-task practice.

Next is the Sims & Jones (2002) Three-Phase Development (3PD) model, which involves consideration of long-term development and collaborative goals. Compared to other models, the 3PD model focuses primarily on a quick initial development cycle to produce functional educational materials, which can be iterated on subsequently. The steps involved in the 3PD are outlined in Figure 6.

This image has been removed due to copyright restriction. Available from (Sims & Jones 2002, p. 4)

Figure 6 - Three-Phase Design model (Sims & Jones 2002)

Next is Dabbagh & Bannan-Ritland's (2005) Integrative Learning Design Framework (ILDF). The ILDF model focuses on a variety of aspects, as well as research methodologies, additionally this approach uses multiple micro-cycles of research, as depicted in Figure 7.

Another distinctive model is the Dick, Carey & Carey (2014) model, which sets itself apart from other models by incorporating an initial needs assessment step and parallel steps (as shown in Figure 8) of instructional analysis and learner analysis. This model is particularly beneficial for E-Learning when a comprehensive and systematic design is required, ensuring that all instructional components align with the learning outcomes.

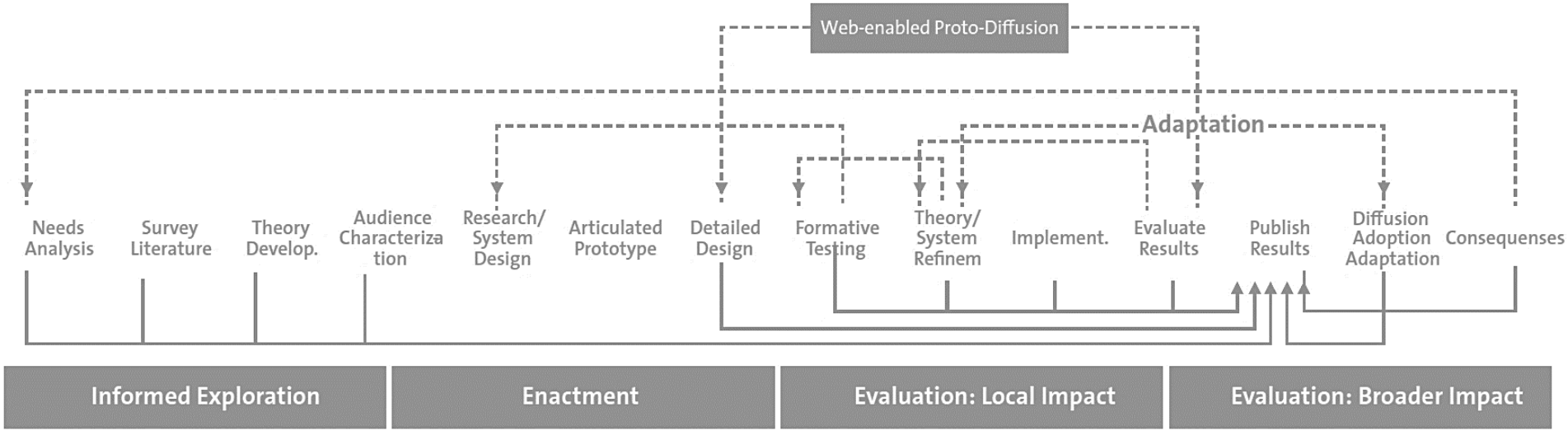

Figure 7 - Integrative Learning Design Framework (Bannan 2013)

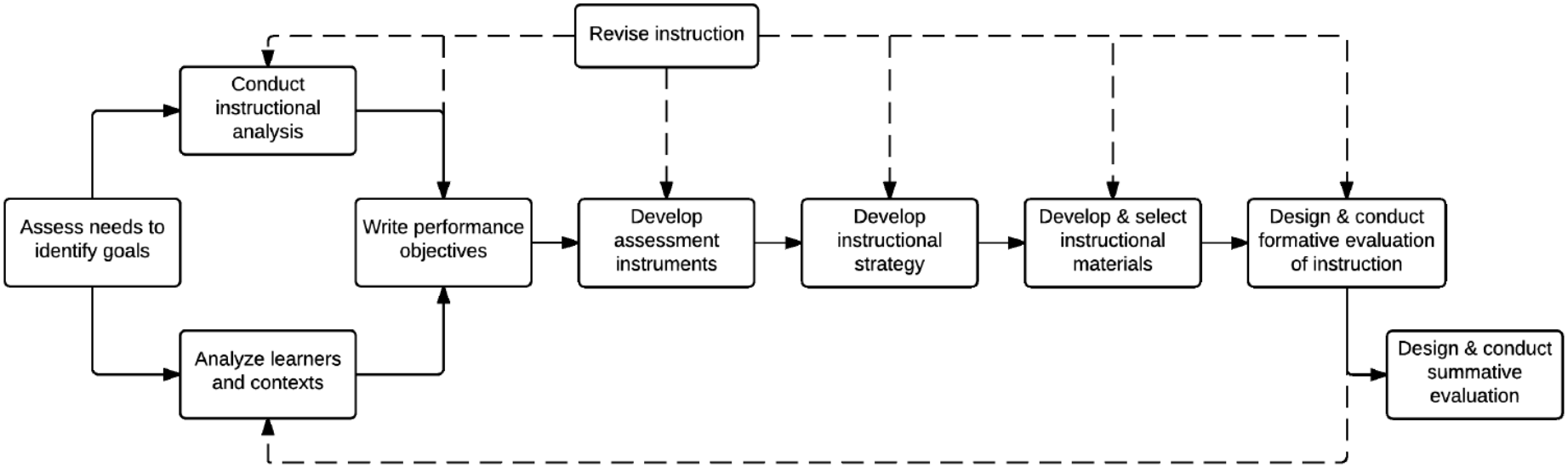

Figure 8 - The Dick, Carey, and Carey model (Dick, Carey & Carey 2014)

The final model to be examined is Merrill's (2002) Pebble-in-the-Pond design model. As shown in Figure 9, this model involves a series of concentric activities centred around the initial problem, in this case, the development of an E-Learning solution. Being task-based, this approach can be instrumental for E-Learning solutions, with its emphasis on problem- centred instruction, promoting active learning and engagement.

This image has been removed due to copyright restriction. Available from (Merrill 2002, p. 40)

Figure 9 - Pebble-in-the-Pond instructional development model (Merrill 2002)

Rationale for selection of instructional design model

The next step is to identify which one of the models mentioned would be best suited for E-Learning tasks and adaptable for use in this research. The criteria for selection were primarily based on three features; Effectiveness in E-Learning situations, Customisability and Adaptability, and the ability to extend upon based on requirements of this research.

Regarding suitability and effectiveness in E-Learning, the ADDIE model has demonstrated effectiveness across various E-Learning environments, as highlighted in a comparative study by Spatioti, Kazanidis, and Pange (2022). Furthermore, an experiment conducted by Almelhi (2021) tested the ADDIE model's efficacy in teaching creative writing to ESL students within the Blackboard LMS. The findings indicated improvements in student performance and in the development of learning materials.

Regarding customisability and adaptability; Spatioti, Kazanidis, and Pange (2022) also discussed the flexibility ADDIE to be able to accommodate a variety of different applications across different domains. The ADDIE model is a systematic instructional design framework used to guide the creation of educational and training programs. As shown in Figure 3, the acronym ADDIE stands for the five key stages in the process:

Analysis - The analysis stage identifies the learning problem, goals, and objectives, considering both the learners' needs and the learning environment.
Design - Involves planning the learning experience, including the instructional strategy, learning objectives, delivery methods, and assessment strategies.
Development - Involves the creation of learning materials, which can include digital content and the integration of technology.
Implementation - The course or training program is delivered to the learners, involving setting up the LMS and ensuring all materials and technology are implemented correctly.
Evaluation - Assesses the effectiveness of the instructional design by collecting feedback from learners and instructors to evaluate whether the learning objectives were met and to identify areas for improvement.

Finally, regarding the ability to customise the design framework for use in this research is considered. This involves only the adjustment to one of the phases of the process, namely the analysis phase. As previously mentioned, the analysis stage considers a wide variety of factors regarding the participants of the E-Learning environment (such as the learners, and educators). Additionally, this stage employs a wide variety of methodologies tailored to different tasks and industries. Piskurich (2015, pp. 103-4) outlines multiple types of analysis suitable for instructional design. However, Audience Analysis and Delivery Analysis are the most pertinent to discussions about E-Learning, as they respectively identify the learner characteristics and teaching/delivery methodologies. While these methodologies are described by Piskurich (2015) from the perspective of corporate training and are presented as separate analyses, combining them could be beneficial for an E-Learning approach.

Figure 10 shows an extended version of the analysis phase of ADDIE, described by Bąkała & Bąkała (2020), which incorporates the additional learner and delivery characteristics mentioned. The extension was developed by Bąkała & Bąkała (2020) using the Business Process Modelling Notation (BPMN) 2.0, which is a broadly accepted standard for modelling the execution flow of business processes (Meland & Gjære 2012).

This extended analysis phase outlines six steps for project managers to follow, with the middle two steps being particularly relevant to this research and the customisation of materials: 'Analyse Learners' and 'Audit Available Resources.' The 'Analyse Learners' step involves identifying individual differences among students (as discussed in the upcoming Section 2.5 Individual differences in learners) and tailoring content and/or delivery accordingly. Subsequently, the 'Audit Resources' step includes analysing the teaching team's pedagogy (as discussed in Section 2.4 E-Learning pedagogies and educational frameworks) and identifying the most appropriate teaching pedagogy for the given situation.

The structured yet flexible approach of ADDIE suggests its appropriateness. While other methodologies have similar structures, ADDIE is relatively simple and allows for adjustments at each stage. It provides a clear, step-by-step framework that can guide the entire process of instructional design, from analysis to evaluation. Branch (2009, p. 168) suggests that the structured approach of ADDIE ensures thoroughness and aids in managing complex E-Learning projects. Despite its structured nature, ADDIE is adaptable and, as mentioned by Molenda (2003), can be customised to fit various learning contexts and needs. It doesn't prescribe specific methods or tools, allowing educators to choose the best strategies and technologies for their projects.

One notable application of ADDIE is combining it with the Classical Waterfall Model (Leach 2018, p. 14) used in software Engineering methodologies. As discussed by Wan Ali & Wan Yahaya (2023), ADDIE was able to be combined with the Waterfall Model in the integration of a digital video learning application into an on line learning platform.

Another factor favouring the ADDIE model is its popularity and current use in various industries. As mentioned by Allen (2006) ADDIE is one of the most well-known and widely used models in the field of instructional design. Its prevalence means that a greater number of educators are already aware of its structure, making any customisations easily understandable. Additionally, Spatioti, Kazanidis & Pange (2022) suggest that the ADDIE model has been successfully used in a wide range of instructional design projects, including E-Learning, corporate training, and educational programs. More recent studies such as that from Li & Abidin (2024) also show the application of the ADDIE model to have improved the teaching skills of educators in experiment groups compared to control groups not using the ADDIE methodology.

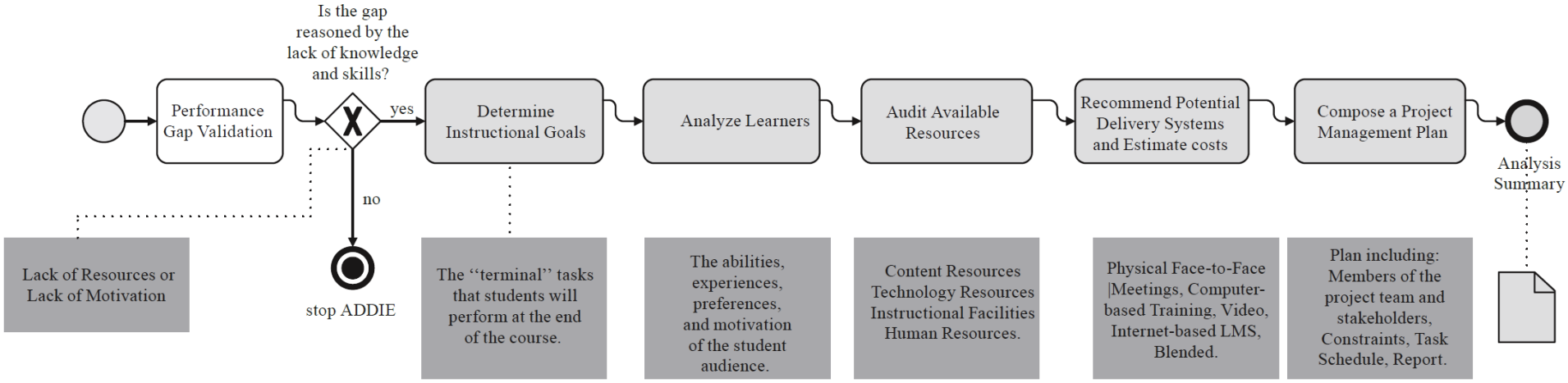

Figure 10 - ADDIE analysis phase BPMN 2.0 standard (Bąkała & Bąkała 2020)

2.4.5. Pedagogy and instructional design summary

This section synthesises key findings from the exploration of prevalent E-Learning pedagogies and instructional design models. It underscores the pedagogical spectrum spanning from behaviourist to cognitivist approaches, each with their own benefits and applications. The behaviourist approach emphasises observable inputs and outputs, and delivering information to learners, whereas, cognitivist approaches, favouring an internal, student-centric learning processes, requiring a more customised approach to teaching.

The section also distinguishes between instructivist and constructivist approaches, defining the role of the instructor and the student in the learning process. Furthermore, it delves into the nuances of connectivist versus collaborativist approaches, especially relevant in determining the nature of student interactions and the formation of knowledge networks or collaborative group work.

Finally, the section revisits instructional design models, highlighting the ADDIE model's prominence due to its structured yet adaptable framework. The model's historical relevance, systematic approach, and widespread acceptance make it a preferred choice in various instructional design scenarios, including E-Learning, corporate training, and educational programs. The adaptability of the ADDIE model, coupled with its systematic nature, allows it to meet a diverse range of learning needs and contexts, making it an optimal choice for this research.

In conclusion, the exploration of pedagogies and instructional design models in this section provides a comprehensive framework for understanding the complexities of E-Learning. It offers a comprehensive investigation into various teaching styles and instructional design methodologies, setting the stage for selecting and applying the most effective strategies for E-Learning environments.

2.5. Individual differences in learners

When discussing how students learn it is important to understand that students are not all the same, and that they will require some level of customisation (in learning content, or teaching style), to help deliver educational content to them in a way that they will accept. This level of customisation can come in the form of tailoring content to a student's specific learning style, a student's level of self-efficacy in the area being studied (or the technology used for teaching), the level of cognitive load that the student has while studying, and preferred delivery timing of the content. While individual differences between students are not available with the dataset for this research, the concepts are still worth discussing, as they have the potential to be discussed regarding differences between colleges.

2.5.1. Learning style

The first individual difference in learners is that of a preferred learning style, which affects the way in which a student will interact with materials, and with the LMS itself. Understanding what learning style, most learners in an E-Learning course will require can help to better personalise the course to suit the needs of learners.

The first method of identifying learning styles of students that will be mentioned is the Felder Silverman learning style model (Felder & Silverman 1988). This model categorises learners across four dimensions, each with two values of each dimension: Perception (sensing and intuitive), Processing (active and reflective), Input (visual and verbal), and Understanding (sequential and global). Felder & Silverman (1988) describe four primary questions used to define a student's learning style. First, what sort of information do they prefer to receive, something more visual or tactile (sensing), or something more abstract requiring greater internal thought and investigation (intuitive)? Second, do they prefer the use of visual representation of information such as pictures, diagrams, demonstrations, or videos (visual), or do they prefer written or audio instructions (verbal)? Third, do they prefer to actively participate in physical activity or discussion of the material (active), or do they prefer a more introspective activity (reflective)?

Finally, do they prefer their information sequentially, in increasing difficulty to better understand each piece, or to examine the problem(s) as a whole (global)?

Another theory for identifying student learning styles is that of Gardner's theory of multiple intelligences, developed in the early 1980's by Howard Gardner (1983), and further elaborated by Davis et al. (2011). For this theory, the identified intelligences were divided amongst 8 different intelligence domains; linguistic (involving oral and written language), logical-mathematical (solving abstract mathematical problems), spatial (dealing with large scale spatial images), musical (handling patterns of sound), bodily-kinesthetic (use of their own body to solve a problem), naturalistic (dealing with plants or animals in solving problems), interpersonal (understanding the moods, motivations, and intentions of others), and intrapersonal (understand their own moods, motivations, and intentions).

The final method of identifying learning styles that will be mentioned, is that of the Learning Style Inventory (LSI), outlined by Kolb (1981). The LSI describes four primary learning styles, each style being a combination of preferring abstract conceptualisation or concrete experience, and active experimentation or reflective observation. These types are described as Convergers (abstract conceptualisation and active experimentation), Divergers (concrete experience and reflective observation), Assimilators (abstract experience and reflective observation), and Accommodators (concrete experience and active experimentation).

While it is possible to categorise learners into different learning style categories, it has been observed that these styles can be flexible depending on the task, and the situation. Rasheed & Wahid (2021) found that depending on the difficulty of the task, the learning style may switch from verbal-linguistic to visual-spatial, especially with high difficulty tasks involving simulations, case studies, and application-level questions. Additionally, Rasheed & Wahid (2021) found that 75% of learners shifted to an interpersonal style from an intrapersonal style when it was close to submission and assessment dates. This would suggest that not only do students have preferences in their methods of learning, the timing and level of difficulty of learning materials may also be a factor in the optimal methods of delivery.

Learning styles have been effectively applied in modern LMS platforms to enhance the delivery of educational content. For example, Kaiss, Mansouri & Poirier (2023) demonstrated the use of the 'LearningPartnerBot', a chatbot integrated into MOODLE, to recommend learning materials tailored to the Felder & Silverman (1988) Learning Style Model. This allowed learners to access content such as videos for visual learners, or textual content for verbal learners, appearing to help improve overall engagement and comprehension.

Detection of learning styles has been greatly improved upon by the incorporation of modern technologies such as Artificial Intelligence (AI). Kanchon et al. (2024) presented a system that leverages AI algorithms like decision trees and blending ensemble techniques to classify students' learning styles based on LMS activity logs. By incorporating machine learning and NLP, AI systems such as the one developed by Kanchon et al. (2024), can identify visual, auditory, and kinesthetic preferences with high accuracy. These systems also modify content formats, such as converting text to audio or creating mind maps, to better match individual learning styles. Such innovations illustrate how AI can seamlessly integrate into LMS platforms to optimise learner engagement and outcomes.

Once learning styles can be identified, it is then important to understand what can be achieved with this knowledge, and what the relationship between learning style and academic performance is. Al-Roomy (2023) identified significant correlations between the use of preferred learning styles and improved GPA among health sciences students, revealing that auditory learners often outperform their peers in certain fields. Additionally, the proper alignment of learning style to LMS implementation has been shown to enhance engagement and satisfaction. For example, kinaesthetic learners perform better in hands-on, simulation-based activities.

Understanding and leveraging individual learning styles is pivotal in designing effective educational experiences. The integration of theories like the Felder & Silverman (1988) Learning Style Model into LMS platforms, combined with AI-driven tools, has demonstrated substantial improvements in learner engagement and academic performance.

2.5.2. Chronotypes

In addition to learning style preferences, the time preferences of students can affect optimal learning experiences. For example, identifying a way to separate students into groups that prefer the delivery of content at certain time periods.

The development of research into time preferences began with work by Horne & Ostberg (1976) on human circadian rhythm differences; identifying 'morning types' and 'evening types', through questionnaires asking subjects to describe the best fit for both sleep and active times. The next development was by Roenneberg, Wirz-Justice & Merrow (2003), developing a standardised questionnaire to record the temporal structure of daily life; the Munich ChronoType Questionnaire (MCTQ), distributed in both Germany and Switzerland.

These early studies focus primarily on sleep quality and timing of sleep and activity periods, however, later studies into chronotypes focus more on the effect of different student chronotypes and learning performance. Using the Roenneberg, Wirz-Justice & Merrow (2003) MCTQ questionnaire, the study by Beşoluk, Önder & Deveci (2011) found that questionnaire results could partially predict academic performance, depending on the time of the teaching period. Additionally, teaching and test start times, with respect to identified student chronotypes, were shown to predict student academic performance.

Research by Ujma et al. (2020), suggests that there is an influence from societal and work-related pressures on both sleep patterns and of learning performance. This study also suggests that individuals with higher intelligence, potentially due to more flexible work or study schedules, may experience a reduced 'social jetlag'; defined by Wittmann et al. (2006, p. 497) as 'the discrepancy between work and free days, between social and biological time'. This indicates that they may experience a less pronounced difference between their normal sleep preferences and their actual sleep times. This perspective emphasises the importance of accommodating not just biological but also environmental factors in optimising learning experiences based on student chronotypes.

2.5.3. Self-efficacy

The next individual difference that will be discussed, relates to the feelings of the individual learner (or educator) regarding their competency with the E-Learning environment. Self-efficacy, according to Bandura (2002) is defined as an individual's belief in their ability to complete a task effectively. This level of self-efficacy in students generally affects how much effort a student will expend, the overall motivation towards completing work, and how long they will persist with difficult tasks (Bandura 1977).

For E-Learning environments, this manifests itself in overall persistence and resistance to dropping out of the topic or just avoiding interactions altogether. Where students with low perceived self-efficacy will be more likely to refrain from attempting learning tasks that they believe they are unable to complete.

Research by Dash et al. (2022) offers insightful findings on the transition to E-Learning, emphasising the role of self-efficacy, interaction, and E-Learning content in enhancing user satisfaction and intention to use E-Learning materials. Their study highlighted the moderating effects of enjoyment and choice on the relationship between these factors and user intention to use the LMS. Notably, the choice to engage in E-Learning, either by force or voluntarily, significantly influences this dynamic, highlighting the importance of autonomy in the learning process.

Similarly, Rankapola & Zuva's (2023) investigation into the impact of E-Learning quality, self-efficacy, and satisfaction during the COVID-19 pandemic reveals that self-efficacy significantly boosts learner satisfaction, which in turn, influences their intention to continue using the LMS. The study's findings, suggest that both the quality of E-Learning services and content, in addition to learners' self-efficacy, are crucial for fostering positive E-Learning experiences.

These studies collectively illustrate the detailed relationship between self-efficacy, the quality of E-Learning environments, and learners' satisfaction and engagement. Dash et al.'s (2022) multi-group analysis further highlight this by comparing the effects of nationality, gender, and respondent type, offering valuable insights for educational providers worldwide. While Rankapola's (2023) emphasis on the positive effects of self-efficacy on user satisfaction and the moderating role of E-Learning quality dimensions provides a strong foundation for enhancing E-Learning practices.

In summary, it becomes evident that both student and educator self-efficacy are critical for the success of E-Learning environments. As suggested by Munir & Waty (2023), fostering educators' self-efficacy through personalised professional development is essential for innovating E-Learning practices. The implications of these studies highlight the need for E-Learning platforms and policymakers to consider self-efficacy and its moderating factors in the design and implementation of E-Learning solutions, ensuring a supportive and effective learning experience for all participants.

2.5.4. Cognitive load

The last individual difference that will be discussed is based more on the cognitive capability of the learner (as well as how the content and presentation of the learning materials affect cognitive processing), as opposed to the previous differences, which relate to preferences in consuming learning materials (learning style preference and chronotypes), and the level of self-efficacy regarding the task in question.

Mayer & Moreno (2003) identify three kinds of cognitive load demands on the brain when presented with information; essential processing, incidental processing, and representational holding. Each type of cognitive demand related to E-Learning materials is different. Essential processing refers to the ability to make sense of images and text in a presentation, whereas, representational holding refers to the ability to hold that mental representation in memory. Finally, incidental processing as its name suggests; the refinement of extraneous content associated with the learning materials (such as colours, and background music).

In multimedia learning, information presented to learners is divided between two channels, verbal and visual. These channels have both a limited capacity (hence the ability to overload them), and actively learning requires substantial cognitive processing (Mayer, Richard E & Moreno 2003). Sweller, van Merriënboer & Paas (2019) expand on this by identifying the concept of germane cognitive load, which is defined as the cognitive load required to learn. This germane cognitive load refers to the cognitive resources devoted to dealing with the essential processing (or intrinsic cognitive load), rather than any extraneous load.

The key issue with regards to the cognitive load imposed on learners from learning materials (especially multimedia heavy materials) is that there is the potential in overloading the learner's cognitive capacity. Consequently, being able to distinguish essential elements from extraneous one to learning is an important task, otherwise, students may encounter situations in which they become overloaded.

Five types of cognitive overload are identified by Mayer & Moreno (2003) including single-channel overload of essential information (type 1), dual-channel overload of essential information (type 2), essential and incidental overloading (type 3), presentation overloading (type 4), and representational holding overload (type 5).

Type 1 generally occurs when both visual and audio materials are present, but one source of information requires extra processing (such as the visual). When the video becomes the focus of the visual processing, and overloads the visual processing ability of the student, the cognitive processing of the text will be diminished. Type 2 involves both the visual and verbal channels to be overloaded at the same time, which can be a common situation when dealing with lecture slides that are narrated over. When there is both visual information, such as imaged/diagrams, and text, as well as an audio component, learners may be overloaded if both text and audio are needed to be processed at the same time (such as the text being simultaneously narrated and written on screen), while the visual channel is also being loaded with additional information.

Type 3 overload involves both essential information as well as incidental information. This type of cognitive overload usually occurs when a large amount of unnecessary information is presented to make the learning materials more entertaining, or to include multimedia aspects for the sake of including them. Type 4 is very similar to type 3, but rather than it being the material itself, it is how the material is presented to the learner that causes additional incidental load. Such as placing materials that should be processed together far apart or presenting concepts out of order, which would increase the level of incidental processing required.

The final type of cognitive overloading (type 5) involves learners overloading their essential processing capacity as well as their representational holding capacity. This can occur when there are large amounts of essential material being presented, that cannot be fully processed until another section of material is presented. An example of this could be a detailed explanation of a phenomenon or process on one slide or page, and a subsequent diagram of said process/phenomenon on another. Instead of having both the description and the diagram on the same page/slide, the learner will need to take in all the information into representational holding, and then attempt to take in another selection of information to process it, while the previous is still not fully processed.

A common solution to dealing with cognitive overloading involves personalising content to learners. The solutions presented by Mayer & Moreno (2003) for types 1 to 5 of cognitive overload all involve in some way, personalising the content. To fix type 1 overloading, it is recommended to offload some of the content from the overloaded channel to the other: customising the desired mix of visual/verbal information. Type 2 overload is solved by allowing for more processing time, or to have materials available prior to the session, both are an intrinsic feature of E-Learning, having the ability to allow the student to decide how long they need to dwell on any given part of the material, and allowing access to additional materials.

Both types 3 and 4 involve dealing with extraneous materials (or identifying what is necessary for learning, and what is not), as well as how to present them in a logical fashion. The solution for both involves customising the materials to better suit how the learner will want to learn (maybe defined by their learning style), as well as identifying what materials will add to the learning outcome, and not be simply an extraneous addition.

Finally, to solve type 5 cognitive load, content that should be processed together needs to be presented at the same time, as well as personalising the content so that the level of processing required is minimised.

A study by Lange (2023) confirms that personalisation/customisation of E-Learning materials is a positive solution; showing a positive relationship between E-Learning personalisation and germane load, a negative relationship between E-Learning personalisation and intrinsic load, and a negative relationship between E-Learning personalisation and extraneous load. Altinpulluk et al. (2019) recommend smaller meaningful units of materials, as well as providing additional materials such as captions and playlists of said materials. Providing this extra functionality would allow students to personalise the content how they want it, making them become more engaged with the content, and reducing overall cognitive load through the additional flexibility provided.

2.5.5. Student engagement

Another important aspect of education that relates to student individual differences is that of a student's engagement with the course content. One of the more influential work on student engagement what that of Kahu (2013), which identified student engagement as a dynamic construct influenced by institutional practices as well as individual factors. This framework distinguishes engagement from its antecedents (motivation and institutional support) and consequences (student learning outcomes), suggesting that a holistic approach to understanding this concept is the most beneficial.

In the context of digital learning, Kahu, Ella R., Thomas & Heinrich (2022) identify tools such as Discord and Teams address key challenges like isolation and limited interaction. Suggesting that these platforms can foster a sense of belonging and camaraderie among students, which are critical to engagement.

The integration of these frameworks and findings highlights the importance of aligning pedagogical tools and practices with the varied needs of students. This alignment not only improves engagement but also supports students' emotional well-being and academic success, underlining the necessity of embedding a clear, evidence-based understanding of engagement in educational research.

2.5.6. Interdisciplinary differences

Research into the teaching methodology and learning behaviour differences between educational disciplines has a long history. Such early research by Gaff & Wilson (1971), Biglan (1973a, 1973b), and Becher (2001) gathered together an educator consensus of the content of topics, and provided a categorisation of disciplines. This early research helped to identify various dichotomies, for example Biglan (1973a, pp. 201-2) refers to dimensions of paradigmatic sciences (fields with a higher consensus on content and method), and non-paradigmatic sciences (fields without a single paradigm, and a lack of consensus on content and method), and pure or applied fields. As well as noting that the paradigmatic/pure (science disciplines) had been well documented as opposed to other fields such as paradigmatic /applied (technology disciplines), non-paradigmatic/pure (humanities disciplines), and non-paradigmatic/applied (social science disciplines).

These differences can also be noted in the use of LMS environments in E-Learning. In a survey by White & Liccardi (2006) it was found that there were distinct differences in the preferences of students with regards to the integration of the LMS into their courses, and how much of its features were useful to them (such as discussion boards, videos, interactive content, and automated assignments/quizzes).

Paradigmatic/pure discipline students were found to prefer the use of online visualisations but not online assessment. Paradigmatic/applied discipline students however, preferred the online assessment, as well as the online visualisation materials. Non-paradigmatic/applied as well as non-paradigmatic/pure discipline students showed a preference for the use of online discussion boards, as the simulated environments but not for the use of online lectures.

One recent research into student individual differences across disciplines is a study by Davidoff & Jayusi (2024), which involved 980 students from diverse disciplines. The study investigated student social-emotional-psychological (SEP) perceptions, as well as the presence and desirability of 14 teaching-learning-evaluation tools. The study revealed significant disciplinary differences in how students engage with and perceive e-learning tools.

For example, in the Davidoff & Jayusi (2024) study, education students reported the highest satisfaction with social interaction and skill acquisition in online environments, benefiting from interactive and collaborative pedagogical approaches. Whereas, business administration and engineering students faced greater challenges, including diminished social interaction and psychological empowerment, alongside difficulties in acquiring practical, applied skills through e-learning environments.

Additionally, the study highlighted distinct preferences for teaching-learning-evaluation tools among disciplines. Students in paradigmatic/applied fields, such as engineering, expressed a strong demand for simulations and professional scenario-based activities that align closely with their applied nature. Meanwhile, students in non-paradigmatic disciplines, such as the social sciences and humanities, valued tools fostering discussion and collaborative engagement, such as Q&A forums and small group activities. However, across all disciplines, a notable gap was identified between the current availability and the desired use of these tools, suggesting the need for more interactive, practical, and student-centred online learning environments.

This finding aligns with earlier work by Biglan (1973a) and Becher (2001) on the varied paradigmatic and methodological characteristics of disciplines, suggesting that these foundational differences significantly influence the effectiveness and perception of e-learning strategies. For example, paradigmatic/pure disciplines like the sciences emphasised the need for structured content and well-organised materials, while paradigmatic/applied disciplines prioritised tools that facilitate professional readiness.

By tailoring e-learning approaches to these disciplinary differences, educators can address the unique challenges and leverage the strengths of each field, thereby improving both engagement and learning outcomes.

2.5.7. Individual differences summary

When addressing individual differences in learners, it is evident that a holistic approach is critical to effectively personalise E-Learning environments. Each of the discussed dimensions (learning styles, chronotypes, self-efficacy, cognitive load, student engagement, and interdisciplinary differences) offers unique insights that, when combined, can create an inclusive and adaptable learning experience.

While various learning styles can be identified to personalise teaching and learning in E-Learning, it is crucial to recognise the overlapping categories within them. For instance, student preferences for short video content align with Felder & Silverman's (1988) learning style model, categorising them as sensing, active, visual, and sequential. Similarly, Gardner's (1983) theory of multiple intelligences suggests a preference towards spatial and logical-mathematical intelligences. In contrast, Kolb's (1981) learning style inventory categorises them as assimilators. Felder & Silverman's (1988) model proves particularly applicable to E-Learning, given its focus on input, output, and sequencing of materials. This model not only identifies preferences for material types and sequencing but also emphasises the significance of timing in education delivery.

Chronotypes, or time-of-day preferences, further illustrate the importance of aligning educational delivery with students' optimal learning periods, suggesting a potential impact on cognitive efficiency and academic outcomes. Optimal methods are divided into sequential or global, considering factors like time of day and semester (Beşoluk, Önder & Deveci 2011). This implies that the timing of learning events, in addition to material type, can impact student performance.

The importance of self-efficacy is highlighted in studies such as Dash et al. (2022) and Rankapola & Zuva (2023), which link learners' confidence in their abilities to their satisfaction and continued use of E-Learning systems. These findings suggest that fostering self-efficacy through intuitive platforms, clear instructions, and support mechanisms is essential for sustaining engagement and reducing dropout rates.

The cognitive load theory underscores the necessity of managing the volume and complexity of information presented to avoid overwhelming learners, advocating for tailored content that facilitates easier comprehension and retention. Moreover, the academic discipline influences learning preferences, with variances noted between paradigmatic and nonparadigmatic, as well as pure and applied fields. This suggests a need for adaptive E-Learning strategies that accommodate discipline-specific learning trends, potentially indicated by preferences for multimedia resources like videos and discussion forums. Notable differences in preferences, such as for videos and forums, are apparent, but log data might reveal more nuanced preferences.

Student engagement, as conceptualised by Kahu (2013), bridges institutional practices and individual factors, highlighting the interplay of emotional, cognitive, and behavioural dimensions. The use of tools like Discord and Teams helping to foster a sense of community and prevent isolation, while promoting collaboration to improve overall student learning outcomes (Kahu, Ella R., Thomas & Heinrich 2022).

Research into discipline-specific preferences (Biglan 1973a, 1973b; Davidoff & Jayusi 2024) reveals that paradigmatic disciplines (such as sciences) prioritise structured content and simulations, while non-paradigmatic fields (such as humanities) value discussion and collaboration. Understanding these differences allows educators to tailor E-Learning strategies to disciplinary norms, ensuring greater alignment with students' academic and professional needs.

In essence, an effective E-Learning environment must account for individual differences in learning styles, chronotypes, self-efficacy, cognitive load, engagement, and disciplinary differences. By integrating these insights into the design and delivery of online education, educators can significantly enhance learner engagement, performance, and satisfaction, ensuring a more personalised and effective learning experience. This holistic approach not only acknowledges but leverages the diversity of learners to foster an inclusive and adaptive educational landscape.

2.6. Predictive data analytics: Machine learning and data mining in E-Learning

This section explores the dynamic interplay between data analytics, predictive analytics, and machine learning within the realm of E-Learning research. It delves into the distinctions and synergies between data analytics insights into historical data and predictive analytics forward-looking forecasts, highlighted by the positive impact of machine learning technologies into the field of predictive analytics. The section highlights the concept and the application of predictive data analytics in E-Learning, from personalising student learning experiences to enhancing educational outcomes through the careful selection and evaluation of machine learning algorithms. Acknowledging the critical balance between technological advancement and the practical usability of these tools, setting the stage for a detailed examination of how these data-driven approaches can innovate and improve E-Learning strategies.

2.6.1. Data analytics

While not directly linked to E-Learning, data analytics offers a vast array of practical applications for managing large Learning Management System (LMS) datasets. Defined by Runkler (2020, p. 2) as any application of a computer system to a large dataset for the purposes of decision making. Its decision-making aspect is often derived from business intelligence, as discussed by Schniederjans, Schniederjans & Starkey (2014, pp. 1-5), where it is applied in a managerial context to improve strategic outcomes.

Data analytics encompasses a wide range of activities aimed at extracting insights from data. It involves the use of statistical methods, machine learning, and algorithmic approaches to analyse and interpret complex datasets (Provost & Fawcett 2013, p. 20). It primarily deals with historical data, offering descriptive insights into past behaviours, trends, and patterns. However, while it can provide insights into past and current trends, data analytics does not inherently focus on predicting future events or outcomes.

Data analytics can be divided into three general areas; these areas, as discussed by Moubayed et al. (2018), include exploratory data analytics, confirmatory data analytics, and qualitative data analytics. Each area of data analytics has a distinct purpose and set of tools/techniques that are followed. Moubayed et al. (2018) describe the area of exploratory analytics as being separate to machine learning techniques; having the same purpose (identifying patterns in the data), but analysing the data to come up with a model, rather than finding a model and then analysing it's parameters. Tools described for this area of data analytics involve graphical elements such as histograms, and quantitative methods such as confidence intervals, and measures of variance. Moubayed et al. (2018) describe the area of confirmatory data analytics as techniques that are used to confirm prior hypothesis of the data, with tests such as Analysis of Variance (ANOVA), and Chi-square test for variance. Confirmatory tests can be utilised in conjunction with exploratory data analysis methods to provide a robust analysis of key hypothesis' held of the data in question. Finally, Moubayed et al. (2018) describe the area of qualitative data analytics; Involving the analysis of generally non-numeric, descriptive data, that can be in the form of multimedia, interviews with students, and other data that is not able to be quantitatively analysed.

Examples of data analytics in E-Learning

Learning Analytics

The purpose of Learning Analytics is described by Doug Clow (2013) as enabling both students and teachers to benefit from access to a large datasets of LMS logs and other datasets of student related LMS interactions. Learning Analytics is one of the most basic usages of data analytics in E-Learning, and benefits from computer based algorithmic methods of data analysis (Lang et al. 2017). This analysis of student engagement data within the LMS allows researchers to identify patterns in course access, learning resource usage, and discussion forum participation. This can be used to help both researchers and educators understand how students interact with what materials and to better identify areas where students may need additional support or what materials are being used the most. The importance of LMS analytics data for educational research is highlighted by Gašević, Dawson & Siemens (2015), suggesting that learning analytics should be further integrated into existing educational research, and how it can help to improve teaching practice as a whole.

Student Performance Dashboards

Integrating further with the LMS, 'Dashboards', are interfaces that allow for the administrators, educators, and in some circumstances users of the LMS to aggregate historical academic performance data. This can help to provide educators with insights into student progress, and potentially highlight trends in grades, assignment completion rates, and exam scores, enabling targeted interventions for students at risk. Common measures of student engagement as discussed by Henrie, Halverson & Graham (2015), have been focused on self-reported surveys and interviews by students, as well as to assessment scores and behaviour counts from LMS data. An automated, LMS data analysis approach would not only be the simplest (not requiring direct student surveys) but would potentially provide quicker feedback.

Content Engagement Analysis

Expanding on general LMS analytics, content engagement analysis allows for the examination of how students engage with online learning content in much greater, and more granular detail. This can include time spent on specific LMS pages (or on the LMS as a whole), views of specific content types (such as videos or assignments), and interaction patterns with interactive elements such as quizzes. Insights from this analysis can guide the development of more engaging and effective educational content.

This is highlighted in research by Arnold & Pistilli (2012), that utilises real-time data analytics to provide detailed interaction information, as well as current performance levels to student and educators (utilising predictive analytics, to predict future performance). Arnold & Pistilli (2012) also note that early feedback of learner analytics can allow for actionable information to educators at a much faster rate, and was found to be very helpful overall.

Course Recommendation Systems

At a more administrative level, historical LMS data may be beneficial for future students, with information on past student course selections and outcomes, allowing the system to recommend courses to students based on their interests, academic history, and career goals. This can be used to better help students customise their learning journey to their personal and professional aspirations. An example of such a system is discussed by Thai-Nghe et al. (2010), who identify a novel recommender system that can be additionally used to predict student performance.

2.6.2. Predictive analytics

Predictive analytics, applied across various domains, offers significant opportunities to advance E-Learning literature and explore disciplinary nuances. Utilising historical data, predictive analytics forecasts future events through advanced statistical techniques and machine learning models (Eckerson 2007, pp. 4-8). The field has evolved from data mining to incorporate machine learning, enhancing its predictive capabilities.

In E-Learning, predictive analytics applies statistical algorithms, machine learning techniques, and data mining principles to analyse educational data, aiming to predict student performance, learning outcomes, and tailor content to individual needs (Baker & Inventado 2014, pp. 63-9). This approach facilitates the development of adaptive learning systems that customise content and teaching strategies based on student behaviour and preferences. However, while predictive analytics is forward-looking, focusing on future trends, it may not provide immediate insights for present decisions without integrating these predictions into a broader analytical framework.

Additionally, there are several distinct sub-groups of predictive analytics, depending on the nature of the objectives, the stakeholders involved, and the methodology used (Ranjeeth, Latchoumi & Paul 2020). These sub-groups, as described by Ranjeeth, Latchoumi & Paul (2020), are academic analytics, education data mining, and learning analytics, each having unique stakeholders and objectives.

Academic analytics is primarily used by educational institutions for purposes such as enrolment prediction, marketing, and decision-making. Education data mining is mainly utilised by teachers and students to enhance learning processes. Finally, learning analytics is employed by teachers, students, and educational institutions for prediction, recommendations (such as course/content recommendations), admissions, marketing, and customisation. Another key differentiator among these three sub-groups is their methodology: Academic analytics predominantly employs statistical methods, education data mining uses clustering, association, and classification, while learning analytics utilises quantitative methods in addition to clustering, association, and classification.

By integrating machine learning, predictive analytics transcends data analysis to enable adaptive learning environments that can personalise education at scale, by leveraging patterns identified through a variety of data mining and learning analytics techniques (Romero & Ventura 2020), to forecast individual student performance, optimise learning paths, and enhance educational decision-making processes. This approach not only facilitates a more responsive and customised learning experience but also contributes to the broader academic discourse by offering insights into effective teaching strategies and learning processes (Dietz-Uhler & Hurn 2013).

Research into predictive analytics over the past decade has involved a variety of different topics of interest. In a systematic review of predictive learning analytics by Sghir, Adadi & Lahmer (2023), documenting research from 2012 to 2022, and collecting 74 research papers in predictive learning analytics, and identifying five key areas of research; Student enrolment, student performance, identifying at-risk students, student engagement, and student satisfaction with the LMS.

Examples of predictive analytics in E-Learning

Predictive Modelling of Student Performance

One of the more important applications for predictive analytics, as mentioned by Sghir, Adadi & Lahmer (2023), is the prediction of student educational outcomes. This can be achieved through employing machine learning algorithms on historical LMS data, such as student behaviours and a variety of LMS engagement metrics to predict future student performance (Qiu, Zhang, et al. 2022). Performance can be evaluated in a number of ways; monitoring of continual academic performance, overall achievement at an individual level, or overall achievement of classes or cohorts of students (Sghir, Adadi & Lahmer 2023). More recent approaches include utilising Deep Learning Models and basic sets of demographic data, academic records, and LMS interaction information, was shown to have high levels of accuracy in student performance prediction (Fazil, Rísquez & Halpin 2024).

Adaptive Learning Pathways

Another example is by utilising predictive analytics to tailor the learning experience to individual student needs, which would be considered as part of the enrolment area of research into predictive analytics, mentioned by Sghir, Adadi & Lahmer (2023). By analysing past performance and learning behaviours, adaptive learning systems can modify content delivery, suggest additional resources, or adjust difficulty levels in real-time. An example of such system is the KT-IDEM (Knowledge Tracing Item Difficulty Effect Model) outlined by Pardos & Heffernan (2011), that incorporates difficulty metrics, which can be adjusted based on the student's performance.

Early Warning Systems

In a similar fashion to predicting student performance, and detecting the levels of difficulty, there is an opportunity to use LMS data, and predictions made to help identify at-risk (potentially failing, or not engaging with the LMS). This is another area of concentrated research into predictive analytics, as mentioned by Sghir, Adadi & Lahmer (2023). This is discussed from the point of view of helping to address rates of retention and to help boost the levels of student commitment to their education.

This can be achieved through implementing predictive models to identify students from usage data, regarding engagement, current (or predicted) grades, or likelihood of dropout. Research by Jayaprakash et al. (2014) showed in mixed results in implementing such features, but that the systems were possible, and did show some benefits.

Learning Outcome Forecasting

Primarily related to the enrolment aspect of predictive analytics research mentioned by Sghir, Adadi & Lahmer (2023); Involving analysing historical data to predict learning outcomes for courses or programs. This can help institutions in curriculum planning, resource allocation, and setting realistic expectations for student success rates. As shown in research by Marbouti, Diefes-Dux & Madhavan (2016), which made student performance data available to course instructors during the semester, for the purposed of creating course-specific performance forecasting, which was shown to help these students improve their performance.

Learning Analytics Dashboards

Another method of integrating predictive analytics into LMS implementations is the use of Learning Analytics Dashboards. This is outlined by Ramaswami, Susnjak & Mathrani (2023), and describes their implementation the 'SensEnablr', which displayed the student risk profile (probability of non-completion in a course), and it was found that 87% of students responding, described 'SensEnablr' as positively impacting their attitude towards their study. This direct method of displaying predictive analytics results to students is more student focused and would be additional to having outcome forecasting or early warning prediction information made available to educators and administrative staff.

2.6.3. Predictive data analytics

Data analytics aims to uncover insights from past and present data, focusing on what has happened and why. Predictive analytics, however, looks forward to what might happen in the future, using historical data to make informed predictions. Employing a variety of statistical and analytical techniques to process and analyse data, predictive analytics specifically leverages statistical models and machine learning algorithms to predict future events. In E-Learning, a combination of data analytics, and predictive analytics might be used to understand both the past student performance trends and engagement levels, as well as forecasting future student performances or identifying potential dropouts, enabling personalised learning experiences.

While data analytics alone provides comprehensive insights into existing datasets, it does not inherently predict future trends. On the other hand, predictive analytics by itself, focuses on forecasting and might not provide the depth of analysis on past data that data analytics offers.

The term predictive data analytics is defined by Kelleher, Mac Namee & D'arcy (2015, p. 1) as the utilisation of past data to identify patterns that can be used to create models to predict future outcomes. The potential is great for this combination of data analytics and predictive analytics (into predictive data analytics), to play crucial roles in extracting value from E-Learning data.

With each field addressing different aspects of data analysis. Data analytics provides a solid foundation for understanding and interpreting data, which is essential for any predictive analysis. Predictive analytics, leveraging insights gained from data analytics, extends the capability to forecast future scenarios, especially critical in dynamic sectors like E-Learning for personalising and optimising learning paths. By understanding the distinctions and complementary nature of these fields, organisations and educational institutions can better leverage data to inform decisions and strategies.

2.6.4. Machine learning

The potential of predictive analytics has existed and evolved over half a century, with roots in data mining and then machine learning, with early research resulting from attempting to replicate human biological processes to simulate learning. The earliest of these research areas began in the 1950's, with the concept and technical feasibility of a 'Perceptron' machine, developed by Frank Rosenblatt (1957). This work built off mathematical concepts presented by McCulloch & Pitts (1943) regarding nervous system activity. This would represent the first implementation of machine learning and become the predecessor to modern Artificial Neural Networks (ANNs) such as the Multi-Layer Perceptron (MLP) classifier (Rumelhart, Hinton & Williams 1986).

In addition to statistical and neural network machine learning, there are generally two other branches of machine learning, namely association rule mining, such as Agrawal & Srikant's (1994) Apriori algorithm, and decision trees such as Quinlan's (1993) C4.5 algorithm (which itself is an extension of Quinlan's (1986) ID3 Algorithm), and simplified versions of Classification and Regression Trees by Breiman, Friedman, Stone, et al. (1984).

Finally, there is ensemble machine learning, which is used to enhance the predictive performance of machine learning algorithms and mitigate against overfitting. These machine learning algorithms generally fall into one of two categories: boosting, which involves combining different models, and bagging, which involves aggregating the results of multiple models. The first example of a boosting algorithm was Freund & Schapire's (1996) AdaBoost. This algorithm involved combining multiple weaker performing algorithms to create a strong performing algorithm. The first example of a bagging algorithm was Breiman's (1996) Bagging Predictor.

This algorithm involved using multiple versions of the same algorithm across various subsets of the original dataset, then aggregating their output together to improve overall performance. This was then iterated upon by Breiman (2001), to develop Random Forests, which utilised multiple decision trees to compete with similar boosting or bagging techniques, but not requiring the progressive change of the training set.

2.6.5. E-Learning applications

Applying the concepts of predictive data analytics into E-Learning involves utilising the most appropriate tools from the realm of machine learning. This section will discuss the techniques and algorithms in machine learning that are applicable for E-learning, and for the use in this research particularly.

Modern uses of machine learning in E-Learning utilise a wide variety of techniques, including algorithms from all machine learning branches mentioned above, creating many potential tools to draw from for any educational data mining or learning analytics work. Therefore, it is important to identify which algorithms are best suited to the task, and what is commonly used. In a survey by Yuniarti, Winarko & Musdholifah (2020), of E-Learning research papers relating to data mining and student assessment in E-Learning (published from 2016 to 2020), Yuniarti, Winarko & Musdholifah (2020) identified 12 primary data mining methods utilised: J.48 (a Java implementation in WEKA by Frank, Hall & Witten (2016) of Quinlan's (1993) C4.5 algorithm), naïve bayes (John & Langley 1995), random forest (Breiman 2001), logistic regression (Cox 1958), K-nearest neighbour (KNN) (Fix & Hodges 1989), association rule mining (Agrawal & Srikant 1994), linear regression (Pearson 1901b), artificial neural networks (ANN) (Krizhevsky, Sutskever & Hinton 2012), multi-layer perceptron (Rumelhart, Hinton & Williams 1986), support vector machine (Vapnik 2000), K-means (MacQueen 1967), and deep long short-term memory (LSTM) (Pearson 1901b).

Algorithms chosen

Table 1 illustrates the machine learning algorithms selected for the project. These include several algorithms mentioned by Yuniarti, Winarko & Musdholifah (2020), as well as some that were not. The following section outlines the rationale for their selection.

Table 1 - Machine learning algorithms chosen

Algorithms	**Ensemble algorithms**
RepTree	RandomForest
DecisionStump	RotationForest
RandomTree	NBTree
J48	AdaBoostM1
NaiveBayes	
simpleCART	

Tree algorithms

The first of the tree-based (non-ensemble) algorithms chosen is the DecisionStump. This algorithm constructs a one-level decision tree, consisting only of one internal node connected to the terminal nodes (Iba & Langley 1992). DecisionStump is also a weak learner compared to the rest of the algorithms and is commonly used in conjunction with ensemble methods, providing a good baseline for comparing the other algorithms.

The next two are standard tree-like algorithms: J48 and simpleCART by Breiman et al. (1984). The J48 algorithm, an open-source Java implementation of Quinlan's (1993) C4.5 algorithm, boasts a relatively high complexity level. It can handle missing values and incorporates functionality for post-pruning to prevent overfitting (Mark Hall et al. 2009; Quinlan, R 1993). The simpleCART algorithm, also capable of managing missing data and equipped with pruning features to inhibit overfitting, is comparably complex and computationally demanding.

The final two tree-based (non-ensemble) algorithms are REPTree and RandomTree. REPTree is a comparatively rapid algorithm Implemented in WEKA by Frank, Hall & Witten (2016), that offers enhanced performance optimisation while including pruning (specifically, reduced-error pruning with backfitting) to mitigate overfitting. Conversely, RandomTree developed by Breiman (2001) considers a random subset of features per split, enhancing its robustness against noise in the data and reducing its susceptibility to overfitting.

Tree-based ensemble algorithms

Ensemble algorithms were incorporated to assess the performance benefits and complexity increments associated with such methods, although they are not utilised in the analysis presented in this thesis, due to their challenging interpretative nature.

RandomForest developed by Breiman (2001) represents the 'bagging' style ensemble algorithms. It constructs multiple decision trees during training and can reduce overfitting and enhance accuracy by averaging the results of the multiple trees generated.

Conversely, the AdaBoostM1 algorithm developed by Freund & Schapire (1997), representative of the 'boosting' style algorithms, employs multiple weaker learning algorithms (such as DecisionStump or REPTree) to forge a strong learner. It focuses on prior misclassifications and adjusts weights to concentrate on more challenging cases.

The final ensemble method is RotationForest, developed by Rodriguez, Kuncheva & Alonso (2006), which leverages Principal Component Analysis (PCA). This method is even more computationally intensive than RandomForest but maintains accuracy on training data while augmenting individual tree diversity, enhancing its generalisation capabilities.

Additional algorithms

Two supplementary algorithms were incorporated for comparison, both involving the use of the Naïve Bayes classifier. The NaiveBayes algorithm, as detailed by Yuniarti, Winarko & Musdholifah (2020) is significantly different from the tree-based algorithms, and are challenging to interpret, but it will serve as a useful benchmark against the other algorithms. Similarly, the ensemble NBTree algorithm developed by Kohavi (1996) was selected as it harnesses both Naive Bayes and the structure of Decision Trees to enhance performance, enabling comparison with other ensemble methods.

For investigative purposes, it's noted that Neural Networks, Support Vector Machines (both not selected), and ensemble methods are significantly more complex to interpret. Therefore, their utilisation is only justified when they offer substantial advantages over simpler methods, such as predictive accuracy.

2.6.6. Algorithm performance evaluation

A key factor to consider is the effectiveness and reliability of different algorithms compared to others. One characteristic of E-Learning data is the potential for imbalanced datasets. For example, student performance data might be skewed, with few students falling into certain grade ranges, such as Fail grades or High Distinctions.

These categories represent minority classes. As He & Garcia (2009) highlight, it's crucial to appropriately handle these minority classes without adversely impacting the classification of majority classes. While not as serious as the classification of cancerous cells mentioned by He & Garcia (2009), it would defeat the purpose to ignore small classes of students that highlight specific cases of student drop-out.

Traditionally, model accuracy is considered, calculated as the sum of true positives and true negatives divided by the total number of observations, along with error rate (which is one minus the accuracy). However, Chicco & Jurman (2020) note that these measures can be misleading in unbalanced datasets due to their sensitivity to data distributions, with metrics such as F-Measure (discussed below) or Accuracy often inflating results.

The next two possibilities are that of precision and recall. Precision measures the accuracy of positive predictions and is defined as the proportion of true positives out of all predicted positives (true positives plus false positives). It indicates how many of the instances the model classified as positive are positive. Recall (or sensitivity) assesses the identification of true positives, which is the ratio of true positives to all actual positives (true positives plus false negatives), indicating the model's ability to find all relevant cases in the dataset. While each metric offers valuable insights, they have limitations when considered in isolation. It's important to note that neither precision nor recall, and thus their arithmetic, geometric, or harmonic means such as the F-measure, consider both false negatives and false positives, meaning the F-measure doesn't take into account all aspects of the confusion matrix and can be misleading in unbalanced datasets (Chicco & Jurman 2020; He & Garcia 2009; Powers 2015).

To address these limitations, Informedness (recall plus inverse recall minus one) and markedness (precision plus inverse precision minus one) for the dichotomous case, as defined by Powers (2003, 2011), could be used instead to evaluate the effectiveness of data mining algorithms. These metrics consider both directions of information flow, providing a more comprehensive evaluation.

Another metric that would be useful for measuring algorithm performance/effectiveness is Cohen's (1960) Kappa statistic. Cohen's Kappa is particularly valuable in this research due to its effectiveness in contexts with imbalanced datasets. It measures the extent of agreement between the model's predictions and the actual classifications, correcting for the agreement that could happen by chance. However, it is suggested by Powers (2012a) that the Kappa statistic is not as robust in dealing with bias, and instead Matthew's (1975) Correlation Coefficient is suggested as a better-suited metric.

The Matthews Correlation Coefficient (MCC) is a more comprehensive metric, offering a single correlation coefficient based on all four values in the confusion matrix (true positives, true negatives, false positives, and false negatives). It evaluates the quality of binary classifications, providing a balanced measure that is particularly useful in the presence of class imbalance (Chicco & Jurman 2020). The MCC's strength lies in its ability to encompass all aspects of the confusion matrix, serving as a trustworthy indicator of a model's overall performance. Additionally, MCC represents the geometric mean of Informedness and markedness in situations of dichotomous classification, a generalisation that encapsulates both the Matthews and the Pearson definitions of correlation. This combination enhances its utility, providing a nuanced evaluation that reflects both the model's ability to correctly identify each class (Informedness) and its precision and reliability in making these classifications (markedness).

The final metrics that will be mentioned is that of the Receiver Operator Characteristic (ROC) and the Area Under the Receiver Operator Characteristic (AU ROC). The concept of ROC was developed in World War 2, for radar engineering, however, the concept was defined by Hanley & McNeil (1982) for use in radiology. Calculating the ROC involves calculating the true positive rate (defined as the number of true positives divided by the sum of true positives and false negatives), and the false positive rate (defined as the number of false positives divided by the sum of false positives and true negatives). The AU ROC involves identifying the area under the plotted x and y coordinates of true positive rate and false positive rate respectively, against a set number of thresholds. The AU ROC is beneficial as it enables users to assess trade-offs between the benefits and costs of classification regarding various data distributions (He & Garcia 2009). Additionally, as Powers (2011, 2012b) explains, the vertical distance of the ROC operating point above the chance line represents Informedness, a metric that quantifies how well a classifier distinguishes between classes. The area under the triangular single operating point curve (connecting a specific operating point to (0,0) and (1,1)) represents Certainty (also referred to as Balanced Accuracy), which can be calculated as half the Informedness value plus 0.5. This calculation accounts for the area below the chance line and ensures that Certainty is typically within the range of 0.5 to 1 for a functional classifier.

The area under the ROC curve formed by multiple operating points, or the convex hull (ROCCH), represents the Area Under the Curve (AUC). This value combines both the Certainty and Consistency of the classifier. Consistency is shown in the area between the multipoint ROC curve (or ROCCH) and the triangular single operating point curve. Consistency measures the robustness of the classifier across varying thresholds or differing data prevalence rates, ensuring that the classifier's performance remains reliable despite changes in operational conditions.

It's important to note that any areas below the reference curves (either the chance line or the single operating point curve) contribute negatively. The convex hull approach removes such concavities, as they represent suboptimal threshold points. By excluding these, the convex hull ensures that the ROC curve reflects only optimal thresholds, providing a more accurate and actionable representation of classifier performance.

Performance metric weighting by class attribute

In the context of unbalanced datasets common in E-Learning, where certain classes (Fail or High Distinction) may be underrepresented, it's essential to ensure that the evaluation metrics chosen are fair and do not skew towards the majority class. Therefore, weighting performance metrics by the chosen class attribute will be necessary to deal with this challenge and ensuring a fair and informative evaluation of the chosen algorithms.

In WEKA version 3.8.6 (Frank, Hall & Witten 2016), the software enables the computation of weighted average metrics for the selected performance measures; F-measure, Cohen’s Kappa, Matthews Correlation Coefficient (MCC), and the Area Under the Receiver Operating Characteristic (AU ROC) curve. For the F-measure, WEKA calculates the harmonic mean of precision and recall for each class, and weights this proportionally to the number of instances in each class within the dataset. This approach adjusts the contribution of each class based on its prevalence, providing a better view of the model's performance across classes of varying sizes, which is common in E-Learning data.

Cohen’s Kappa and the Matthews Correlation Coefficient are also adjusted to reflect weighted contributions, by first calculating these metrics for each class individually and then computing an average weighted by the class sizes. This method helps to account for the different impacts of class sizes on the model's predictive accuracy and the agreement beyond chance.

For the AU ROC metric, WEKA calculates the area under the receiver operating characteristic curve for each class. It then computes a weighted average of these areas, again based on the proportion of instances for each class in the dataset. This ensures that the final AU ROC metric reflects the model’s ability to discriminate between classes, weighted by the significance of each class in the overall dataset.

Through these weighted calculations, WEKA provides a more balanced assessment of model performance, particularly in datasets with imbalanced class distributions. This approach allows for better investigation into the effectiveness of models more accurately, considering both the model's precision and its ability to handle classes of differing sizes.

Performance metric summary

For the purposes of this project several metrics were chosen to best judge effectiveness, rather than relying on any one metric which may have its own bias, or unsuitability:

- Model Accuracy.
- Weighted average Cohen’s Kappa.
- Weighted average Matthews Correlation.
- Weighted average AU ROC.

With regards to E-Learning, an algorithm with Accuracy suggests that it is effective at identifying relevant learning patterns while minimising irrelevant ones. However, as discussed, this alone does not necessarily indicate that the model is proficient in differentiating between various learning outcomes. Additionally, given the nature of this dataset (an imbalanced dataset), a high Cohen’s Kappa score would potentially indicate that the algorithm demonstrates a high level of agreement between its predictions and actual learning patterns, accounting for agreement that might occur by chance.

Next the Matthews Correlation Coefficient is also particularly useful in the presence of class imbalance; a high score suggests that the algorithm effectively identifies learning outcomes for both majority and minority classes. Finally, a high Area Under the Receiver Operating Characteristic (AU ROC) value suggests that the algorithm is capable of effectively differentiating between positive and negative learning outcomes across a range of decision thresholds.

2.6.7. Significance of results

Identifying a high-performing algorithm is a crucial first step, but it's equally important to assess how reliably the algorithm performs across different data subsets. Additionally, the validity of the results in repeated tests is essential. This assessment helps determine the algorithm's generalisability, a key factor given the vast and diverse data typically encountered in E-Learning. An algorithm that performs well only on a specific dataset or is overly tailored to data characteristics may not be useful in broader applications. Therefore, ensuring that an algorithm is not just effective, but also reliable, valid, and generalisable, is vital for its successful application in E-Learning environments.

To this end, utilising the repeated cross-validation t-test by Bouckaert & Frank (2004), will be beneficial. This approach applies the Gosset's Student (1908) t-test across repeated training and testing subsets of the data, helping to ascertain the model's performance across multiple variations of the underlying data. Such a method is important in reducing the risk of overfitting the model to a specific dataset. While it may sound good that a model is a perfect fit for the data, overfitting is a genuine problem in machine learning, as discussed by Domingos (2012). In the context of E-Learning, this issue might manifest in a model that performs exceptionally well on test student data but fails to deliver similar results when applied to a different university, or even a different year at the same university.

2.6.8. Usability, understandability, and visualisation

One of the most important features of any algorithm chosen for this project, along with accuracy, reliability, and significance, is its interpretability and suitability for presentation to both educators and data scientists.

Therefore, it's necessary to identify a machine learning algorithm that balances performance with usability in a simpler context. In a study by Matzavela & Alepis (2021), on M-Learning applications, decision tree algorithms were found to be particularly useful in an educational context. This is attributed to their simplicity and easily understandable representation. While their study wasn't focused on E-Learning applications, the principles are still applicable to this project.

Additionally, Souza et al. (2022), note the potential to enhance interpretability by limiting a decision tree's features, such as reducing its depth and complexity. This approach makes models more explainable and accessible to users without extensive backgrounds in machine learning.

The decision tree methodology was selected for its intuitive and visually accessible approach, facilitating an easier understanding and identification of associations and rules within E-Learning data. Additionally, various ensemble methods were incorporated. Despite their inherent complexity, these methods were chosen to leverage performance-boosting techniques such as bagging and boosting algorithms. This approach aims to explore potential enhancements in predictive performance, rather than for interpretative purposes. This dual approach strikes a balance between the interpretability of decision trees and the advanced predictive capabilities of ensemble methods, aiming to maximise the efficacy of our E-Learning data analysis.

Decision tree methodologies

Among the visually interpretable algorithms selected for the study (REPTree, Decision Stump, Random Tree, J48, and simpleCART), each has its own methodology for creating decision trees, a process also known as Decision Tree Induction. However, they all share common processes, as described by Han, Kamber & Pei (2012, p. 275): Each consists of a list of attributes, an attribute selection method, and a series of data to build the tree from.

Most of the differences between the algorithms mentioned revolve around attribute selection and decisions on how to split the data at various decision points. A primary distinction among these algorithms is the choice to use either information gain/variance reduction, as used by Quinlan (1993) for C4.5, or Gini impurity, as used by Breiman, Friedman, Stone, et al. (1984) for CART and simpleCART. Information gain is described in simple terms by Han, Kamber & Pei (2012, p. 280) as measuring how much information is gained by splitting on a particular attribute, compared to the information gained from just the proportion of classes in the dataset. On the other hand, the Gini index is described by Han, Kamber & Pei (2012, p. 283), as a measure of the impurity of the splits within the dataset, with more homogeneous observations in a split resulting in 'purer' splits.

For example, REPTree uses information gain for its splits and employs reduced-error pruning to avoid overfitting. Decision Stump simplifies the process to a single attribute split, making it a fast and straightforward method, often used as a weak learner in ensemble techniques. Random Tree introduces randomness by selecting a subset of attributes at each decision point, which helps in creating diverse trees when used in ensemble models like Random Forests. J48, an implementation of C4.5, uses the gain ratio for attribute selection, offering a balance between the number of splits and the information gain, and includes mechanisms for dealing with missing values and pruning to improve generalisation. Lastly, simpleCART makes binary splits based on Gini impurity, focusing on simplicity and computational efficiency, while still allowing for complex decision boundaries through recursive binary splitting.

For the evaluation of decision trees in this study, the primary aspects of the trees that will be discussed are the components of the trees themselves, the decision nodes, the leaf nodes, and the branches, as shown in Figure 11.

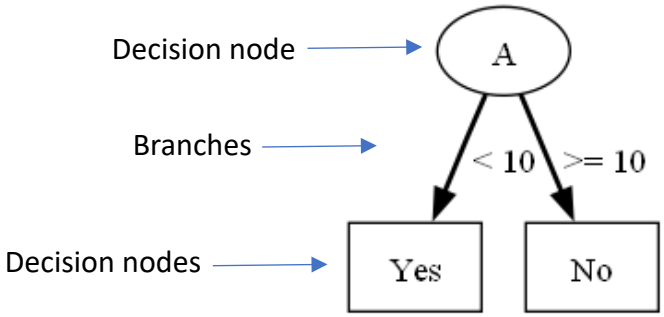

Figure 11 - Example decision tree

Decision nodes

Decision nodes serve as critical junctures within a decision tree where the data is partitioned using specific attributes, informed by methodologies such as information gain or the Gini index (Han, Kamber & Pei 2012). The selection of attributes at these nodes is governed by heuristics designed to maximise the purity of the subsequent partitions. For instance, the information gain heuristic evaluates how much uncertainty in the class distribution is reduced after a split, while the Gini index measures the homogeneity of the partitions created (Breiman et al. 1984; Quinlan, R 1993). The type of data (numerical vs. categorical) also influences the decision-making process, with different strategies employed to handle each. Numerical attributes may involve creating splits based on thresholds, whereas categorical attributes could lead to partitions based on category membership. Additionally, the handling of missing values at decision nodes is an important aspect, with techniques such as surrogate splitting being utilised to maintain the integrity of the tree structure when data is incomplete (Quinlan, R 1993). As shown in Figure 11, 'A' represents a decision point, indicating the attribute from the dataset that has been selected. In the context of E-Learning data for this research, this might be 'Videos Viewed' or other similar attributes.

Branches

Branches in a decision tree represent the outcomes of splits made at decision nodes, typically reflecting binary choices based on the evaluated attributes (Han, Kamber & Pei 2012). The determination of these splits is crucial, as it directly impacts the tree's complexity and its ability to generalise. Criteria for splitting can include maximising information gain or minimising Gini impurity, depending on the chosen algorithm. The complexity introduced by the branches is a double-edged sword; while it can lead to more accurate models, it also risks overfitting. Therefore, strategies like pruning are employed to trim branches that contribute little to the model's predictive power, thereby enhancing its generalisability (Breiman et al. 1984). Examples of thresholds and conditions used for branching should illustrate the practical application of these criteria, highlighting how decisions at branches navigate the trade-off between model complexity and interpretability.

In Figure 11, this is illustrated by labels on each arrow (e.g., "less than 10" or "greater than or equal to 10"). This split would be determined by the previously mentioned information gain or Gini index methodology, depending on the algorithm used.

Leaf nodes

Leaf nodes represent the conclusion of the decision-making process in a decision tree, where each node assigns a class label to the instances that reach it. The assignment is typically based on the majority class among the instances in the leaf, although probabilistic approaches can also be employed to account for uncertainty (Han, Kamber & Pei 2012). While the research primarily focuses on classification, it's worth noting that in regression trees, leaf nodes predict numerical values rather than classes. Pruning strategies are particularly relevant to leaf nodes, with both pre-pruning (stopping the tree growth early) and post-pruning (removing non-essential nodes after the tree has been fully developed) approaches being utilised to prevent overfitting and ensure that the tree remains as simple as possible without sacrificing accuracy (Quinlan, R 1993). These nodes represent the classes in the classification process and may require multiple levels of decision points to achieve a leaf node, as opposed to the simplified one-level structure shown in Figure 11. The results presented in this research will analyse both full decision trees, as well as ‘paths’ through the decision trees.

These paths will encompass relevant decision nodes, leaf nodes representing different outcomes, and the branches to and from the decision nodes and leaf nodes.

2.6.9. Predictive data analytics summary

The journey of predictive data analytics within the sphere of E-Learning highlights the significant advances made in machine learning and data mining over the decades. From the inception of the Perceptron in the 1950s to the advent of modern machine learning ensemble techniques, this section has explored the evolution of technologies that now enable the personalisation and optimisation of E-Learning paths at a much larger scale. Key to this exploration has been the distinction and intersection between data analytics and predictive analytics. Data analytics, with its roots in analysing historical data to provide insights into past behaviours, trends, and patterns, sets the foundation for predictive analytics. The latter extends this analysis into the future, employing statistical models and machine learning algorithms to forecast potential outcomes and trends.

In the realm of E-Learning, the application of predictive data analytics encompasses a broad spectrum of practices. From predicting student performance and engagement to tailoring educational content to meet individual needs, predictive analytics offers a forward-looking approach that complements the descriptive insights provided by data analytics. This synergy between predictive and descriptive analytics is crucial for developing adaptive learning systems that can respond dynamically to the needs of students, thereby enhancing the educational experience and outcomes.

The selection of appropriate machine learning algorithms is pivotal to the success of these endeavours. This chapter has delved into the various algorithms at our disposal, including decision trees, ensemble methods, and neural networks, each with its own strengths and limitations. The evaluation of these algorithms, through metrics such as Accuracy, Cohen's Kappa, Matthews Correlation Coefficient, and AU ROC, provides a framework for assessing their effectiveness in addressing the unique challenges presented by E-Learning data. These challenges often include imbalanced datasets and the need for algorithms that can perform reliably across diverse data subsets.

Additionally, the significance of results extends beyond algorithmic performance, encompassing the reliability, validity, and generalisability of the algorithms in real-world E-Learning environments. This involves assessing how well these algorithms can adapt to different data characteristics and maintain their performance across various educational contexts. The goal is to identify high-performing algorithms that are not only effective but also robust and versatile.

However, technological efficacy is only one aspect, the usability, understandability, and visualisation of these predictive models are equally important. Decision tree methodologies have been highlighted for their intuitive representation, making them accessible to educators and data scientists alike. The ability to visually interpret and understand the decision-making process of these models is essential for their practical application in E-Learning. It allows educators to make informed decisions and to tailor educational strategies effectively to the needs of their students.

In conclusion, the integration of predictive data analytics into E-Learning represents a combination of data analysis, machine learning, and educational theory. By harnessing the power of these technologies, educators and researchers can unlock new potentials in personalised learning, student performance prediction, and the overall enhancement of the educational experience. The careful selection, evaluation, and application of predictive models, grounded in an understanding of their usability and interpretability, will continue to play a critical role in shaping the future of E-Learning.

2.7. Chapter summary

The origins of E-Learning can be traced back to correspondence courses during the industrial revolution, evolving through radio and TV, and later computer-based platforms in the 1960s. Since then, significant milestones in development have included the PLATO system and Project Athena, leading to the emergence of modern Learning Management Systems (LMS) like Blackboard and Moodle, which are the primary focus of this thesis. However, the diverse definitions of E-Learning emphasise the lack of a standardised definition and the overlap between terms like distance learning, online learning, and E-Learning. It discusses various E-Learning categories, such as Course Management Systems (CMS), LMS, and Knowledge Management Systems (KMS), each serving different educational functions.

More important than the definition of an LMS is how it is used as an educational tool. Therefore, it is important to consider the student perspective, which include individual differences in learners, such as learning styles, chronotypes, self-efficacy, and cognitive load, which are examined for their impact on E-Learning. This underscores the importance of customising E-Learning to accommodate these differences for effective education delivery. Identifying what and how to personalise in E-Learning can be done using machine learning and data mining. This thesis has reviewed various algorithms and their effectiveness in predicting E-Learning success, emphasising the importance of choosing algorithms that balance accuracy with usability and interpretability for educators and data scientists, which is why the current decision tree focus has been chosen.

In the realm of instructional design, the chapter has underscored the pivotal role of models such as ADDIE in guiding the structuring and delivery of E-Learning content. The historical relevance, systematic methodology, and adaptability of the ADDIE model make it an invaluable tool in the instructional designer's arsenal, aptly catering to diverse learning needs and contexts. It elucidates how instructional design models, especially ADDIE, provide a structured yet flexible framework, enabling educators to craft learning experiences that are not only systematic and coherent but also responsive to the unique dynamics of various educational domains.

In summary, the literature offers an in-depth analysis of the evolution, current practices, pedagogical approaches, individual learner differences, and technological advancements in E-Learning. It highlights the field's complexity and points out that there are further gaps in research to explore, particularly in how predictive analytics can be useful in E-Learning.

3. Methods

3.1. Chapter overview

The aim of this chapter is to provide a comprehensive overview of the research methodologies utilised in this study. This chapter will outline the systematic approach adopted for data collection, preprocessing, exploratory analysis, and subsequent experiments. It will detail the hardware and software environments utilised, the specific methods for data handling and analysis, and the rationale behind the selection of statistical tests and machine learning models. This chapter will provide a clear understanding of the procedural steps, analytical strategies, and the methodological rigour that underpins the research findings presented in subsequent chapters.

The experimental stages are broken down into six main categories, as shown in Figure 12, these stages include:

Data preprocessing and transformation – The initial stage where the dataset is extracted from the base data files, any irregularities are addressed and is processed into the final form necessary for subsequent stages.

Exploratory analysis – Before the main series of machine learning experiments, the dataset is evaluated with statistical methods to identify any patterns that may be further explored via machine learning.

Attribute reduction – Given the size of the dataset, attribute reduction will be attempted to ensure that the chosen number of attributes represents the dataset efficiently. Principal Component Analysis will be used for this step and will contribute to the analysis of the dataset from any patterns found.

Experiment 1: Predicting student grade – The first machine learning experiment will run all selected algorithms to classify by grade, evaluate their performance, and assess chosen decision trees for rule extraction.

Experiment 2: Predicting college affiliations – The second machine learning experiment will perform the same steps as experiment 1 but classify based on the college attribute.

Experiment 3: College and grade analysis – Finally, the third machine learning experiment will run all selected algorithms to classify by grade, for each college separately, and assess performance, and evaluate decision trees for college-based rules.

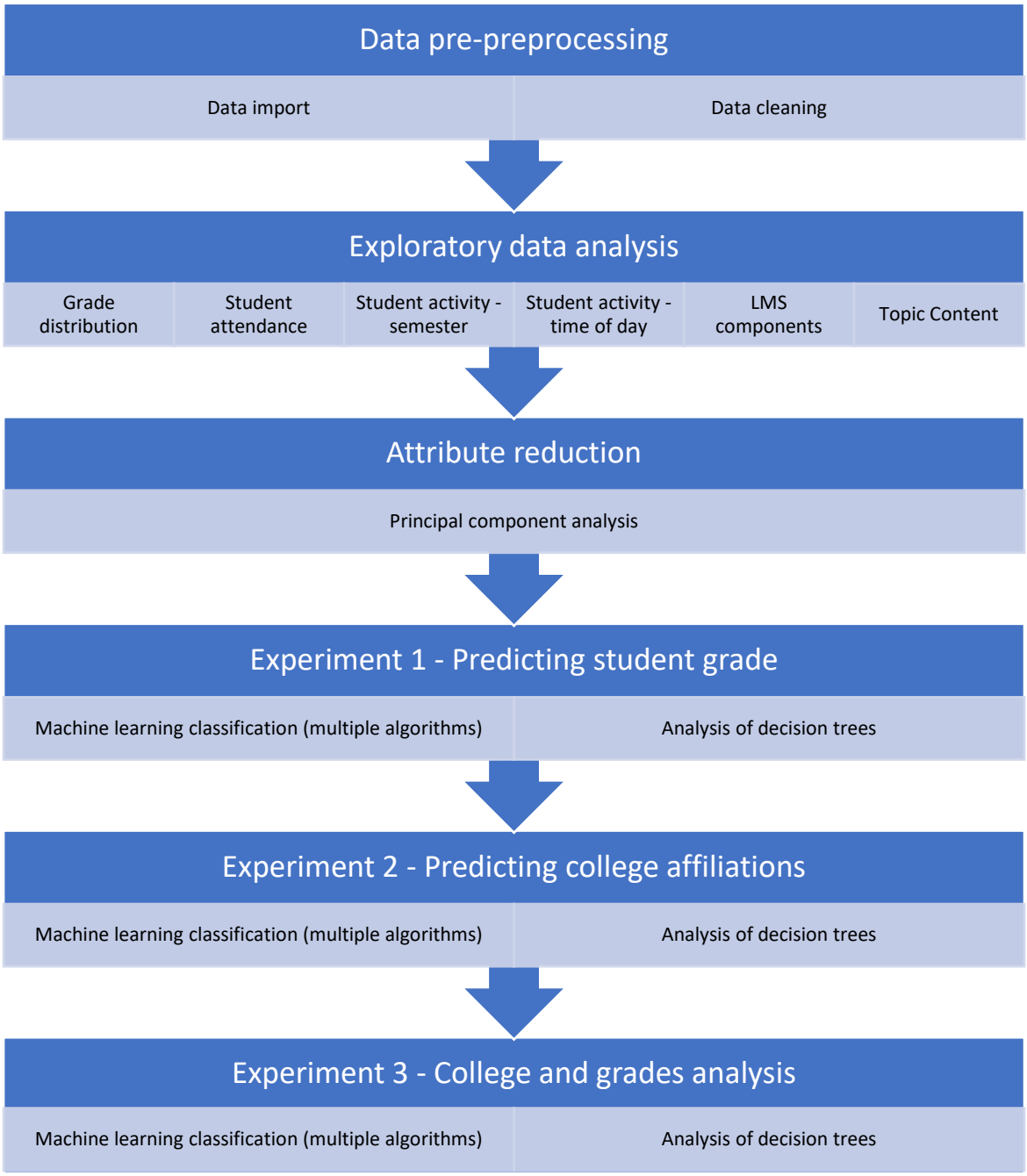

Figure 12 - Experiment methodology

3.2. Data ethics approval information

Approval for the use of Flinders University was granted on 20 July 2018 (Project #7987), the research was deemed by Flinders university Social and Behavioural Research Ethics Committee (SBREC) as being low risk. Only logs of direct interactions with the FLO LMS are required, no personal information from students was requested. This does remove some potential forms of analysis, as no information on student year level, or method of enrolment (such as on-campus/off-campus, or part-time/full-time) is available.

The data is de-identified and stored on secure university research drives. Data used in WEKA analysis are summarised .csv files, with no personally identifiable information (such as student name, student ID etc.), or specific topic information (such as a full topic code, e.g. COMP1001). The only identifiable material kept, is that of semester and daily activity, and general college level attendance in the form of truncated topic codes, mapped to the college that offers those groups of topics (e.g. COMP1001 truncated to COMP, and mapped to the College of Science, and Engineering).

Re-identification of Student Data

As mentioned by Yacobson et al. (2021), student data can often be re-identified by comparing the de-identified data to publicly available resources, such as social media accounts. However, in this case, specific topic details are not retained, and semester and time data are kept in grouped blocks of weeks and time of day respectively. Therefore, it would be less likely a student schedule or daily activity pattern can be matched to the data stored in this research.

Available Dataset context

While the dataset provides limited information about students, it does reveal some details. The dataset includes student interactions with FLO on various topics across several years: 2013 (3 students, 3 topics), 2014 (4,278 students, 788 topics), 2015 (15,309 students, 1,179 topics), and 2016 (17,877 students, 1,745 topics).

3.3. Hardware and software

3.3.1. Computing hardware

The workstation used across experiments was a Windows 10 Education (22H2) x64 workstation, with an AMD Ryzen 9 5950X CPU (16 cores, 32 threads 3.4Ghz base clock 4.9Ghz boost), 128Gb of PC4-28800 (28,800 MB/s peak) ram, and a Samsung 980 Pro NVMe PCIe 4.0 2280 SSD (7,000MB/s max read, 5,100MB/s write).

3.3.2. Software requirements

Due to the large size of the initial dataset .csv , the use of Microsoft SQL Server 2017 Enterprise (Microsoft 2018a), and SQL Server Management Studio (Microsoft 2018b), enabled the creation of a database of the required size. With regards to which version of Microsoft SQL Server to use, there are two main considerations, the amount of ram and the number of CPU cores the software can utilise.

First, considering the ram limitation, according to Microsoft's (2023d) product description for the free editions of SQL Server 2017, there was a limitation of 1.4 GB memory, which would significantly hinder usability for this use case. Whereas the Standard and Enterprise editions' limits (128 GB, and OS max respectively) would allow for this dataset to be created with no issues.

Finally, considering CPU cores, according to the product description for each SQL 2017 version (Microsoft 2023d), the limitation of 4 cores for the free Express version, 24 for the Standard, and OS max for Enterprise. To fully utilise the 16 physical cores, and 32 logical cores of the AMD Ryzen CPU, at least SQL Server 2017 Standard would be necessary.

Therefore, Microsoft SQL Server 2017 Enterprise was chosen, considering it was the only version that supported all the required features. The ability to handle large datasets, and ability to utilise the hardware made available for the research was a critical aspect of this initial software package, for such a large dataset.

The primary software utilised in this research for data mining and machine learning tasks is WEKA version 3.8.6 (Frank, Hall & Witten 2016). This widely recognised, free-to-use suite of machine learning algorithms for data mining tasks is favoured in the research community for its accessibility and comprehensive resources.

Additionally, MATLAB version 9.12.0.1975300 R2022a (The MathWorks Inc. 2022) and R Statistical Software version 4.2.3 2023-03-15 ucrt (R Core Team 2023) utilising the RStudio: Integrated Development Environment for R "Ocean Storm" version release 33206f75, 2023-12-17 (RStudio Team 2023), were used for Principal Component Analysis (PCA) and various visualisations across each experiment.

3.4. Data pre-preprocessing and transformation

3.4.1. Aim

The aim of this stage is to efficiently transform the dataset from a large .csv file into a more manageable and processable format, suitable for analysis. The primary tasks include preprocessing and cleaning tasks, followed by aggregating the data into structured student-topic-interaction rows. These processes significantly reducing the dataset's size for improved handling. Additionally, this stage involves conducting a comprehensive feature analysis of the interaction logs. This analysis aims to identify, extract, and refine pertinent features, ensuring that the dataset is optimally prepared for subsequent processing stages. Each step of this transformation will be meticulously logged to maintain transparency and facilitate a clear understanding of the data manipulation processes involved.

3.4.2. Setup

After deciding on using the data gathered, the first step was to ensure that the data was in a format that will be usable for analysis, and additionally in a format that will be quick and efficient to perform initial analysis. Given the size of the initial logs file (FLO_S12015_16.csv), a manual review was impractical. This size also hampered initial analysis, as software like Microsoft Excel was unable to handle the large file, often leading to crashes, a challenge echoed by various other software suites.

The FLO_S12015_16 data was imported directly into Microsoft SQL Server 2017 Enterprise; this was done using the built-in import functionality in the software.

Purpose

This initiation stage is where source data is imported from comma separated text files stored securely on the Flinders University research drive, into the Microsoft SQL Server 2017 database.

Input Files

FLO_S12015_16.csv	- Usage logs of FLO system from 2015-2016 (14.2 GB).
raw_grades_27_08_18.csv	- Student grades from 2009-2017 (69.7 MB).
topics_data.csv	- Flinders University (2018) Topic information (349 KB).
Sem_dates.csv	- Semester information (9.70 KB)

Data Context

Due to the limitation of the lack of specific context regarding aspects of the data, specific information will not be available, such as the student engagement status (proportions of full- and part-time students), location of study (on- or off-campus). Additionally, any topic that does not utilise the LMS for content, will not be tracked. Only direct interactions with the LMS are available for analysis.

Tables are manually generated, with appropriate data types and lengths. This step was complicated for FLO_S12015_16, with the excessively large size of certain attributes (such as 'name'). While most of the 'name' entries were small, there were outliers which would break the bulk import process, so an overly large maximum length was chosen.

Components are identified through selecting all distinct component names from the FLO_S12015_16 database, and then assigning them a component category depending on what type of component it is (e.g. lecture video, quiz, assignment, etc.).

For the FLO_S12015_16 data, there were 46 distinct component names, which were able to be divided up into seven separate component types (video, assignment, quiz, support, participation, forum, and other). Resulting tuples from the insert process consisted of the component name, and the component type associated with the component name.

The seven types: Video, Assignment, Quiz, Support, Participation, Forum, and Other (for everything else not classified), were chosen to summarise the components into a more easily analysed form and reduce the overall complexity for later stages.

Video

Components under the video category consist of objects such as lecture videos (Table 2).

Table 2 - Video type LMS components

mod_kalvidres
mod_lecture

Assignment

The assignment components included all objects that were involved with assignment participation, submission, and feedback (Table 3). For the purpose of this project, an assignment is an activity that is submitted through FLO (or feedback is received through FLO for a physical assignment hand-in) and is differentiated from a quiz by its general lack of interaction.

Table 3 - Assignment type LMS components

assignsubmission_comments	mod_feedback
assignsubmission_file	mod_kalvidassign
assignsubmission_onlinetext	mod_workshep
mod_assign	mod_workshop

Quiz

Quiz components are similar to those of the assignment components (mentioned above); however, both involve a more interaction with the participant (Table 4). Both types of quiz components shown below involve a LMS managed learning activity (or series of activities). Consisting of teacher generated content, and feedback, with back-end functionality to (in most cases) automatically grade and give feedback to participants in a regulated manner.

Table 4 - Quiz type LMS components

mod_activequiz
mod_quiz

Support

Support components primarily consist of objects involved in more generic aspects of a course, which are not specifically tailored to an online teaching space (Table 5). Aspects such as written materials (such as lesson materials not put into video form), static pages, and topic grading materials. TurnItIn content was also categorised as support, with it not specifically being related to the use of the LMS regarding functionality. Information on specific support material for student well-being, or other specific support material is not known with the data and is therefore a limitation that should be noted.

Table 5 - Support type LMS components

booktool_print	mod_folder	mod_turnitintool
gradereport_grader	mod_glossary	mod_turnitintooltwo
gradereport_overview	mod_lesson	report_log
gradereport_user	mod_page	report_outline
mod_book	mod_resource	report_stats

Participation

Participation consisted of objects related to participant interactions with the topic using the LMS functionality (Table 6). Objects such as attendance marking, topic-coordinator communication, group selection, group management, simple feedback collection, wiki management, and scheduling. While there is some limited participant to participant interaction (through the group selection), and participant to topic-coordinator (through the dialog options), the main feature of this component type is interaction, or 'participation' with the topic through the functionality of the LMS.

Table 6 - Participation type LMS components

block_comments	mod_groupselect
core	mod_oublog
mod_attendance	mod_ouwiki
mod_choice	mod_scheduler
mod_dialogue	mod_wiki

Forum

The forum component type is essentially how participants communicate using the functionality of the LMS (Table 7). This consists of direct messaging (chat), and through the managed forums. The intention of this component type is to identify all social interactions using the LMS, to analyse how this affects the outcomes of participants.

Table 7 - Forum type LMS components

mod_chat
mod_forum

Other

The 'Other' component tag was used to classify any other component not already classified (Table 8). This was not used for any analysis but was included for completeness.

Semester blocks are manually generated by organising the semester periods (found in the sem_dates table), into five separate groups. Each grouping depends on a combination of the semester (either 'S1' or 'S2'), and the semester period. The resulting tuple with consist of the semester, the semester period, and a numerical value to represent the block number.

Table 8 - 'Other' type LMS components

mod_data	mod_scorm
mod_imscp	mod_subcourse
mod_lti	mod_url
mod_mapleta	

3.4.3. Data transformation phase 1

Objective

The primary goal at this phase is to convert the large .csv dataset into a structured and manageable format suitable for analysis. The procedure commences with cleaning and preprocessing tasks on the data, followed by the aggregation of student-topic-interaction data. This approach significantly trims the dataset for enhanced manageability. An exhaustive feature analysis of the interaction logs is also undertaken to pinpoint, extract, and refine critical features, thereby preparing the dataset for the following stages of processing. Each transformation step is documented in detail to preserve the integrity and transparency of the data manipulation process.

Tables Created

FLO_S12015_16	- Unmodified import of the usage logs.
raw_grades	- Unmodified import of the student grades.
topic_data	- Unmodified import of the topic data.
components	- Helper table to map component values to component types.
sem_blocks	- Helper table to map week values into ‘blocks’ of weeks.
sem_dates	- Helper table to provide date boundaries to map week values.

All tables except the FLO_S12015_16 have primary keys associated with each unique tuple. This is due to the nature of the dataset; timestamps for each action appear to have been truncated prior to being received, therefore multiple sequential interactions have been shown to have been given identical timestamps. This was manifested with integrity constraints being broken part way into the importing process, causing initial attempts to import the data to fail.

Without a way to uniquely identify each tuple, in the raw logs FLO_S12016_16 is unable to conform to First Normal Form (1NF) due to no combination of any attribute (singular of composite) being able form a Primary key. However, this issue will be remedied in a later stage, at this stage, FLO_S12015_16 will have a list of participant interactions (identifiable by id, topic, date, and semester).

The remaining tables consist of raw_grades table will be a Third Normal Form (3NF) table containing student grades, topic_data that will be a 3NF table containing topic information extracted from the Flinders website, and 3 helper tables in 3NF containing data generated to help describe the log data (component, sem_blocks , and semdates).

Data Inserted

A bulk insert operation is performed on FLO_S12015_16, raw_grades, topic_data, and sem_dates from external comma separated values files (FLO_S12015_16.csv, raw_grades_27_08_18.csv, topics_data.csv, and sem_dates.csv respectively).

Tuples from sem_dates consist of the year in question, and a series of date values representing the beginning of each semester period. The naming convention of the semester periods is that of the semester in question (such as 'S1'), an underscore, and the week of that semester period. Depending on the semester, the valid numerical ranges for weeks can be between 1 to 7 for non-semester 1 (NS1), 0 to 14 for semester 1, and 1 to 13 for semester 2. Non-numerical week values include mid-semester breaks (MB1, and MB2), end of semester breaks (B1, and B2), and exam periods (E1, and E2).

Manual insert operations are performed on components, and sem_blocks, with data directly added in the SQL script.

3.4.4. Data transformation phase 2

Objective

This phase is dedicated to refining the dataset, enhancing it with additional attributes, and purging redundant entries. It aims to forge a dataset that mirrors student engagement on the E-Learning platform with precision, readying it for predictive analytics.

Tables Created

FLO_PROCESS - FLO logs, with additional attributes, and irrelevant tuples deleted.

topics - Processed topic information from raw data, of only relevant topics.

The table FLO_PROCESS is manually generated with appropriate data types and lengths. Primary Keys are still not present at this point (data is not in 1NF). While rows are not able to be uniquely identified, they will be used for summary purposes in the next stage.

The topics table is manually generated with appropriate types and lengths. The table is in 3NF, with a composite Primary key of (topic_shortcode, study_period, year).

Data Inserted

An insert operation on FLO_PROCESS is performed by selecting topic information from FLO_S12015_16, with null or zero values inserted as placeholders for ‘week’ and category respectively. The insert is constrained to only insert rows that contain a student_id that exists in the raw_grades table, this eliminates any non-participants (such as admin users, or topic-coordinators).

The timestamp attribute is transformed into integer values for year_val, month_val and day_val. The function ‘datepart’, will extract the specified part of the date attribute ‘timestamp’ from FLO_S12015_16 tuples. Each required attribute for the insert process (year_val, month_val, and day_val) are extracted using the appropriate datepart variables, further information on the function is available on the Microsoft SQL documentation website (Microsoft 2022).

In addition to year, month, and day components of the timestamp being extracted, the hour_period value the timestamp is extracted. Hour_period is an integer value between 1 and 8, representing three-hour blocks of time beginning from 12:00am to 11:59pm.

Microsoft T-SQL is able to utilise more advanced functionality such as case statements (Microsoft 2023a); this functionality is well suited for quick logical analysis of attribute data. Once again, the 'datepart' function is utilised to extract the hour value from the timestamp, then the case statement allows for the extracted hour value to essentially be converted into an integer value to represent the required three-hour block as described above.

The last insert operation is performed on the topics table, selecting topic information from FLO_PROCESS, with zero values inserted as placeholders for future updated values of category, and various component counts (lecture, quiz, assignment, etc.).

Data Updated

For the FLO_PROCESS table, the first value that needed to be updated was the week_value attribute.

The T-SQL code described above fulfils several purposes. Firstly it converts the 'timestamp attribute from a 'datetime2' down to a 'date' attribute using the 'cast' function (Microsoft 2023b). The purpose of this conversion is to truncate the time data (hours and minutes) from the timestamp, which will allow for a differentiation between timestamps.

The second purpose is to create boundary dates for the case logic to work correctly. This is performed by taking the date lower boundary of the semester period stored in the sem_dates table, and then utilise the dateadd function (Microsoft 2023c), to add seven days to that value to create the upper boundary for the semester period.

The case functionality mentioned earlier allows the timestamp to be categorised into the relevant semester period, then have the text value of that period stored in the 'week' attribute.

The College category number is then updated in the topics table, this will represent the Flinders University College that is associated with each of the topics.

Topic information stored in topic_data refers to a Flinders University topic as a whole and does not differentiate between years or semesters. Assigned College categories were therefore allocated to all versions of the topic regardless of year or semester. Historical information on specific cross-college topics that are not identified in this information is unavailable and is noted as a limitation on the data.

To assist with the category matching between topics and the topic_data table, the T-SQL function 'left' is used (Microsoft 2023e). This simply extracts a specified number of characters from the supplied text attribute (in this case, eight characters are extracted, to represent the four-character topic code, and four-digit topic number). This process will also exclude any year or semester information in the process (those values are stored as additional characters after the main topic code and number, separated by an underscore.

The College information is then added to the FLO_PROCESS tuples to associate the College to the interactions.

The next step in the process was to summarise all the interaction features of the topics involved. This was one of the most important steps with regard to data collection, and analysis. This process is lengthy, and similar for each required attribute. Therefore, only the relation showing the distinct videos for each topic is described below, all subsequent instances not outlined are available in the appendix.

At the end of this update process the total number of unique component types are generated and stored in topics. This is done by counting the total number of distinct component ids (cmid) for each component type.

That this causes a limitation with the information gathered for the topics; the unique components and total components recorded for each topic is only that of the observable features of the topics found in the logs themselves. In other words, if there are additional features that the topics have (such as videos or activities that are not documented in the logs), then they will not be recorded for the purpose of this study. Since I was not given access to each individual topic and could not verify the existence of any content that was not in the logs provided, it will be assumed that only what is shown in the logs exists.

For the purpose of this investigation, this is an appropriate assumption to make, considering that if no student makes use of a component for the topic, then we cannot assume its effectiveness. While this may cause the numbers of content for some topics to not reflect the actual amount, the purpose of this analysis is not to analyse topics themselves, but the use of said topics by participants.

For future studies, it would be highly recommended to have separate logs containing accurate information on the composition of each topic individually. However, the lack of such information, is not a detriment, and as long as it is stated that information is gathered by observations, the analysis stage would still be measuring all participants using the same metrics (with observations of content from all participants being used, to measure individual students).

Data Deleted

A delete operation is performed on FLO_PROCESS to remove all rows from the dataset that are from out-of-scope topics (such as non-semester or summer topics, rather than standard semester 1 and 2 topics).

Additionally, rows which contain timestamps that fall outside of normal topic duration are removed, given students appear to have access to the topics long after completion of the topic, this would not have any effect on the analysis.

This is to ensure all data is easily comparable (non-standard topics, have variable duration, and are often split across different physical datasets), and to reduce overall workload for the query. Given the initial number of rows imported from FLO_S12015_16.csv (88,212,495 rows), and the substantial processing time required to perform any operations on this table, excluding irrelevant tuples at the earliest possible stage was an imperative.

3.4.5. Data transformation phase 3

Purpose

The final stage of transformation aims to summarise the data into a streamlined format, explicitly structured for machine learning applications and statistical analysis. This process finishes in a dataset that best details student interactions within the LMS, optimised for the subsequent analysis stages.

At this stage, the FLO_interactions table will be a fully 3NF table, containing unique student, topic, semester, and year tuples. The data summarises all interactions, interaction type, time of day performed, and semester period performed, that a student makes for a specific topic, in a specific semester and year.

The grades table is a reduced version of raw_grades, only containing the relevant student information required for FLO_interactions.

Having this form of granular and indexable information on student/topic interactions is critical for the final analysis stage; showing the characteristic of all forms of participant interactions and allows for mapping of outcomes for said participants from the relevant grade entry.

Tables Created

FLO_interactions	- Unique participant/topic table, with interaction information.
Grades	- Relevant participant grades, allowing for outcome mapping.

The FLO_interactions table generated with appropriate types and lengths, the table is in 3NF, with a composite Primary Key of (student_id, topic_shortcode, study_period, year).

The Grades table is manually generated with appropriate types and lengths, the table is in 3NF, with a composite Primary key of (student_id, topic_shortcode, study_period, year).

The choice of keys for FLO_interactions cause the tuples to be uniquely identifiable based on an individual participant, for a specific topic, during a specific semester of a specific year. Which is the same as what was provided by the initial raw_grades, and the updated Grades tables.

This is exactly what is required for analysis, as a summary of any other breakdowns of the data, such as all interactions of a single participant for all topics (or all attempts by said participant), or all participants for a single topic, would not allow the mapping of outcomes from the Grades table.

Data Inserted

An insert operation on FLO_interactions was performed by inserting summaries of participant interactions from FLO_PROCESS (grouped on topic_shortcode, category, study_period, year, and student_id). The aggregate function SUM was used to create numerical summaries of all interaction types, a non-exhaustive list of said operations is as follows; Individual interaction component types (based on the 'component' attribute), and grouped interaction component types (grouped into Assignment, Video/Lecture, Support, Activity, Social, Participation, and Other). An insert operation is performed on Grades from FLO_PROCESS, inserting all observed student and topic combinations. This ensures that only relevant rows are recorded, and any combinations not observed, potentially due to the non-use of the FLO system (making the outcome irrelevant for the purpose of this investigation), being excluded from the operation.

Data Updated

Grades rows are updated with relevant raw_grades grade values, additionally, they are updated with numerical representations of all grade labels. The numerical representation is based on the Flinders standard grading metric, with 7, 6, 5, and 4 being High Distinction, Distinction, Credit, and pass respectively (and all being a positive outcome overall). Grade values that would be a negative outcome are unequivocal fails ('F'), and indirect fails such as 'F/A', 'F/M', 'FAS', 'FCP', 'W/F', are considered negative outcomes and are all given numerical grade values of zero.

'NA', or 'NoGrade' are all removed.

FLO_interactions is updated with grade values from the grades table, along with a selection of topic related information and topic/component related information.

Days_active, which is calculated by taking just the day, month, and year component of the date timestamp, and performing a distinct operation, and counting the output, effectively finds all the different date values found in all timestamps for the participant/topic combination, and then counting how many returns.

Distinct component types, which are similar to the information recorded in the Topics table, however, instead recording all distinct components only the participant views/interacts with.

Topic related information such as component type counts, which will allow for the calculation of 'percentage of' for component types, for a given topic in the next stage.

Data Deleted

Participant summary tuples (in FLO_interactions) that do not show a definitive outcome state (such as a final grade that will be recorded), are removed during this stage.

Grades such as 'CO' (Continuing), 'I' or 'I/M' (incomplete, and incomplete deferred medical assessment), indicate that this attempt is not the final attempt recorded for this participant, and is overwritten at a later stage. Additionally, grades of 'WN' signified a withdrawal from the topic before census date (at approximately 20% of the topic duration).

Due to the lack of a proper outcome (positive or negative) for the participants, tuples containing grade values of 'CO', 'I', 'I/M', or 'WN', are deleted. This does not negatively affect the analysis, in fact, it will ensure that outcomes can be measured more effectively.

3.4.6. Data selection

For the data to be useful in predicting student outcomes, it would need to include either a numerical grade or a pass/fail outcome associated with each user in the dataset. It should also be large enough to provide sufficient data for the machine learning algorithms, and to better ensure significant results. Log data needs to be identifiable to both the user, to the topic in which they are interacting with, and the period in which the event occurred (such as semester and year). The data needs to have an identifiable id for the student, and an identifiable topic code with accompanying teaching period and year. Demographic factors would be ideal, however, due to the nature of the need for de-identification, this would not be possible, and would be limited to potential future research.

3.4.7. Selection criteria & data handling

The data that was gathered for use included a large log file of interactions made between the users and FLO, with associated tags and other meta information. It additionally included student outcomes (in the form of grades for topics), which would be able to be mapped onto the interactions from the log file. All data that was from students (i.e. had mappings available from the student outcomes), was selected, and processed. Data from non-students (e.g. teachers or LMS administrators), was discarded.

The data is de-identified and stored on secure university research drives. Data used in WEKA analysis are summarised .csv files, with no personally identifiable information (such as student name, student ID etc.), or specific topic information (such as a full topic code, e.g. COMP1001). The only identifiable material kept, is that of semester and daily activity, and general college level attendance in the form of truncated topic codes, mapped to the college that offers those groups of topics (e.g. COMP1001 truncated to COMP, and mapped to the College of Science, and Engineering). As mentioned by Yacobson et al. (2021), student data can often be re-identified by comparing the de-identified data to publicly available resources, such as social media accounts. However, in this case, specific topic details are not retained, and semester and time data are kept in grouped blocks of weeks and time of day respectively. Therefore, it would be less likely a student schedule or daily activity pattern can be matched to the data stored in this research.

3.5. Exploratory data analysis

Exploratory data analysis, in this research, refers to a critical phase within data analytics, particularly within the realm of an LMS and its impact on student academic environments. Exploratory data analysis is a process pioneered by John W. Tukey (1977), aimed at understanding and summarising the main characteristics of a dataset, often visually, before formal modelling or hypothesis testing (Tukey 1993, pp. 1-5). This approach helps in uncovering the underlying structure of the data, identifying outliers, anomalies, patterns, and trends without making any initial assumptions about the data's distribution or outcome relationships.

In this specific context, exploratory data analysis serves as the initial step that allows researchers and educators to pull apart and investigate different aspects of the data (allowing for inspection of specific attributes/features) of the data generated by the LMS. By analysing the student engagement and academic performance data from multiple angles, this exploratory data analysis helps to uncover patterns and relationships that might not have been found in other circumstances. Behrens (1997) discusses the methodologies and benefits of exploratory data analysis, while discussed from the perspective of psychological research, the methodologies and benefits outlined are applicable to E-Learning data analysis and emphasise the importance of exploratory data analysis for answering research questions (RQ1.1, RQ1.2, RQ1.3) focused on understanding how different aspects of LMS usage, such as interaction with specific features and tools or access patterns based on student demographics, correlate with educational outcomes.

As mentioned in Section 2.6.1 Data analytics, and discussed by Moubayed et al. (2018) the process of exploratory analytics involve a wide variety of statistical techniques and visualisation methods, such as those mentioned by Wickham & Grolemund (2016). The goal of this stage is to provide a statistical foundation for further predictive analytics, enabling a more data-informed decision-making process on what LMS features, and student metrics can be utilised to predict student performance, and to help enhance teaching and learning processes.

By integrating the insights gained from this exploratory data analysis, including investigating the detailed information on LMS usage and student engagement information, this exploratory data analysis stage aims to offer a more in-depth perspective on the usage of LMS environments in supporting educational achievement across a variety of different disciplines.

3.5.1. Grade distribution

Aim

The initial analysis begins by examining the distribution of student grades across different colleges. The goal is to identify whether there are notable variations in grade outcomes that may indicate disparities among colleges. This approach leverages student and LMS data to substantiate findings, offering a data-driven alternative to reliance on questionnaires and self-reporting methods. This analysis seeks to pinpoint disparities in grade outcomes among colleges, thereby addressing RQ1.1 and providing a foundational understanding of how LMS usage varies across disciplines and its potential implications on student performance as indicated in RQ1.2.

Setup

The overall grade distribution between colleges was plotted in R using the RStudio interface. To facilitate a fair and balanced comparison, the data was normalised to adjust for size disparities between colleges. In addition, the distribution was shown as percentage values to adjust for different college sizes.

A Pearson's (1900) Chi-squared test was conducted to examine the differences in grade distributions across colleges (Montgomery 2017, p. 31). This test compares observed frequencies against expected frequencies in each category of a contingency table. It is used to test the hypothesis of independence between categorical variables, in this case, to determine if the distribution of grades is independent of the colleges.

An inspection of the expected frequencies in R, based on the contingency table, was performed to check if expected cell frequencies were 5 or greater. This step is crucial to validate the assumption for the Chi-squared test, ensuring the appropriateness of the test for the given data.

Finally, given that the conditions for the Chi-squared test were met (i.e., all expected cell frequencies were 5 or greater), there was no need to resort to Fisher's (1922) Exact test. Although Fisher's Exact Test is preferred in certain scenarios, particularly because it calculates the exact probability of observing the data assuming the null hypothesis is true, it is generally not feasible for large datasets due to its computational intensity. In contrast, Pearson's (1900) Chi-squared test is more suitable for larger samples where the expected frequency assumptions are met.

3.5.2. Student attendance

Aim

The aim of this stage is to analyse the patterns and levels of student attendance and activity within the LMS to understand how these factors correlate with academic performance. By categorising student activity into distinct levels and examining the distribution of active days across different colleges, this analysis seeks to identify trends and disparities in student engagement. The insights gained will be crucial for evaluating the impact of student attendance on learning outcomes and for identifying potential areas for intervention to enhance student engagement and academic success. By analysing attendance and activity patterns, this stage aims to establish a correlation between LMS engagement and academic performance, feeding into the predictive analytics model (RQ1.3).

It also sets the stage for a deeper dive into the behavioural aspects of LMS usage across colleges (RQ2.2), paving the way for tailored interventions.

Setup

The overall distribution of days active between colleges was plotted in R using the RStudio interface. The distribution was shown as whole values to as each college utilises the same semester structure.

In addition to measuring the absolute number of days active, student activity is categorised into three distinct levels: 'High', 'Medium', and 'Low'. These categories are defined based on equal intervals relative to the maximum number of active days observed: High - students active for more than two-thirds of the maximum number of active days, Medium - students whose activity falls between one-third and two-thirds of the maximum observed, Low - students active for up to one-third of the maximum number of active days.

In addition to the distribution of total days active for students across colleges, the relationship between grade and days active was also plotted in RStudio.

A Pearson's (1900) Chi-squared test was conducted to examine the differences in student activity levels across colleges.

Considering that the 'grade' attribute is of an ordinal nature, the data preparation process, particularly the categorisation of 'days_active' into 'High', 'Medium', and 'Low', was informed by the principles outlined by Agresti (2010, p. 37). This categorisation is a critical step in analysing the relationship between an ordinal independent variable and an ordinal outcome. The 'days_active' variable was divided into three ordered categories based on its distribution, aligning with the methodology's emphasis on respecting the ordinal nature of the data. This approach facilitates a more nuanced analysis by ensuring that the inherent order in the 'days_active' variable is appropriately considered in subsequent statistical modelling. The categorisation sets the stage for any further analysis, such as ordinal logistic regression, where these ordinal categories serve as a predictor for the ordinal outcome 'grade'.

3.5.3. Average student activity across the semester

Aim

The aim of this stage is to examine the average student activity throughout the semester, separating it by college and academic grades. This involves a detailed investigation into how student interactions with the LMS vary across different times of the semester, such as orientation weeks, mid-semester breaks, and exam periods. By dissecting the temporal distribution of student interactions, this analysis aims to reveal trends and patterns that could inform strategies to enhance student engagement and academic success during specific segments of the semester. This temporal analysis aims to identify critical periods impacting student engagement and performance. Insights from this stage will inform the development of predictive models (RQ1.3) and contribute to understanding college-specific pedagogical approaches (RQ2.3), particularly in managing student workload and activity throughout the semester.

Setup

The overall level of student activity across the semester between the colleges was plotted in RStudio. For a uniform analysis of both semester 1 and semester 2 interactions (given the uneven allocation of mid-semester breaks, and total weeks), the week values were grouped into semester blocks as shown in Figure 13.

Semester block	Semester 1	Semester 2
1	Week 0 (orientation week) Week 1 Week 2	Week 1 Week 2
2	Week 3 Week 4 Week 5 Week 6	Week 3 Week 4 Week 5 Week 6
3	Week mid break 1 Week mid-break 2 Week 7 Week 8	Week 7 Week 8 Week mid break 1 Week mid-break 2
4	Week 9 Week 10 Week 11 Week 12	Week 9 Week 10 Week 11 Week 12
5	Week 13 Week 14 Week Exam 1 Week Exam 2	Week 13 Week Exam 1 Week Exam 2

Figure 13 - Semester block allocation table

To assess the data's distribution, an Anderson-Darling (1952) test was performed to determine if the dataset follows a normal distribution (Montgomery 2017, p. 80). This test is advantageous as it is less sensitive to data ties and not heavily influenced by sample size limitations. It was applied to assess the normality of total interactions grouped by college, grade, and semester period.

Subsequently, the Kruskal-Wallis (1952) was employed to compare median semester interaction levels across colleges. This non-parametric method is used for testing whether samples originate from the same distribution (Montgomery 2017, p. 123). It's particularly useful as an alternative to the one-way ANOVA when the data does not meet ANOVA's assumptions, such as normality and homogeneity of variances. Furthermore, the Kruskal-Wallis (1952) test is suitable for analysing data that are not normally distributed, aligning well with situations where the Anderson-Darling test indicates non-normality.

To further delineate these differences, post-hoc pairwise comparisons were conducted using Dunn's (1964) test to identify which groups differ. This was achieved by comparing the sum of ranks between groups and then adjusting for multiple comparisons. In conjunction with this, Cohen's (1992) d test was utilised to measure the effect size, indicated by the mean difference between two colleges. Combining both tests facilitated the identification of differences between colleges and the quantification of the magnitude of these differences. Given the extensive dataset, only results with a p-value less than 0.05 and a Cohen's (1992) d value greater than 0.5 or less than -0.5 were considered significant.

Finally, the average number of interactions per college at each semester period was plotted in RStudio, to identify any high or low periods across the semester for each college.

3.5.4. Average student activity across daily time periods

Aim

The aim of this stage is to evaluate the average student activity within the LMS across different daily time periods. This analysis aims to identify peak periods of student engagement and explore how activity levels fluctuate throughout the day. Understanding these patterns will help in assessing the alignment of LMS interaction with students' daily schedules and identifying potential temporal factors that may influence academic performance and engagement. By dissecting student activity throughout the day, this stage seeks to align LMS interactions with students' daily routines, offering insights into how time-of-day factors into student engagement and performance, thereby addressing aspects of RQ1.2 and RQ2.2.

Setup

The overall level of student activity across time of day between the colleges was plotted in RStudio, and time values were grouped into 3-hour blocks.

The Anderson-Darling (1952) test was performed. This test was applied to total interactions grouped by college, grade, and time periods.

The Kruskal-Wallis (1952) test was also employed to compare median daily interaction levels across colleges.

Dunn’s (1964) test along with Cohen's (1992) d test were conducted to measure the effect size of the differences, with only results with a p-value less than 0.05 and a d value greater than 0.5 or less than -0.5 being retained.

Finally, the average number of interactions per college at each time period was plotted in RStudio, to identify any high or low periods across the day for each college.

3.5.5. LMS components

Aim

The purpose of this stage is to examine the utilisation patterns of various LMS components by students and to assess the relationship between these patterns and academic performance. This analysis aims to identify which LMS components (e.g., videos, quizzes, assignments, forums) are most engaged with by students, understand how the usage of these components correlates with student grades, and determine if specific components have a more pronounced impact on academic outcomes. This insight will help in understanding the efficacy of different LMS components in supporting student learning and identifying areas for potential improvement in online learning environments.

This stage aims to correlate specific LMS components with academic outcomes, providing a direct link to RQ1.2 by identifying which features of LMS use are significant predictors of student performance and enhancing the understanding of how specific LMS interactions contribute to the predictive models.

Setup

Student usage of various LMS components were compared with grade outcomes, as well as to draw correlations between the various components.

A Pearson (1895) correlation test was performed, to test the strengths of relationships between the percentage viewed of each LMS component and the corresponding numerical grade values. Numerical grades were mapped from categorical values ('HD' = 5, 'DN' = 4, 'CR' = 3, 'P' = 2, 'F' = 1) to facilitate quantitative analysis. LMS components tested included: interactions, days active, percentage of videos, percentage of quizzes, percentage of assignments, percentage of participation modules, percentage of support modules, percentage of 'other' modules, forum usage (percentage of topic), forum posts (number student posts).

The Anderson-Darling (1952) test was performed on each of the LMS components to test for normality.

Next, the average interaction types across colleges are compared using the Kruskal-Wallis (1952) test, with p-values adjusted for multiple comparisons. The adjustment methodology utilises the Bonferroni (1936) method, a widely used statistical method to adjust for multiple comparisons mentioned by Dunn (1961), and used to address the problem of multiple comparisons. The Bonferroni (1936) adjustment controls the Family-Wise Error Rate (FWER) by lowering the significance threshold for each individual test. It should be noted that the use of the Bonferroni (1936) adjustment does increase the likelihood of type II errors, however, this can be reduced with the addition of the Hochberg (1988) step-up and Holm (1979) step-down methods that are less conservative and more modern compared to the use of just the Bonferroni (1936) adjustment.

3.5.6. Topic content

Aim

The aim of this stage is to analyse the content structure of different academic topics within the LMS across various colleges. This involves examining the diversity and quantity of LMS content components (e.g., lecture videos, assignments, quizzes) utilised in each topic, and assessing how these components are distributed across different subject areas and colleges. The goal is to uncover patterns in the educational content delivery, understand the relationship between topic content complexity and student engagement or performance, and identify any discrepancies or notable trends in content provision among different academic disciplines. This analysis seeks to unpack the complexity and diversity of topic content across colleges, addressing RQ2.3 by examining how content structure and variety influence pedagogical approaches and student performance, thereby offering insights into discipline-specific instructional design.

Setup

The same tests as the previous were performed; The Anderson-Darling (1952) test, The Kruskal-Wallis (1952) test, and Dunn's (1964) test along with Cohen's (1992) d test. Each were repeated for the observed components of each topic; total unique number of LMS content that students interact with grouped by individual topic codes.

While not an exact measure of actual number of LMS components per topic, this is an accurate measure of what content was interacted with and would contribute to any student outcomes for the period investigated.

3.5.7. Controlling for multiple testing

Aim

The purpose of this stage is to implement rigorous statistical controls to address the potential for Type I error inflation due to the extensive nature of the dataset, which includes 147,780 observations and involves numerous pairwise comparisons. The objective is to apply appropriate Family-Wise Error Rate (FWER) control methods such as the Bonferroni (1936) method to maintain the overall integrity and reliability of the statistical analysis made from the dataset. This stage of the analysis is critical for accuracy in answering the research questions (specifically, RQ1.1 to RQ1.4), and ensuring the following experiments draw from credible information.

Setup

Given the size and complexity of the FLO dataset, which include a significant number of observations, and the necessity for many pairwise comparisons across the dataset, it critical to address the potential for a significant number of Type I errors (known as false positives). To mitigate this risk, a conservative approach to statistical testing was taken, techniques such as the Bonferroni (1936) correction, were applied to adjust the significance thresholds appropriately, to ensure the integrity of the statistical results.

For the pairwise comparisons between colleges, which involve a significant number of comparisons due to the number colleges involved (six colleges), a significance level of $p < 0.001$ was set. This threshold was chosen, taking into account the Bonferroni (1936) correction method, which is used to adjust the alpha level based on the number of comparisons made. This approach is further supported by procedures such as the Hochberg (1988) step-up and Holm (1979) step-down methods, which provide a balanced approach to controlling the FWER, which is able to identify any significant differences while minimising the risk of Type I errors.

For the individual predictions (for example, student-topic comparisons), a more lenient alpha level of 0.05 is deemed appropriate, given the increased number of comparisons/results. This level is also adjusted for multiple testing using FWER, considering the context of the analysis which includes multiple hypotheses being tested concurrently. This customised approach helps to ensure that the analysis remains reliable, and the findings statistically significant.

3.6. Attribute reduction (Principal Component Analysis)

3.6.1. Aim

The aim of this stage is to refine and optimise the dataset for subsequent analytical and machine learning processes by reducing its dimensionality. This involves identifying and eliminating redundant, irrelevant, or less significant attributes to streamline the data. The objective is to enhance the efficiency and interpretability of the analysis if possible.

The attribute reduction method employed in this experiment is Principal Component Analysis (PCA), as described by Han, Kamber, and Pei (2012, p. 483). PCA, originally developed by Pearson (1901a), is executed using Singular Value Decomposition (SVD). SVD is a mathematical technique relating to eigenvector analysis that, according to Stewart (1993), has evolved through contributions from numerous mathematicians from the early 1800s to the mid-1950s. Additionally, Latent Semantic Analysis (LSA) was applied to help better visualise the resulting information. LSA is a refactoring of SVD described by Deerwester et al. (1990), commonly used for information retrieval purposes.

Through this process, the goal is to discover a reduced number of latent features that contribute meaningfully to understanding and predicting patterns of student performance and engagement. Consequently, this will help to ensure a more focused and effective analysis. This stage directly addresses RQ1.4 by assessing the necessity and efficacy of dimensionality reduction in capturing the core aspects of LMS use and enhancing model performance. It seeks to strike a balance between model simplicity and predictive power, ensuring that the attributes retained are most indicative of student performance and engagement patterns. Additionally, this analysis can offer support for research question RQ2.1 offering insight via identifying patterns in the data relating to LMS attributes.

3.6.2. Data selection

The initial analysis encompassed all students enrolled in courses with recorded grade outcomes. It was observed that certain student interactions with course content extended beyond the standard timeframe, indicating that some students may have retaken the course in a subsequent semester. Future research will consider datasets filtered to include only single-semester students, only multi-semester students, and unfiltered data to examine these different cohorts.

3.6.3. Setup

Manual attribute review

The first step involved a manual review to remove features that were evidently irrelevant or redundant. The initial SQL query, ambitious in its breadth, captured an extensive range of LMS components and learning activity metrics. However, due to constraints in processing capacity and the time allocated for analysis, it was essential to refine the dataset.

For instance, specific LMS component interactions, detailed in section 3.2.4 and encompassing 46 distinct component names, were deemed extraneous as their information was already encapsulated within broader collected component types. This reduction effectively decreased the attribute count to 147. Similarly, LMS usage types—Create, Read, Update, and Delete, coded as ‘c’, ‘r’, ‘u’, and ‘d’—along with numerical representations of grades (which duplicated categorical grade information), were omitted. Additional details such as 'topic_shortcode', 'study_period', 'year', and 'student_id' were excluded for their redundancy and minimal contribution to nuanced analysis. These exclusions resulted in a more manageable set of 138 features for in-depth processing.

While a detailed breakdown of LMS component usage and specific component interactions within timeframes and semesters could provide a wealth of information, it also significantly increases the volume of data to be analysed. The initial SQL result contained 193 columns and 147,780 rows. Given the exponential increase in processing requirements with additional attributes, a strategy involving dimensionality reduction was adopted to streamline the dataset for subsequent exploratory analyses.

Data Import and Preprocessing

The dataset, stored in a .csv format, was imported via the MATLAB interface. The script for this process, along with other operations, is documented in Section 7.3 Appendix C: MATLAB script for PCA, clustering and figures.

Principal Component Analysis (PCA)

The following section will describe the application of SVD in the Principal Component Analysis phase, analysing the singular values to determine the number of principal components to retain, based on the explained variance.

The primary aim at this stage is to condense the dataset into an optimal number of attributes that best describe the dataset's variance. Additionally, there is an opportunity to utilise PCA, and to a lesser extent LSA, to probe the complex relationships and characteristics inherent in the data. This process involves not merely reducing dimensionality but also interpreting the principal components to comprehend the latent structures and dynamics within the dataset. Advanced visualisation techniques, such as biplots for PCA, which show both the principal components and the original variables in the same plot, and topic modelling visualisations for LSA, will be utilised to clarify these relationships and patterns. This approach provides a more profound understanding of the intricate nature of student interactions within the LMS.

First, the data was standardised to ensure that each feature contributed equally to the analysis, eliminating disproportionate influences from features with larger numeric ranges. Standardisation involved adjusting each feature to have a mean of 0 and a standard deviation of 1.

SVD is applied to the demeaned and standardised data matrix (Enrolments x Attributes), decomposing it into three matrices: U (left singular vectors), S (singular values), and V' (right singular vectors). This decomposition allows for the identification of principal components, facilitating data analysis and visualisation. SVD as applied to Enrolments x Attributes (ExA), is $E \times A = U \times S \times V'$.

In LSA, derived from the SVD, the singular value matrix S is reinterpreted through a transformation into L and L', where $L = S\frac{1}{2}$. This enables the representation of both enrolments and attributes in a common latent space by facilitating a unified scaling. The equation becomes $E \times A = U \times L \times L' \times V'$, where $S = L \times L'$, and this decomposition assists in visualising the data's structure in the latent space. The matrices U and V serve as rotations, transforming the data to a new basis without changing its geometric relationships. S, or L in the context of LSA, provides scaling, adjusting the significance of each dimension. This scaling is crucial for interpreting the data's structure, with $U \times S\ (or\ U \times L\ in\ LSA)$, and $V \times S\ (or\ V \times L\ in\ LSA)$ allowing for the visualisation of enrolments and attributes, respectively, in different or common latent spaces.

Both U and V are orthonormal matrices, ensuring that the transformations they represent preserve the original data's distances and angles, which is important for retaining the essence of what the dataset contains. The singular values in S, arranged in descending order, indicate the dimension's importance. For LSA, reducing the dimensionality involves selecting a subset of these singular values (and the corresponding vectors in U and V), focusing on the most significant latent dimensions.

Based on the cumulative variance explained, a selection of principal components was made for detailed analysis. This approach ensured that enough components were visualised up to an appropriate amount of variance explained, while capturing the most significant patterns and relationships within the data.

Loading information and attribute name mappings from the selected components were analysed, with attributes sorted according to their absolute loadings to highlight those with the greatest impact on each component. The top 20 attributes were retained for each component, facilitating a focused visualisation and interpretation of the most influential factors.

Finally, the identified components were systematically compared to identify commonalities or distinct patterns among them. This analysis aimed to identify underlying themes or relationships across the components, providing deeper analysis into the structure of the dataset. Specific attention was paid to overlapping attributes, trends in loadings, and how each component contributes to explaining the dataset's variance.

3.7. Experiment 1 - Predicting student grade

3.7.1. Aim

The aim of Experiment 1 is to investigate and quantify the relationship between student LMS behaviour and their academic outcomes, specifically focusing on one-semester students. This includes examining various LMS interaction metrics, their frequency, and types to understand how these behaviours correlate with the final grades. Additionally, this experiment aims to assess the influence of college affiliation on student performance, potentially identifying unique behaviour patterns or engagement strategies associated with different colleges. While the primary focus is on understanding the predictive factors of academic performance (addressing RQ1.1 and RQ1.2), insights from this experiment can also provide a foundational understanding of behavioural patterns that may differ across colleges. These insights will be instrumental for Experiment 2, where the prediction of college affiliation based on LMS data becomes the central focus.

3.7.2. Data selection

The same dataset from the pre-processing and analysis stage is used for this and all subsequent experiments. The dataset is filtered by one semester students (not repeating students), and split into several subsets for analysis, the subsets for experiment 1 are shown in Table 9.

Table 9 - Experiment 1 attribute subset information

subset	**observations**	**attributes**	**size**
Predict grade excluding college attribute	129,643	130	51,433KB
Predict grade including college attribute	129,643	131	51,697KB

3.7.3. Setup

While the primary research interest lies in a tree-based, easily interpretable algorithm (such as the Reduced Error Pruning Tree) that can quickly produce meaningful rules, additional algorithms were also compared for a comprehensive analysis. The choice of algorithms and the metrics for performance evaluation have been discussed in detail in previous chapters. The algorithms used in subsequent experiments with filtered versions of the dataset are as follows:

Algorithms

1. RepTree (minimum number per leaf = 20).
2. DecisionStump.
3. RandomTree.
4. J48.
5. J48 (with REPTree pruning).
6. NaiveBayes.
7. simpleCART.
8. RandomForest (with 10 trees).
9. RotationForest (using REPTree with minimum number per leaf = 20).
10. NBTree.
11. AdaBoostM1 (using REPTree).

Evaluation Criteria

1. Weighted average recall.
2. Weighted average Cohen's KAPPA.
3. Weighted average Matthew's Correlation.
4. Weighted average AU ROC.

Additional performance criteria were included.

5. Training time.
6. Tree size (size of tree algorithms).
7. Size of serialised representation.

To ensure the accuracy of tests, k-fold cross-validation was utilised, this is process is described by Yadav & Shukla (2016). The dataset is split into 'k' number of randomly assigned and equally sized subsets of the dataset. To which each of these sets is used as a hold-out to test the model, while the rest are used as training data for each hold-out.

This process is repeated 'k' times. The benefit of this process is every observation is used to both train and testing the dataset during analysis. The results from each of the tests is then averaged. The number of folds ('k') necessary for accuracy on large datasets can be set relatively low, as noted in results from Yadav & Shukla (Yadav & Shukla 2016); with instances between 5,000 and 100, 000 folds of size 5-6 being performing accurately, while not being too computationally costly.

Each algorithm was used to classify all subsets of the dataset and was run with 10-fold cross-validation, given the extra computational hardware available (5-fold may have been suitable, but for improved accuracy, 10-fold was chosen). WEKA was initialised in multiple instances to run concurrently in the same workstation.

An initial test of all algorithms on limited datasets and a single run took approximately 3 days and 3 hours using one instance of WEKA. Given that this limited run took an excessive amount of time, and the overall CPU usage and physical memory usage only reached approximately 25% and 40% respectively, it was decided to run multiple WEKA instances simultaneously to better utilise system resources. Analysis processing times varied, with each of the tree classifiers taking between 2 and 6 hours, and the forests taking a similar time due to distribution across 10 threads.

For this experiment and Experiment 2 - Predicting college affiliations, and Experiment 3 - College and grades analysis, 'paths' through the decision trees will be discussed. As depicted in Figure 14 leaf nodes are characterised by both a class value and a numerical accuracy value (for instance, X with 0.75, and Y with 0.50). This numerical value calculates accuracy by dividing the number of misclassified instances within the leaf by its total instances, following Eibe's (2015) explanation of the REPTree methodology for calculating leaf weightings. This

measure reflects the path's classification precision within the tree, indicating the proportion of correctly classified instances reaching the leaf.

From Figure 14 it can be suggested that with the path to ‘X’ showing 0.75 (75%), compared to the path to ‘Y’ 0.50 (50%), the path to ‘X’ has greater accuracy in classification.

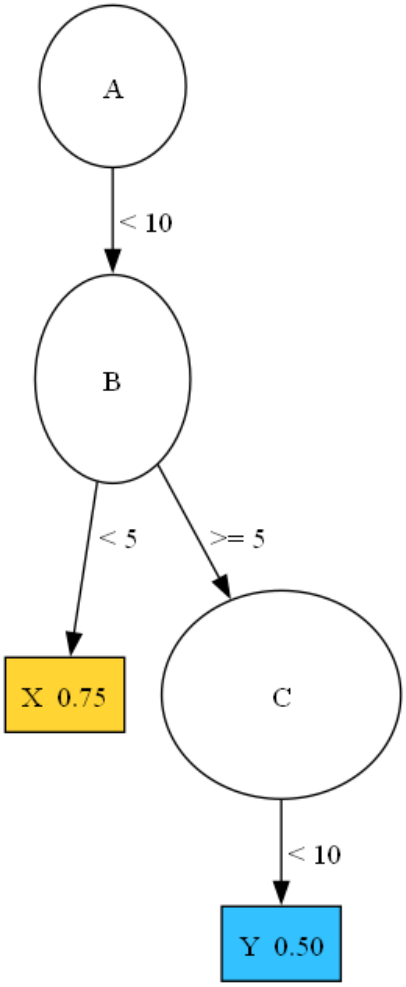

Figure 14 - Path through decision tree example

3.8. Experiment 2 - Predicting college affiliations

3.8.1. Aim

The aim of Experiment 2 is to explore the predictive capacity of LMS behaviour metrics concerning the college affiliation of a topic. This involves analysing LMS interaction data to identify distinctive behavioural patterns and engagement levels associated with topics from different colleges. The objective is to identify whether the college affiliation of a topic influences the way students interact with the LMS. By addressing RQ1.3, this experiment aims to identify if LMS behaviour can predict college affiliation, exploring the unique educational and engagement patterns associated with different colleges. The findings from this experiment are expected to reveal the nuances of college-specific engagement and inform the development of targeted strategies for enhancing the LMS experience tailored to the needs and preferences of each college.

3.8.2. Data selection

The same dataset with subset filtering as experiment 1 is used for this experiment. The dataset is filtered by one semester students (not repeating students), and split into several subsets for analysis, the subsets for experiment shown in Table 10.

Table 10 - Experiment 2 attribute subset information

subset	observations	attributes	size
Predict college excluding college attribute	129,643	130	51,350KB
Predict college including college attribute	129,643	131	51,697KB

3.8.3. Setup

The same algorithms and performance metrics are used for this and the following experiment. Each algorithm was used to classify all subsets of the dataset and was run with the same 10-fold cross-validation as the previous experiment.

3.9. Experiment 3 - College and grades analysis

3.9.1. Aim

The aim of Experiment 3 is to utilise the insights gained from understanding the predictive relationship between LMS behaviour and academic outcomes (Experiment 1) and the capacity to predict college affiliation based on LMS data (Experiment 2). The primary objective is to develop and evaluate college-specific predictive models that not only forecast student grades but also consider the intricacies of college affiliation as a significant factor. This experiment aims to refine the predictive models by focusing on college-specific factors (addressing RQ1.3 and Q1.5), enhancing the accuracy and relevance of the models for individual colleges. It seeks to uncover unique influences on student performance within each college, offering insights that could guide custom interventions and pedagogical strategies (RQ2.1, RQ2.2, RQ2.3).

3.9.2. Data selection

The same dataset with subset filtering as experiment 1 is used for this experiment. The dataset is filtered by one semester students (not repeating students), and split into several subsets for analysis, the subsets for experiment 3 are shown in Table 11.

Table 11 - Experiment 3 college subset information

subset	observations	attributes	size
Predict grade BGL college	25,269	130	9,739KB
Predict grade EPS college	29,735	130	11,519KB
Predict grade HAS college	21,034	130	7,915KB
Predict grade MPH college	10,323	130	4,295KB
Predict grade NHS college	18,623	130	7,770KB
Predict grade S&E college	24,659	130	10,224KB

3.9.3. Setup

The same algorithms and performance metrics are used. Each algorithm was used to classify all subsets of the dataset and was once again, run with the same 10-fold cross-validation as the two previous experiments. It is worth noting, that even though the observations are significantly lower per-college (maximum of 29,735 for EPS), it is still above the 5,000 threshold mentioned by Yadav & Shukla (2016) of 5,000. Again, for accuracy, and given the computing hardware available, 10-fold was retained as per the last two experiments.

3.10. Chapter summary

This chapter provided a detailed exposition of the methodologies employed in this investigation, aimed at exploring the influence of LMS usage on student performance across varying disciplines and colleges. It commenced with an outline of the robust hardware and software configurations, establishing a solid computational foundation for the research. Following this, the chapter described the meticulous process of data preprocessing, transforming voluminous raw data into a structured format conducive to analysis. This initial stage was crucial for addressing RQ1.3 and RQ1.4, which focus on predictive analytics and the necessity of dimensionality reduction, respectively.

The exploratory analysis stage provided significant information for the study, identifying multiple LMS/University patterns and correlations, providing information that can be used to directly, and indirectly be used to answer the primary research questions. Techniques such as grade distribution analysis, and student attendance analysis, this section was important in uncovering the nuances of LMS interaction across colleges, and directly linking to RQ1.1 and RQ1.2 by identifying multiple significant predictors of student performance, enrolment, and activity. Additionally, the analysis of LMS components and content delivery approaches directly link to RQ2.3, identifying how specific LMS features and pedagogical strategies impact student outcomes.

The attribute reduction stage, specifically through Principal Component Analysis (PCA), and Latent Semantic Analysis (LSA), aimed to assess the need for streamlining the dataset, and to identify patterns in the dataset. This stage was directly aligned with RQ1.4, addressing the need for managing the dataset's complexity while not losing any valuable information.
The chapter also detailed a series of machine learning experiments to further investigate the dataset, and to identify college specific patterns within the data.

Experiment 1 primarily focuses on predicting student grades, which identifies factors in the dataset that relate to student academic performance. This experiment directly addresses research questions RQ1.1 and RQ1.2.

Experiment 2, using similar techniques as Experiment 1, was aimed at predicting college affiliations based on LMS data. This experiment identifies educational patterns across colleges, which directly addresses research question RQ1.3.

Finally, Experiment 3 returned to the same methodology as Experiment 1, however, is applied on a college-by-college basis. This provides college-specific factors that affect student outcomes, directly addressing research question RQ1.3 by refining the predictive models to directly address colleges individually. Additionally, this series of experiments were critical for addressing RQ2.1, RQ2.2, and RQ2.3, by identifying specific patterns, and roles of attributes used in prediction, which can be used to provide a customised educational approach.

Throughout the methodology chapter, a transparent, and rigorous methodology was outlined to ensure that data handling and statistical analysis are reliable, consistent, and statistically significant. The methods selected were specifically chosen to best address the outlined research questions, and to ensure that the study's objectives were met, setting a solid foundation for the subsequent analysis and discussions in subsequent chapters.

Overall, this chapter not only detailed the methods employed but also highlighted the detailed approach taken in this research to ensure that the findings are correct and will contribute meaningfully to the field of E-Learning, to better help with understanding student learning behaviours in online environments.

4. Results

4.1. Chapter overview

This chapter outlines the results of the following experiments.

Exploratory data analysis: This section presents the outcomes of all statistical tests performed on the dataset, highlighting observed patterns.

Principal Component Analysis (PCA): Describes the results from each stage of the PCA process, including calculating variance explained, comparing individual principal components, and comparing multiple principal components. This section directly addresses research question RQ1.4.

Experiment 1: Presents results from the machine learning tasks, comparing the performance and complexity of each algorithm for grade classification. Research questions RQ1.1 and RQ1.2 are directly addressed.

Experiment 2: Presents results from the second series of machine learning tasks, focusing on comparing the performance and complexity of algorithms for college affiliation classification. Research question RQ1.3 is directly addressed.

Experiment 3: Results from the final series of machine learning tasks are presented here. This experiment repeats the methodology of Experiment 1 but is targeted at individual colleges to provide specific college-based grade classification. It directly addresses research questions RQ1.3 and RQ1.5, while also supporting research questions RQ2.1, RQ2.2, and RQ2.3.

4.2. Exploratory data analysis

4.2.1. Grade distribution

Figure 15 presents the percentage of students in each college awarded specific grades, normalised to prevent skewed comparisons between colleges of varying sizes.

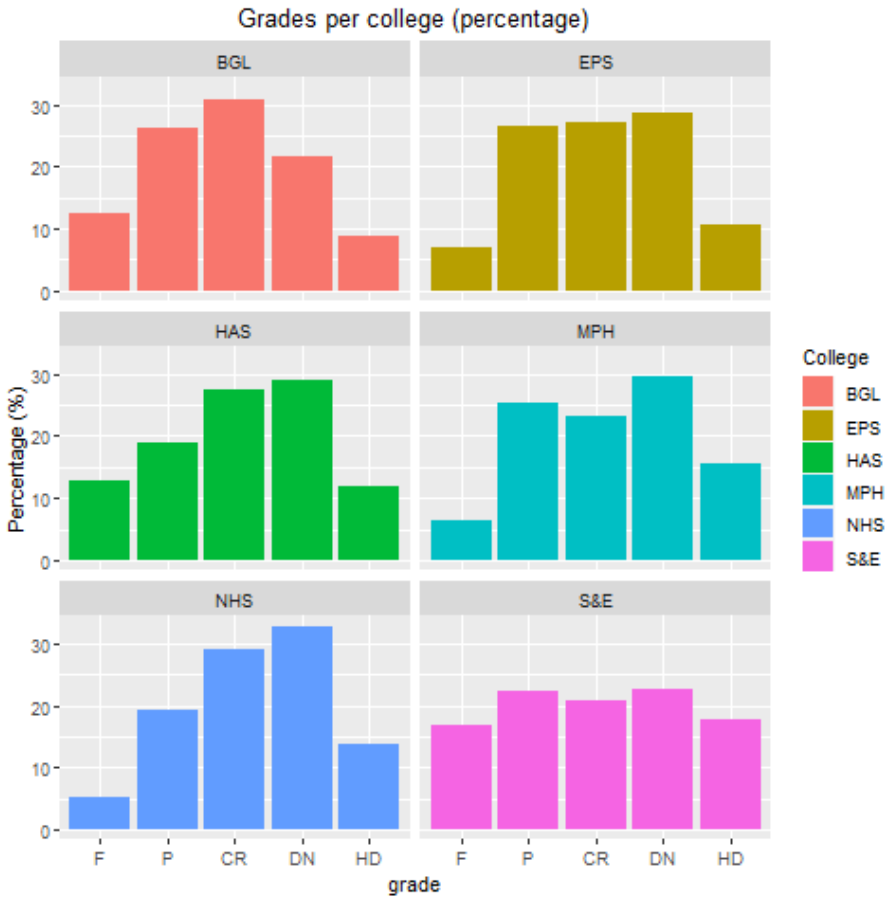

Figure 15 - Student grades grouped by college

Table 12, summarises the overall grade distribution. While visual inspection suggests variations between colleges, further statistical analysis is necessary to confirm if these differences are significant.

Table 12 - Grade distribution across colleges

college	F	P	CR	DN	HD
BGL	3640	7744	9090	6380	2545
EPS	2386	8951	9155	9721	3578
HAS	3072	4597	6666	7031	2885
MPH	785	3072	2818	3587	1875
NHS	1110	4045	6108	6896	2890
S&E	4691	6280	5840	6349	4993

The Pearson's (1900) Chi-squared test yielded a Chi-squared statistic of 5697.85 with 20 degrees of freedom, indicating significant differences in grade distributions. This suggests that the college attended might influence student grade outcomes. Importantly, an inspection of the expected frequencies in the contingency table confirmed that all expected cell frequencies were 5 or greater, satisfying the assumption for the validity of the Chi-squared test.

As such, the conditions for the Chi-squared test were met, and there was no need to resort to Fisher's (1922) Exact test. Table 13 provides a detailed comparison of the observed and expected frequencies of grades for each college.

These results suggest a notable variability in grade distributions across colleges, warranting further investigation into potential causes such as differences in course difficulty, teaching, and assessment methods.

Table 13 - Comparison of observed vs. expected frequencies (from chi-square test)

grade	college	observed	expected	difference
F	BGL	3,640	3,099.27	540.73
P	BGL	7,744	6,854.80	889.20
CR	BGL	9,090	7,840.46	1,249.54
DN	BGL	6,380	7,897.17	-1,517.17
HD	BGL	2,545	3,708.30	-1,163.30
F	EPS	2,386	3,562.37	-1,176.37
P	EPS	8,951	7,879.05	1,071.95
CR	EPS	9,155	9,012.00	143.00
DN	EPS	9,721	9,077.18	643.82
HD	EPS	3,578	4,262.40	-684.40
F	HAS	3,072	2,556.79	515.21
P	HAS	4,597	5,654.97	-1,057.97
CR	HAS	6,666	6,468.11	197.89
DN	HAS	7,031	6,514.90	516.10
HD	HAS	2,885	3,059.22	-174.22
F	MPH	785	1,279.87	-494.87
P	MPH	3,072	2,830.75	241.25
CR	MPH	2,818	3,237.79	-419.79
DN	MPH	3,587	3,261.21	325.79
HD	MPH	1,875	1,531.38	343.62
F	NHS	1,110	2,219.46	-1,109.46
P	NHS	4,045	4,908.87	-863.87
CR	NHS	6,108	5,614.73	493.27
DN	NHS	6,896	5,655.34	1,240.66
HD	NHS	2,890	2,655.59	234.41
F	S&E	4,691	2,968.45	1,722.55
P	S&E	6,280	6,565.45	-285.45
CR	S&E	5,840	7,509.51	-1,669.51
DN	S&E	6,349	7,563.83	-1,214.83
HD	S&E	4,993	3,551.77	1,441.23

Suggested explanations and discussion

While the data itself does not directly suggest any reason for the differences in distribution, potential causes may be due to course difficulty and structure differences across colleges, or differing teaching and assessment methodologies. However, as the data does not indicate anything, these are merely suggestions.

Observed differences

The first observed difference is in the overall Fail grades. The College of Business, Government, and Law (3640 vs. 3099.166), the College of Humanities, Arts, and Social Sciences (3072 vs.2556.79), and the College of Science and Engineering (4691 vs. 2968.45) each showed higher Fail grades that expected. While the College of Education, Psychology, and Social Work (2386 vs. 3562.37), the College of Medicine, and Public Health (785 vs. 1279.87), and the College of Nursing and Health Sciences (1110 vs. 2219.46) each showed fewer Fail grades that expected.

The second observed difference is the number of High Distinctions. The College of Education, Psychology, and Social Work (3578 vs. 4262.40), the College of Medicine, and Public Health (1875 vs. 1531.38), the College of Nursing and Health Sciences (2890 vs. 2655.59), and College of Science and Engineering (4993 vs. 3551.77) all showed an increased observation of HD grades. Conversely, the College of Business, Government, and Law (2545 vs. 3708.30) and the College of Humanities, Arts, and Social Sciences (2885 vs. 3059.22) both showed notably fewer HD grades than expected.

Overall, the data shows that the failure rate appears to vary significantly between colleges, with the College of Science and Engineering appearing to have the highest proportion while the College of Medicine, and Public Health and the College of Nursing and Health Sciences appearing to have the lowest (-494.87 and -1109.46 respectively).

Additionally, the proportion of 'HD' grades varies significantly across all colleges, with the College of Science and Engineering having the greatest positive difference (+1441.23), and the College of Business, Government, and Law having the greatest negative difference (-1163.30). This may indicate a difference in the distribution of high achievers or potentially in differing grading standards across colleges. The Pass grade appears to be the most frequent grade across most colleges, except for the College of Education, Psychology, and Social Work, where the distribution is flatter, and the College of Science and Engineering, where 'CR' and 'DN' are the more common grades.

4.2.2. Student attendance

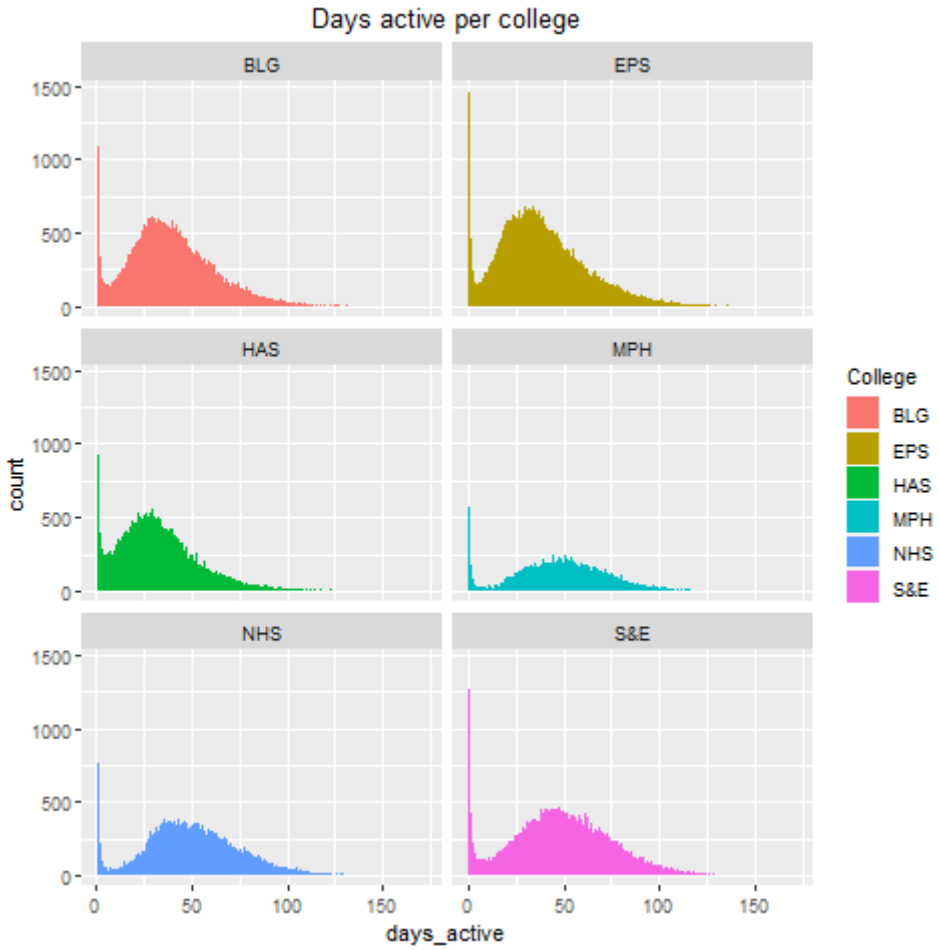

Figure 16 - Student days active in LMS grouped by college

The analysis then turns to student activity within the LMS. Figure 16 displays the total days students were active in the LMS, categorised by college. With the total days active on the x-axis (ranging from 1 to 180 days) and the counts of students on the y-axis. In addition to measuring the absolute number of days active, student activity is categorised into three distinct levels: 'High', 'Medium', and 'Low'. These categories are defined based on equal intervals relative to the maximum number of active days observed: High - students active for more than two-thirds of the maximum number of active days, Medium - students whose activity falls between one-third and two-thirds of the maximum observed, Low - students active for up to one-third of the maximum number of active days.

In addition to the distribution of total days active for students across colleges, Figure 17 illustrates the relationship between grade and days active. The Pearson's (1900) Chi-squared test (X^2 = 7445.7, df = 10, p < 2.2e-16) confirms significant variations in student activity levels across colleges. This suggests that there is a statistically significant association between the colleges and the categorised levels of days_active ('Low', 'Medium', and 'High'). The contingency table, showing the number of students in each activity level across all colleges, is presented in Table 14. The results indicate a difference in activity levels across colleges.

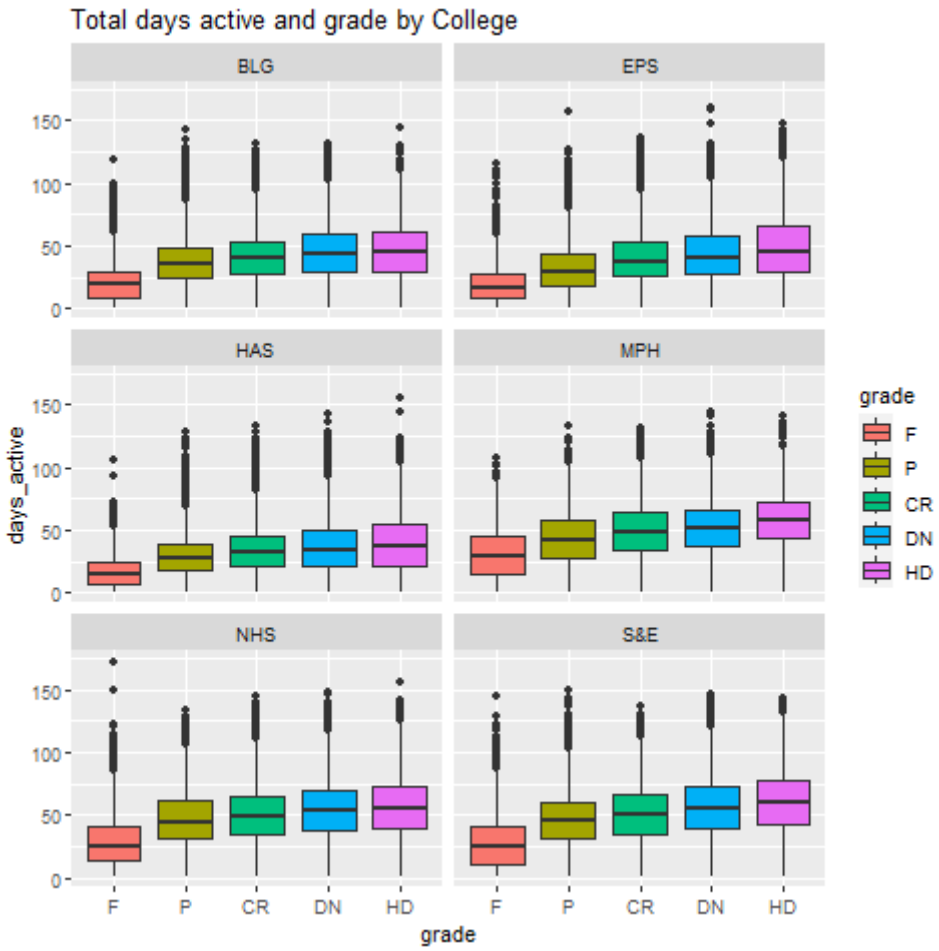

Figure 17 - Total days active in LMS and grade by college

Table 14 - Contingency table of days_active for each college

	Low	Medium	High
BGL	23717	5618	64
EPS	27355	6308	128
HAS	21362	2835	54
MPH	7837	4234	66
NHS	13371	7509	169
S&E	17978	9940	235

Table 15 - Significant residuals from Chi-squared test

College	Activity_Level	Adjusted_Residual
BGL	Low	24.980453
EPS	Low	16.474499
HAS	Low	25.734300
MPH	Low	-27.757803
NHS	Low	-23.925649
S&E	Low	-24.078664
BGL	Medium	-41.678258
EPS	Medium	-28.333591
HAS	Medium	-44.145642
MPH	Medium	48.287705
NHS	Medium	40.700737
S&E	Medium	40.805521
BGL	High	-14.550875
EPS	High	-3.553680
HAS	High	-6.359951
MPH	High	2.073823
NHS	High	8.354873
S&E	High	9.517807

Significant residuals from the Chi-squared test (Table 15) reveal trends in activity levels. Negative residuals for 'Low' activity in colleges like MPH suggest fewer less-active students than expected, contrasting with colleges like BGL. The results from the post-hoc analysis of the Chi-squared test, shown in Table 15, indicate how the odds of being in a higher category of grade are influenced by changes in the 'days_active' category.

The negative residuals for 'Low' activity in MPH, NHS, and S&E (−27.76, −23.93, and −24.08, respectively) suggest fewer students with low activity levels than expected, contrasting with BGL, EPS, and HAS. For the 'Medium' activity category, BGL, EPS, and HAS show negative residuals (−41.68, −28.33, and −44.15, respectively), indicating fewer students in these categories than expected. This could point to either a concentration of students in the 'Low' category or a jump to 'High' activity without a substantial middle group. Negative residuals in the 'High' activity category for BGL, EPS, and HAS (−14.55, −3.55, and −6.36, respectively) suggest that these colleges have fewer highly active students than expected.

The data suggest that there are different levels of LMS engagement across each of the colleges. Notably, BGL, EPS, and HAS have a concentration of students in the 'Low' activity category, while MPH, NHS, and S&E have a more even distribution across each of the activity levels, with fewer students in the 'Low' category and more in the 'Medium' and 'High' categories.

4.2.3. Average student activity across the semester

Average interaction values were calculated grouped by semester period and grade as shown in Figure 18. The results indicate a correlation between activity levels, grade outcomes, and semester periods.

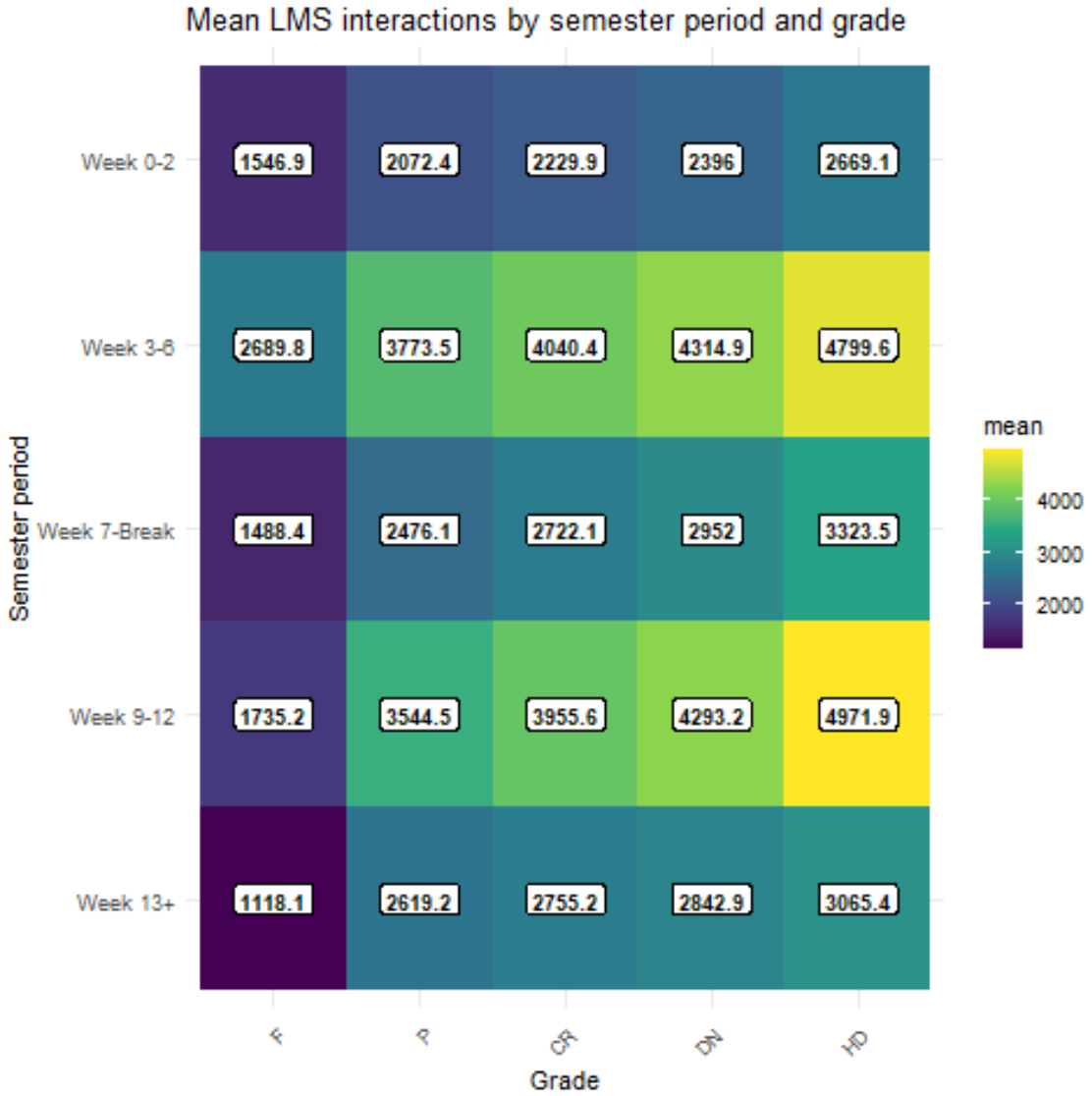

Figure 18 - Average interactions by semester period and grade

The Anderson-Darling (1952) test showed an extremely low p-value (1.097e-05) strongly suggests that the data is not normally distributed across colleges for each semester period.

The Kruskal-Wallis (1952) test yielded a Chi-squared value of 78.807 with 5 degrees of freedom and a highly significant p-value (1.491e-15), indicating substantial differences in interaction patterns across colleges.

The negative Z values from Dunn's (1964) test (shown in Table 16), indicate that the first college in each pair consistently has a lower median interaction count compared to the second. Similarly, the negative values of Cohen's (1992) d suggest that the first college in each pairing tends to have a lower mean interaction count.

Table 16 - Semester periods summary of significant results (Cohen's d test)

Period	College Pair	Z	P-Value	P Adjusted	Cohen's d
Week 0-2	HAS-MPH	-3.232895	0.0006127120	0.009190681	-1.776441
Week 0-2	BGL-NHS	-3.089211	0.0010034436	0.015051654	-2.309708
Week 0-2	HAS-NHS	-3.699869	0.0001078553	0.001617829	-2.327979
Week 3-6	HAS-MPH	-3.196974	0.0006943863	0.010415795	-2.309708
Week 3-6	HAS-NHS	-3.340659	0.0004178996	0.006268494	-2.327979
Week 3-6	HAS-S&E	-3.089211	0.0010034436	0.015051654	-2.191463
Week 7,8, break	HAS-MPH	-3.017369	0.0012748958	0.019123437	-2.309708
Week 7,8, break	HAS-NHS	-3.232895	0.0006127120	0.009190681	-2.327979
Week 7,8, break	HAS-S&E	-3.125132	0.0008886257	0.013329385	-2.191463
Week 9-12	HAS-MPH	-2.837764	0.0022715396	0.034073094	-2.309708
Week 9-12	HAS-NHS	-2.730001	0.0031667106	0.047500659	-2.327979
Week 9-12	HAS-S&E	-2.945527	0.0016120253	0.024180379	-2.191463
Week 13+	HAS-MPH	-3.161053	0.0007859986	0.011789979	-2.309708
Week 13+	HAS-S&E	-3.628027	0.0001427976	0.002141965	-2.191463

In addition to the interactions across semester periods, the data from Figure 19 shows similar grade patterns across each college and semester periods, with all showing consistent increases in grades with increasing interactions, and similar peak times across the semester. However, each college shows a unique pattern of interaction across the semester.

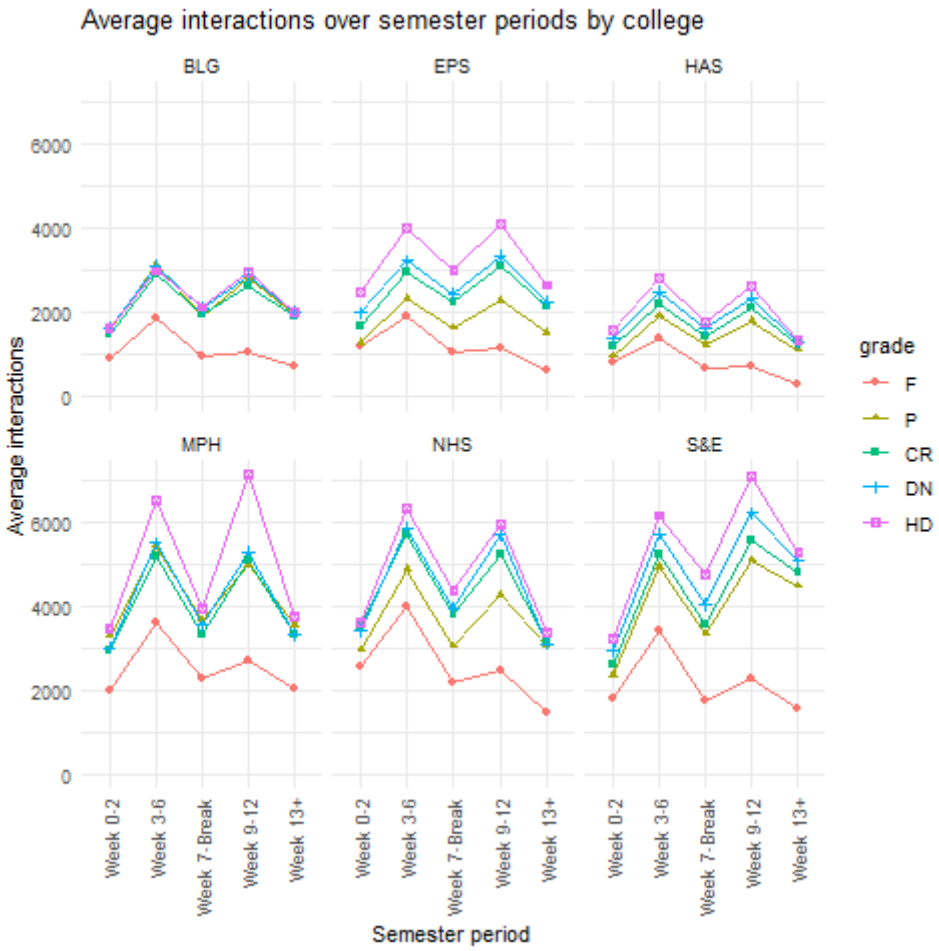

Figure 19 - Average interactions over semester period by grade for each college

Analysis of Interaction Patterns and Differences

From the data, HAS College showed the most significant differences in interaction patterns compared to other colleges across all semester periods. Particularly in the early weeks (Weeks 0-2) and mid-semester (Weeks 7-8, including breaks), these differences are most pronounced. The consistent negative Cohen's (1992) d values indicate that HAS generally exhibits lower interaction counts than other colleges, such as MPH and NHS.

This trend is consistently observed across all semester blocks, highlighting a potential disparity in engagement or resource utilisation within the HAS college compared to others. The analysis of interaction patterns suggests that there may be underlying factors influencing student engagement in the LMS across different colleges.

4.2.4. Average student activity across daily time periods

Average interaction values were calculated grouped by time of day and grade as shown in Figure 20. The results indicate a correlation between activity levels, grade outcomes, and the time of day across the semester.

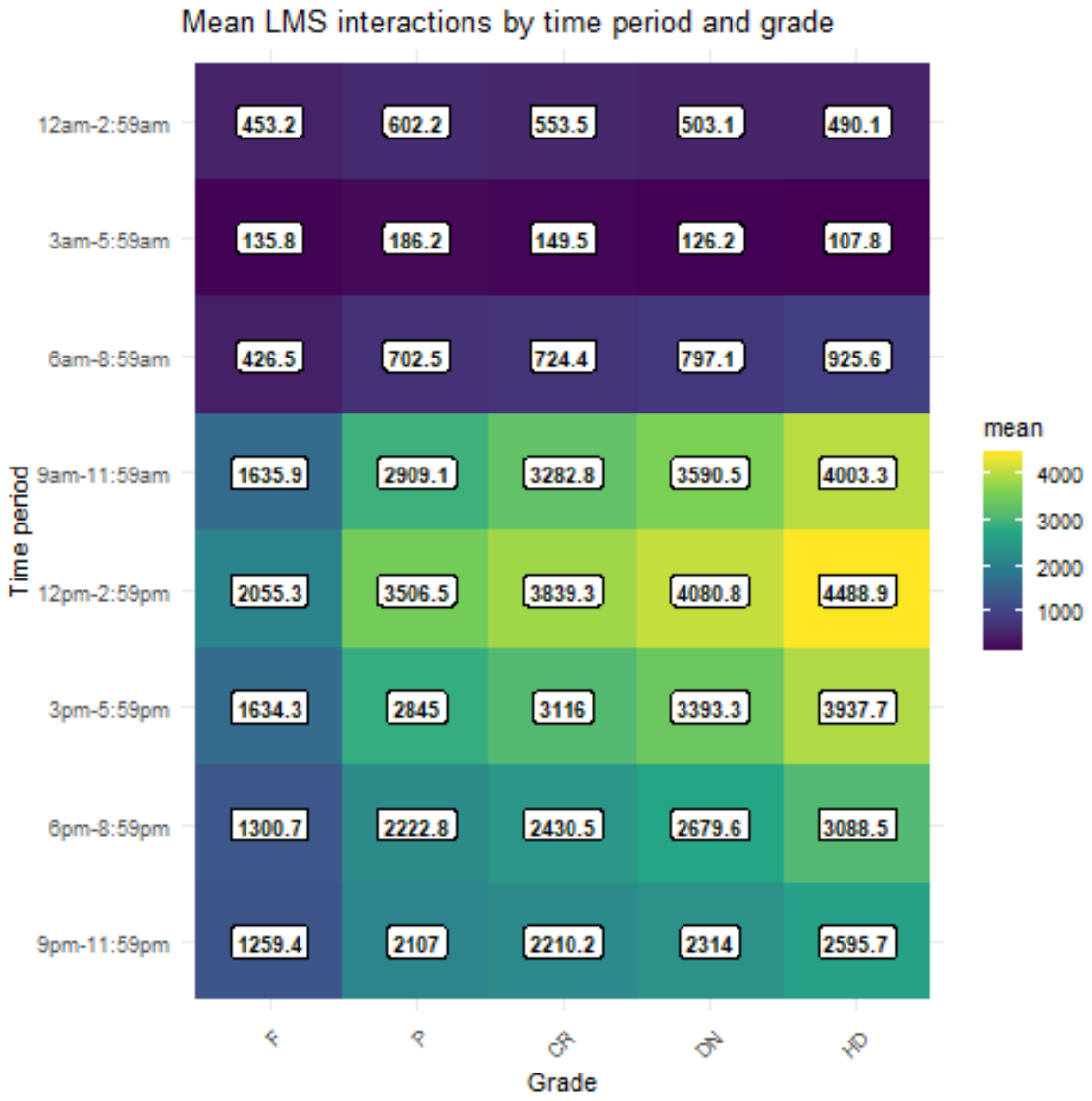

Figure 20 - Average interactions by time period and grade

The Anderson-Darling (1952) test showed an extremely low p-value of 4.579e-15, strongly suggesting that the time data for each college is not normally distributed. Again, the Kruskal-Wallis (1952) test was performed, resulting in a Chi-squared of 35.064, 5 degrees of freedom, and a p-value of 1.461e-06. Similar to the semester periods, the results suggest significant differences across colleges, requiring the post-hoc pairwise comparisons via Dunn's (1964) test to determine which pairs of colleges differ significantly.

In a similar fashion to the semester interaction patterns shown in Figure 19, there is a consistent increases in grades with increasing interactions across time periods (Figure 21), with the 9am to 6pm being consistently the peak of activity across all colleges. Again, like Figure 19, the time of days for each college show distinct patterns of increases in grades across time periods.

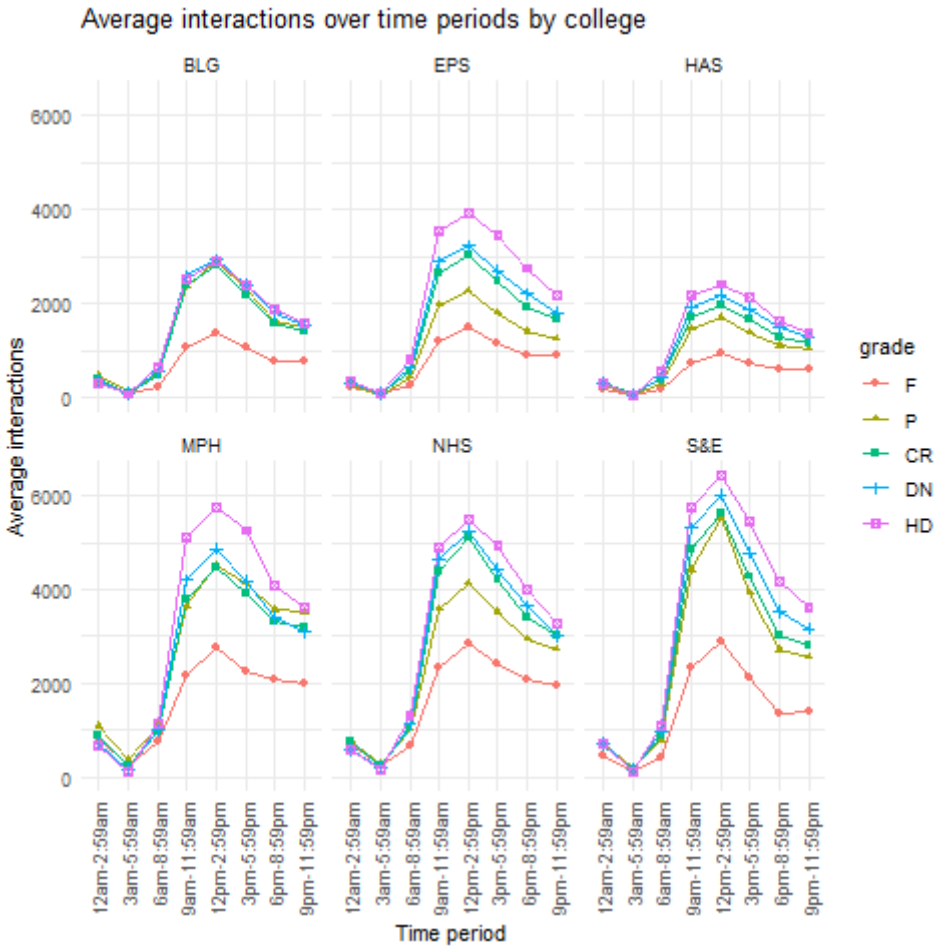

Figure 21 - Average interactions over time period by grade for each college

4.2.5. LMS components

Figure 22 visualises a Pearson (1895) correlation matrix, revealing the strengths of relationships between the percentage viewed of each LMS component and the corresponding numerical grade values.

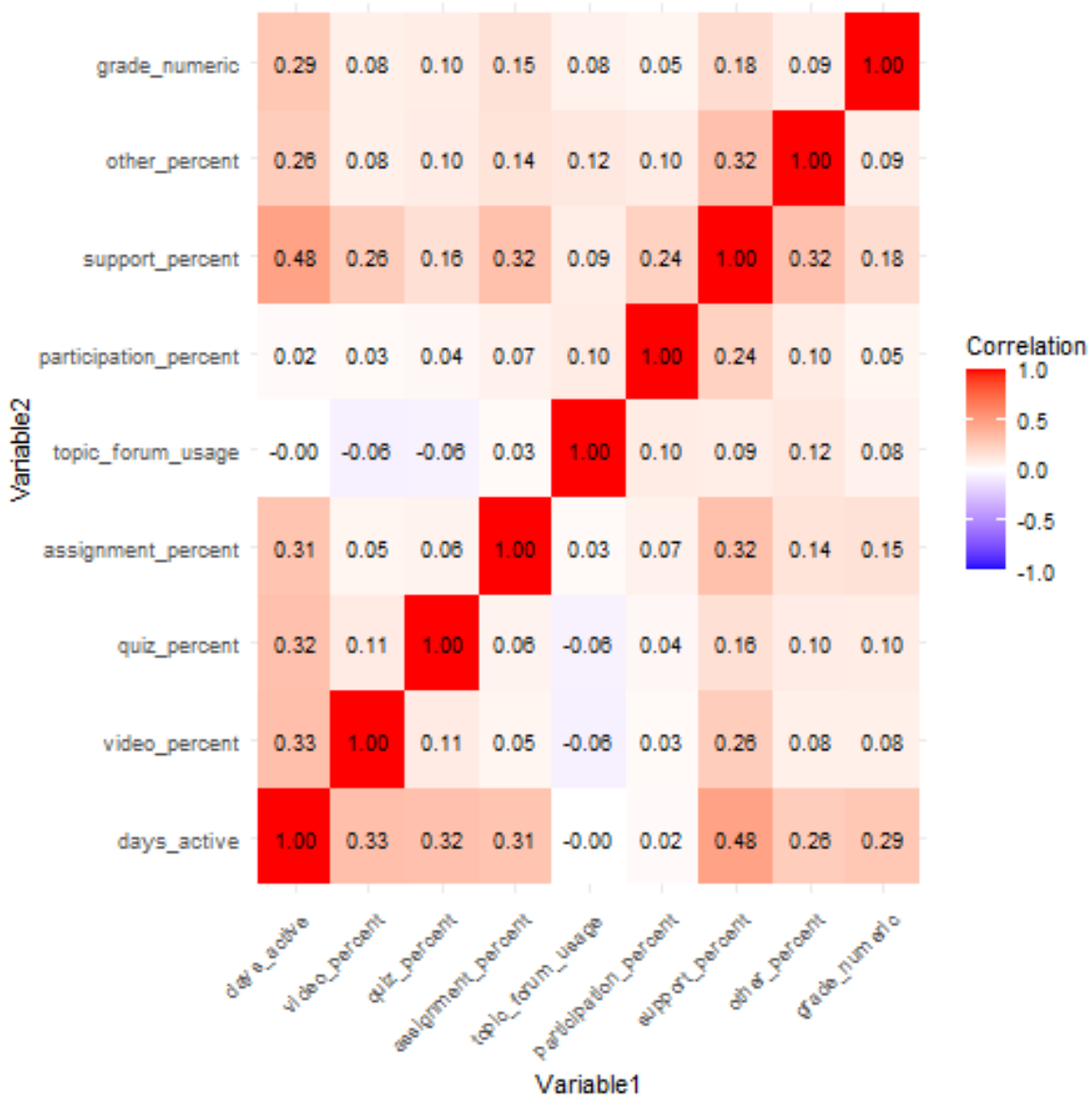

Figure 22 - Attribute correlation matrix

As can be seen in Figure 22, there is only one moderate correlation of 0.48 (days active and percentage of support modules), with the rest of the correlations being weak to very weak. There are several weak correlations primarily focusing on the percentage of support modules viewed. Interestingly, there are several very weak negative correlations, focused on the topic forum usage (the percentage of total posts in the topic, that the student participated in). However, overall, the grade correlation across attributes is low, with slightly stronger correlations with percentage of support and assignment modules viewed.

The results from the Anderson-Darling (1952) test in Table 17 show that for each college, the p-value is extremely low (all being 3.7e-24). This would strongly suggest that the interaction data for each college is not normally distributed.

Table 17 - LMS components Anderson-Darling test

Metric	**P-Value**
Interactions	3.70E-24
Days Active	3.70E-24
Percentage of Videos	3.70E-24
Percentage of Quizzes	3.70E-24
Percentage of Assignments	3.70E-24
Percentage of Participation Modules	3.70E-24
Percentage of Support Modules	3.70E-24
Percentage of 'Other' Modules	3.70E-24
Forum Usage (Percentage of Topic)	3.70E-24
Forum Posts (Number per Student)	3.70E-24

The results from Kruskal-Wallis (1952) test shown in Table 18 show that they are all highly significant. The low p-value (less than 2.2e-16) indicates a statistically significant difference in the distributions of interactions across colleges. With the Kruskal-Wallis (1952) test showing significant differences across colleges for each metric, the next test to perform is post-hoc pairwise comparisons via Dunn's (1964) test to determine which pairs of colleges differ significantly.

Table 18 - LMS components Kruskal-Wallis test

Metric	Chi-Squared	P-Value
Interactions	22747	<2.2e-16
Days Active	11032	<2.2e-16
Percentage of Videos	6807.2	<2.2e-16
Percentage of Quizzes	15234	<2.2e-16
Percentage of Assignments	2874.6	<2.2e-16
Percentage of Participation Modules	13393	<2.2e-16
Percentage of Support Modules	8424.4	<2.2e-16
Percentage of 'Other' Modules	14223	<2.2e-16
Forum Usage (Percentage of Topic)	1183.8	<2.2e-16
Forum Posts (Number per Student)	7087.5	<2.2e-16

The results Table 19 show only the non-significant pairings, with the remaining comparisons from Dunn's (1964) test all exhibiting statistically significant differences in the interaction distributions between those pairs. The results show both the original p-values and the p-values adjusted for multiple comparisons. The adjustment methodology utilises the Bonferroni method, a widely used statistical method to adjust for multiple comparisons mentioned by Dunn (1961).

Table 19 - LMS components summary of non-significant results (Dunn's test)

Content Type	Comparisons	Z Value	P Value	P Adjusted	Significance
Video Percent	BGL - MPH	-0.39048	0.3480907	1.00E+00	Not Significant
Assignment Percent	BGL - MPH	2.585367	0.004863759	7.30E-02	Not Significant
Participation Percent	MPH - S&E	-2.22876	0.01291481	1.94E-01	Not Significant
Forum Activity	BGL - S&E	-1.75824	0.0393537	5.90E-01	Not Significant

The consistent pattern of significant differences across all college pairings would suggest that the distribution of interactions varies significantly from one college to another, which may reflect differences in student engagement, course content, teaching methodologies, or other factors unique to each college. However, there is no direct indication in the data of these factors from the student LMS usage information.

4.2.6. Topic content

Figure 23 shows the total number of video type components for each individual topic, with the x-axis representing individual topics, and the y-axis representing the count of unique videos found for each topic. As can be seen in Figure 23, topics vary widely in number of video type components both between college and between topics. Additionally, there are instances of topics with no video type components. Remaining topic composition attributes can be found in Section 7.1 Appendix A: Additional tables and figures.

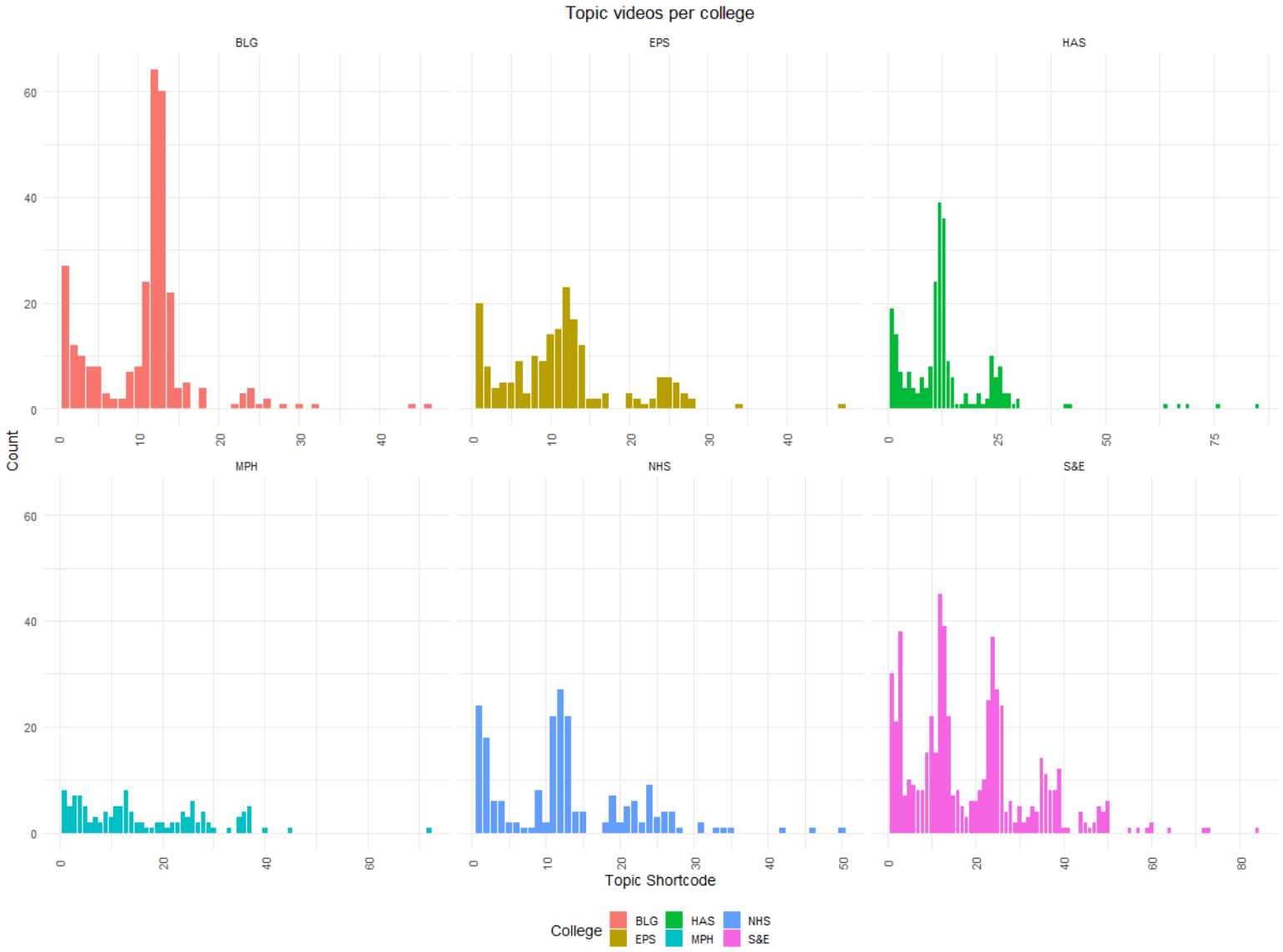

Figure 23 - Topic videos grouped by college

Table 20 summarises the total number of unique video components per college, as well as mean and max number of videos. This shows a great variation in overall number of video type components, and the average number per topic, as well as the maximum any topic in that college has.

Table 20 - Topic videos grouped by college

college	video sum	video avg	video max
BGL	3,120	4.66	46
EPS	2,224	3.19	47
HAS	3,297	2.76	85
MPH	1,973	4.79	72
NHS	2,524	4.37	50
S&E	10,773	13.03	84

Videos was chosen to display in the thesis body, while additional visualisations are shown in Appendix A: Additional tables and figures, due to the space required for each component type visualisation.

Results from the Anderson-Darling (1952) test are shown in Table 21, these show that for each college grouping of topics, the p-value is extremely low (all being 2.2e-16). This would strongly suggest that the topic construction (number of each of the identified LMS components) between each college is not normally distributed.

Table 21 - Topic grouped LMS components Anderson-Darling test

Component	p-value
Lecture Videos	2.2e-16
Quiz Count	2.2e-16
Assignment Count	2.2e-16
Forum Activity	2.2e-16
Participation Activity Count	2.2e-16
Support Materials Count	2.2e-16
Other Interactions Count	2.2e-16

The Kruskal-Wallis (1952) test results shown in Table 22 show that they are all highly significant, indicating a statistically significant difference in the structure of topics across colleges.

Table 22 - Topic grouped LMS components Kruskal-Wallis test

Component	Chi-Squared	df	p-value
Lecture Videos	635.32	5	< 2.2e-16
Quiz Count	250.69	5	< 2.2e-16
Assignment Count	276.23	5	< 2.2e-16
Forum Activity	265.63	5	< 2.2e-16
Participation Activity Count	226.89	5	< 2.2e-16
Support Materials Count	381.49	5	< 2.2e-16
Other Interactions Count	139.45	5	< 2.2e-16

Finally, the results of Dunn's (1964) test in the following tables show each pairing exhibiting statistically significant differences across all components for each college. The Z-scores in Dunn's test results represent the standardised difference between the ranks of the two colleges compared. A positive Z-score suggests that the first group tends to have higher values (e.g., a higher number of videos) than the second group. Conversely, a negative Z-score indicates that the first college tends to have lower values than the second college.

Videos components

The results shown in Table 23 identify significant negative Z-scores when comparing BGL, EPS, and HAS against S&E. This indicates that S&E topics typically have a larger number of lecture videos compared to BGL, EPS, and HAS. Additionally, positive Z-scores for BGL compared to EPS and HAS suggest that BGL tends to have a larger number of videos compared to EPS and HAS. All relationships shown are statistically significant, suggesting that both S&E and BGL have a large number of videos, with S&E having the most of the two colleges.

Table 23 - Topic video components summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - EPS	5.339001	4.673e-08	7.0095e-07
BGL - HAS	8.526007	7.5743e-18	1.136145e-16
BGL - MPH	3.328797	0.000436	0.006542
HAS - MPH	-3.55853	0.000186	0.002797
HAS - NHS	-5.41925	2.992489e-08	4.488734e-07
BGL - S&E	-12.8196	6.36737e-38	9.551055e-37
EPS - S&E	-18.5896	1.951343e-77	2.927014e-76
HAS - S&E	-23.8336	7.497187e-126	1.124578e-124
MPH - S&E	-14.5116	5.11752e-48	7.67628e-47
NHS - S&E	-14.8154	5.82698e-50	8.74047e-49

Quiz components

The results shown in Table 24 identify negative Z-scores when comparing BGL, EPS, and HAS against S&E, suggesting that S&E topics typically have a larger number of quizzes compared to BGL, EPS, and HAS. Conversely, positive Z-scores when comparing S&E to MPH and NHS indicate that MPH and NHS tend to have a larger number of quizzes compared to S&E. All relationships shown are statistically significant, suggesting that MPH and NHS have a larger number of quizzes than S&E, while S&E has a larger number of quizzes than BGL, EPS, and HAS.

Table 24 - Topic quiz components summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - MPH	-9.31472	6.113802e-21	9.170703e-20
EPS - MPH	-10.2277	7.452461e-25	1.117869e-23
HAS - MPH	-10.7799	2.141781e-27	3.212671e-26
BGL - NHS	-8.48357	1.091908e-17	1.637862e-16
EPS - NHS	-9.49205	1.132631e-21	1.698946e-20
HAS - NHS	-10.1493	1.668571e-24	2.502856e-23
BGL - S&E	-6.3728	9.280261e-11	1.392039e-09
EPS - S&E	-7.46058	4.307059e-14	6.460588e-13
HAS - S&E	-8.04556	4.29278e-16	6.43917e-15
MPH - S&E	4.178207	1.46908e-05	0.00022
NHS - S&E	2.776503	0.002747	0.04121

Assignment components

The results shown in Table 25 identify positive Z-scores when comparing BGL and EPS against HAS, suggesting that BGL and EPS typically have a larger number of assignments compared to HAS. Conversely, negative Z-scores when comparing BGL, EPS, and HAS to S&E indicate that S&E tends to have a larger number of assignments compared to BGL, EPS, and HAS. All relationships shown are statistically significant, suggesting that S&E has a larger number of assignments compared to the other colleges, while BGL and EPS have more than HAS.

Table 25 - Topic assignment components summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - HAS	5.458624	2.399199e-08	3.598799e-07
EPS - HAS	5.043069	2.290615e-07	3.435923e-06
BGL - MPH	-6.41868	6.872897e-11	1.030935e-09
EPS - MPH	-6.84572	3.804542e-12	5.706813e-11
HAS - MPH	-11.6484	1.169687e-31	1.754531e-30
BGL - NHS	-7.07167	7.65381e-13	1.148072e-11
EPS - NHS	-7.55544	2.08716e-14	3.130741e-13
HAS - NHS	-13.1227	1.220146e-39	1.83022e-38
BGL - S&E	-4.71946	1.182383e-06	1.773575e-05
EPS - S&E	-5.22862	8.53923e-08	1.280884e-06
HAS - S&E	-11.2505	1.151266e-29	1.726898e-28
NHS - S&E	2.88264	0.001972	0.029577

Forum activity

The results shown in Table 26 identify positive Z-scores when comparing BGL and EPS against HAS, suggesting that BGL and EPS typically have a higher level of forum activity compared to HAS. Additionally, a negative Z-score when comparing HAS to S&E indicates that S&E tends to have a higher level of forum activity compared to HAS. All relationships shown are statistically significant, suggesting that S&E has a higher level of forum activity compared to HAS, while BGL and EPS are more active than HAS.

Table 26 - Topic forum posts summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - HAS	8.37147	2.845191e-17	4.267786e-16
EPS - HAS	10.47532	5.608018e-26	8.412027e-25
HAS - MPH	-9.9072	1.93678e-23	2.905171e-22
BGL - NHS	-5.67993	6.737677e-09	1.010652e-07
EPS - NHS	-4.04981	2.562988e-05	0.000384
HAS - NHS	-14.3379	6.337388e-47	9.506082e-46
EPS - S&E	3.471937	0.000258	0.003875
HAS - S&E	-7.08833	6.787196e-13	1.018079e-11
MPH - S&E	4.069482	2.355886e-05	0.000353
NHS - S&E	7.490695	3.425483e-14	5.138225e-13

Participation components

Results shown in Table 27 identify a negative Z-score when comparing BGL to EPS, suggesting that EPS typically has a higher level of participation activity compared to BGL. Conversely, positive Z-scores when comparing BGL, EPS, and HAS to S&E indicate that BGL, EPS, and HAS tend to have higher levels of participation activities compared to S&E. All relationships shown are statistically significant, suggesting that BGL, EPS, and HAS are more active in participation compared to S&E, while EPS is more active than BGL.

Table 27 - Topic participation component summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - EPS	-8.16089	1.662798e-16	2.494196e-15
BGL - HAS	2.988737	0.001401	0.02101
EPS - HAS	12.29514	4.809706e-35	7.214559e-34
EPS - MPH	9.112517	4.024605e-20	6.036908e-19
EPS - NHS	9.085396	5.166099e-20	7.749149e-19
BGL - S&E	5.380463	3.714728e-08	5.572092e-07
EPS - S&E	14.034	4.827512e-45	7.241268e-44
HAS - S&E	2.992878	0.001382	0.020727
NHS - S&E	3.874092	5.351142e-05	0.000803

Support components

Results shown in Table 28 identify positive Z-scores when comparing BGL and EPS against HAS, suggesting that BGL and EPS typically provide more support materials compared to HAS. Conversely, a negative Z-score when comparing HAS to S&E indicates that S&E tends to provide more support materials compared to HAS. All relationships shown are statistically significant, suggesting that S&E provides more support materials compared to HAS, while BGL and EPS provide more than HAS.

Table 28 - Topic support component summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - HAS	10.32567	2.697097e-25	4.045646e-24
EPS - HAS	9.725972	1.168318e-22	1.752476e-21
HAS - MPH	-10.4737	5.703435e-26	8.555152e-25
BGL - NHS	-4.95761	3.568281e-07	5.352421e-06
EPS - NHS	-5.63269	8.870984e-09	1.330648e-07
HAS - NHS	-15.3901	9.539272e-54	1.430891e-52
MPH - NHS	-2.82009	0.002401	0.036008
BGL - S&E	-4.50085	3.384043e-06	5.076065e-05
EPS - S&E	-5.23947	8.051921e-08	1.207788e-06
HAS - S&E	-16.1947	2.74843e-59	4.122645e-58

'Other' components

Finally, results shown in Table 29 identify negative Z-scores when comparing BGL to EPS and HAS, suggesting that EPS and HAS typically have higher levels of other interactions compared to BGL. Conversely, positive Z-scores when comparing BGL and HAS to S&E indicate that BGL and HAS tend to have higher levels of other interactions compared to S&E. All relationships shown are statistically significant, suggesting that BGL and HAS are more engaged in other interactions compared to S&E, while EPS and HAS are more engaged than BGL.

Table 29 - Topic 'other' component summary of Dunn's test results

Comparisons	Z Value	P Value	P Adjusted
BGL - EPS	-5.41274	3.103386e-08	4.655078e-07
BGL - HAS	-4.19694	1.352705e-05	0.000203
BGL - MPH	-2.91453	0.001781	0.026717
BGL - NHS	-10.918	4.727424e-28	7.091136e-27
EPS - NHS	-5.81957	2.949932e-09	4.424897e-08
HAS - NHS	-8.23501	8.977391e-17	1.346609e-15
MPH - NHS	-6.78719	5.716968e-12	8.575451e-11
BGL - S&E	-7.41238	6.202506e-14	9.303759e-13
HAS - S&E	-4.03825	2.692573e-05	0.000404
MPH - S&E	-3.36499	0.000383	0.005741
NHS - S&E	4.329756	7.463727e-06	0.000112

Analysis of results

The College of Science and Engineering (S&E) prominently features in many of the statistically significant comparisons, particularly in lecture videos, quiz counts, and assignment counts. This trend indicates that S&E generally provides a larger volume of content in these areas compared to other colleges. The analysis reveals considerable variations in topic construction across colleges, with distinct preferences in the number of components utilised.

4.2.7. Exploratory data analysis summary

Grade Distribution Variability

The data showed significant variation in grade distribution across colleges, suggesting potential differences in course difficulty, teaching quality, or assessment methods. For example, colleges like the College of Nursing and Health Sciences showed fewer Fail grades than expected, while the College of Science and Engineering showed a greater number of Fail grades while at the same time a greater number of High Distinctions (HD). This variation suggests differences in high achiever distribution or grading standards across colleges.

Student Attendance and LMS Activity

Marked variations were observed in student activity within the LMS, categorised by college. This could indicate a significant association between the colleges and levels of student engagement with the LMS. For instance, colleges such as NHS and S&E showed more students in the 'Medium' and 'High' activity categories compared to BGL, which had a concentration in the 'Low' activity category. These differences across colleges hint at disparities in student engagement or resource utilisation.

Average Student Activity Across the Semester

Significant differences in interaction patterns were noted across colleges throughout the semester. The college of Humanities, Arts, and Social Sciences (HAS), for example, demonstrated notable differences in interaction patterns compared to other colleges, especially in the early weeks and mid-semester, suggesting different engagement or resource utilisation strategies.

Average Student Activity Across Daily Time Periods

The data indicates varying levels of LMS engagement across colleges during different 24-hour periods. Notable trends include fewer students with low activity levels than expected in colleges like MPH, contrasting sharply with colleges like BGL where low activity levels were more prevalent, reflecting potential variations in course structure or student engagement strategies.

LMS Components Analysis

The analysis compared the usage of various LMS components with grade outcomes and revealed weak to very weak correlations. For instance, while the percentage of support modules viewed showed some correlation with grades, most other components, like forum usage, showed very weak or even negative correlations. This suggests that while the predictability of grades using LMS components appears weak, significant differences in how colleges engage with these components indicate varied approaches to LMS utilisation.

Topic Content Analysis

Significant differences were observed in the structure of topics across colleges, with S&E often providing more extensive content in areas such as lecture videos and quizzes. For example, the number of video components in S&E topics was notably higher than in other colleges like BGL or HAS, reflecting differences in educational focus and resource allocation.

In Conclusion, the findings underscore significant variability in grade distribution, student engagement with the LMS, and interaction patterns across colleges. The weak correlation between LMS usage and grades contrasts with the pronounced differences in how colleges utilise LMS components, pointing to varied pedagogical strategies and resource allocations. The College of Science and Engineering stands out for its extensive provision of content, while the variability in topic construction across colleges highlights a diversity in educational priorities and methodologies. The differences in student engagement and interaction patterns across both daily and semester periods underscore the unique engagement strategies and potentially different pedagogical approaches among colleges. A summary of differences found between colleges can be seen in Table 30, these differences include grade distribution, average days active, interaction patterns in the early weeks of the semester, LMS usage, and volume of topic content.

Table 30 - Summary of general college differences

College	Grade distribution	Avg. activity days	Interaction patterns in early weeks	LMS components usage	Topic content volume
BGL	Greater Fail.	Predominantly low activity could point to engagement challenges or external commitments.	Fewer early semester interactions might reflect slower course start or student adaptation time.	Varied usage.	Lower video content suggests a preference for traditional or text-based learning materials.
	Fewer HD.				
EPS	Fewer Fail.	Moderate activity levels suggest a balanced online engagement among students.	Varied interaction patterns.		Moderate video content levels might reflect a balanced approach to multimedia and traditional teaching methods.
	Fewer HD.				
HAS	Moderate distribution.		Significant interaction differences (high and low) particularly in early weeks and mid-semester might reflect unique course structures or intensive periods.	Moderate usage of LMS components suggests a balanced, possibly traditional approach to online learning.	
MPH	Fewer Fail.	High activity levels indicate strong student engagement and interaction with online resources.	Higher early semester interactions might reflect an intensive start to courses or proactive student behaviour.	High usage of LMS components suggests an extensive and diverse use of online learning tools, possibly reflecting interactive or tech-savvy course designs.	Higher video content indicates a strong emphasis on multimedia and visual learning resources.
	Greater HD.				
NHS	Fewer Fail.	Moderate to high activity levels suggest varying degrees of student engagement.	Varied interaction patterns.	Moderate to high usage of LMS components indicates a comprehensive approach to online learning, possibly with a focus on interactive or supportive materials.	
	Greater HD.				
S&E	Greater Fail.	High activity levels indicate strong student engagement and interaction with online resources.	High interaction levels throughout the semester suggest an intensive, continuous learning approach.	High usage of LMS components indicates an extensive and varied approach to online learning, possibly reflecting innovative teaching methods or tech-integrated courses.	Highest video content levels indicate a strong emphasis on multimedia resources.
	Greater HD.				

4.3. Principal Component Analysis (PCA)

4.3.1. Variance explained

The total variance explained by each of the principal components identified is shown in the scree plot in Figure 24. The plot suggested that the variance explained diminishes significantly after a small number of components, with the first component accounting for the largest variance proportion.

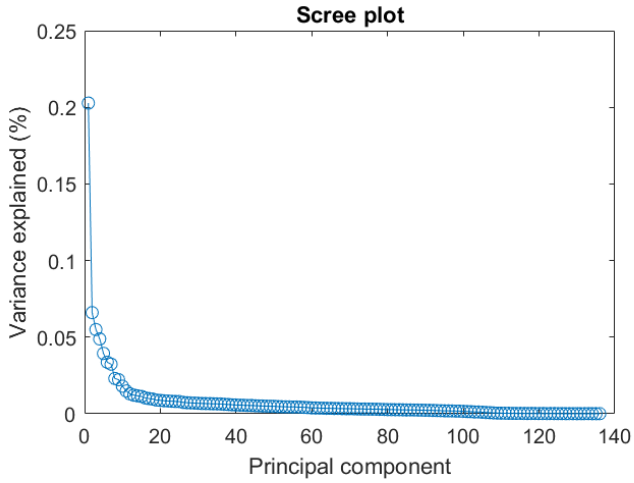

Figure 24 - Scree plot showing dataset variance explained by each principal component

A cumulative variance plot (Figure 25) was subsequently developed to determine the number of components that capture a substantial amount of variance, guiding further analysis.

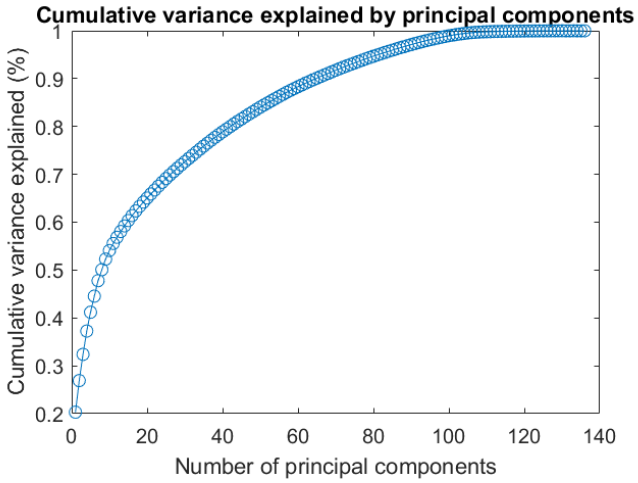

Figure 25 - Cumulative variance plot of number of components

As shown in Figure 25, there is a steep increase followed by a gradual plateau in the cumulative variance explained by the total number of components. Figure 25 shows for the first ten components 20%, 26%, 32%, 37%, 41%, 44%, 47%, 50%, 52%, and 54% variance explained respectively. While at the higher end of variance explained; 90%, 95%, and 99% variance explained would require 65, 82, and 101 components respectively.

4.3.2. Visualising principal components

In Figure 26, principal component one appears to consist of overall LMS engagement across time and content. There are high loadings (negative) across various total semester periods, total periods, and days active attributes. This component seems to capture general engagement across different times of the day (total periods) and weeks of the semester (total semester periods). With negative loadings indicating that a higher scores on this component correspond to lower overall interactions. Which might be an indicator of overall activity or lack of activity in the LMS.

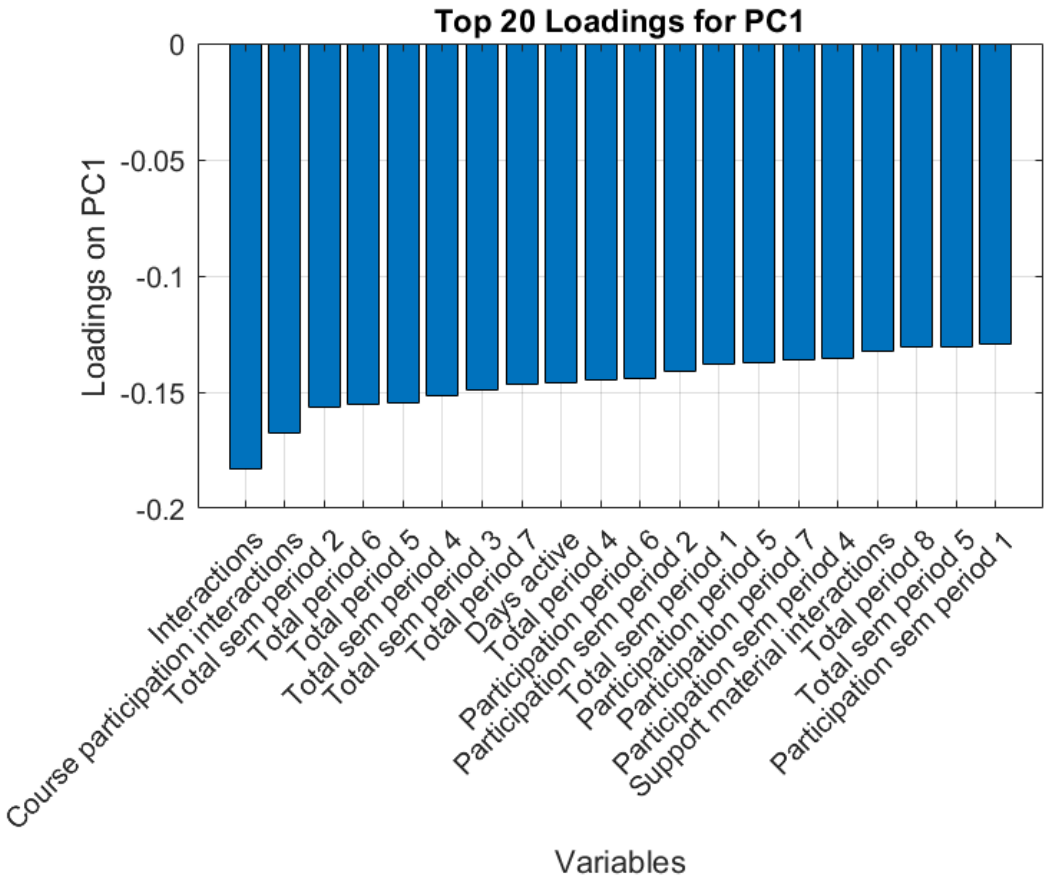

Figure 26 - Top 20 loading attributes for principal component one

In Figure 27, principal component two shows a contrast between video engagement and assignment interactions, with high positive loadings for video interactions and specific video period attributes, and negative loadings for assignment interactions. This component likely represents a contrast between video consumption and assignment interaction. With high scores potentially indicating a preference for video-based learning during specific times of the day or periods of the semester, contrasting with assignment activities.

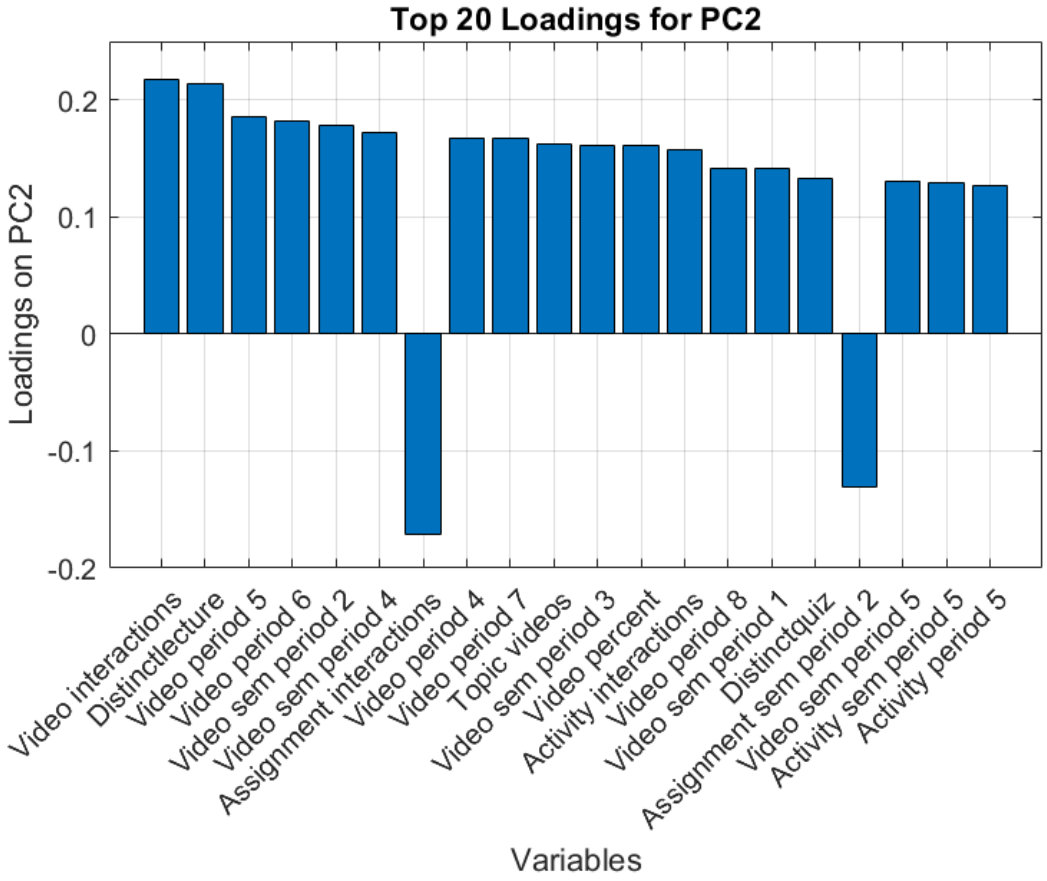

Figure 27 - Top 20 loading attributes for principal component two

In Figure 28, the third component appears to be primarily based on 'Other' types of interactions, with high negative loadings for other interactions and other semester periods. The negative loadings suggest that this component inversely relates to these types of interactions.

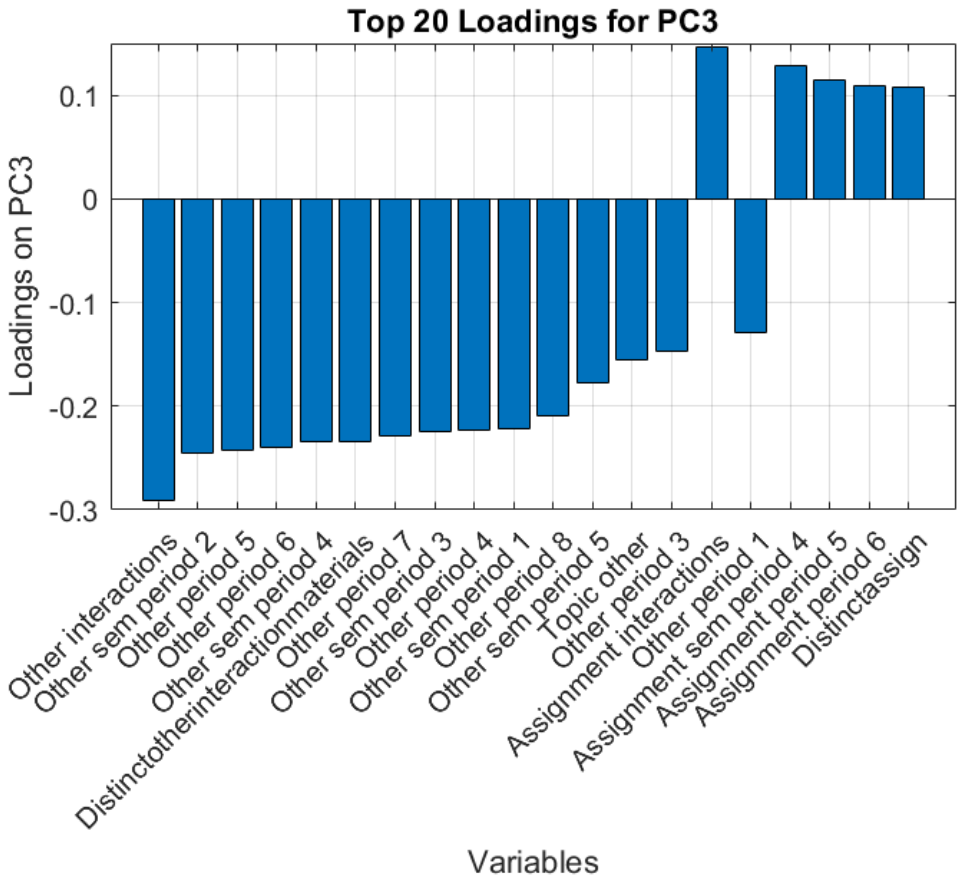

Figure 28 - Top 20 loading attributes for principal component three

In Figure 29, principal component four shows a balance between videos and active participation, showing negative loadings for video interactions and positive loadings for activity interactions. Suggesting that this component might capture a balance between passive (watching videos) and active (forum activities, quizzes) learning behaviours, especially considering the time of day and periods of the semester.

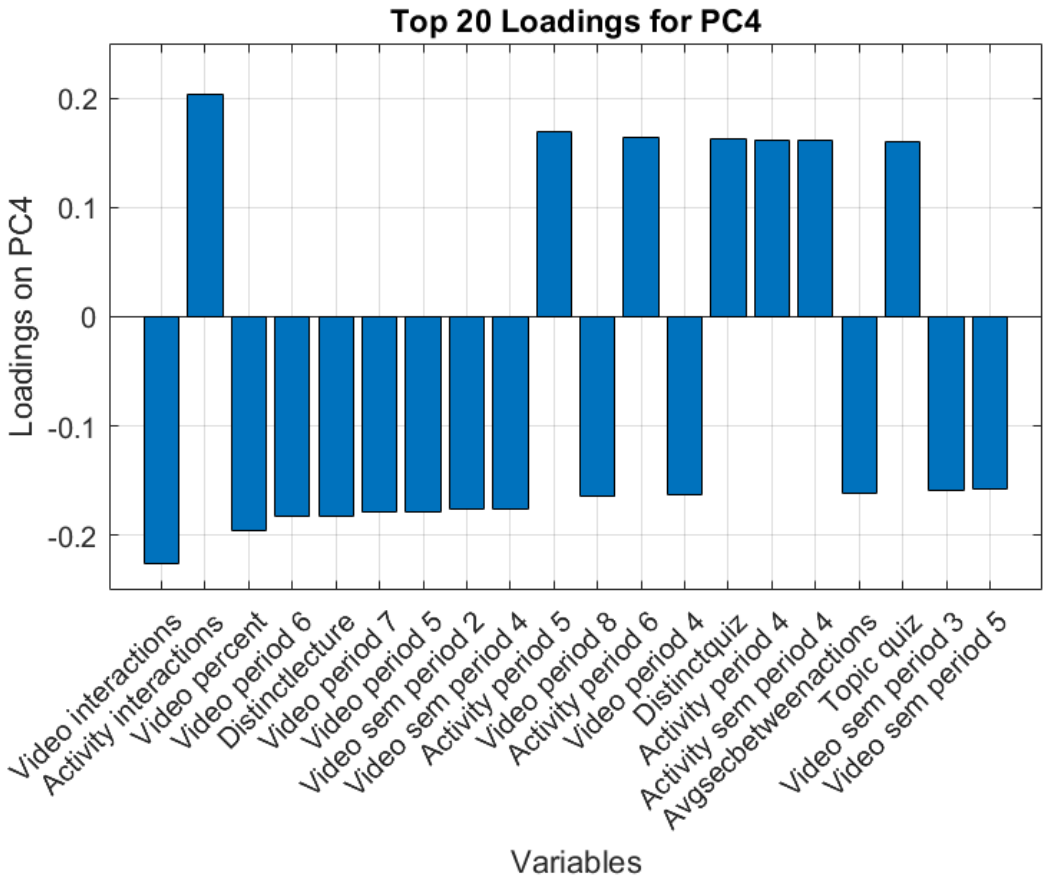

Figure 29 - Top 20 loading attributes for principal component four

In Figure 30, principal component five shows a comparison between support material usage against assignment engagement, with positive loadings for support material interactions and negative loadings for assignment interactions. This component suggests a trade-off between the use of LMS support materials and engagement with assignments. It might represent different learning strategies or preferences for support materials over direct assignment interaction.

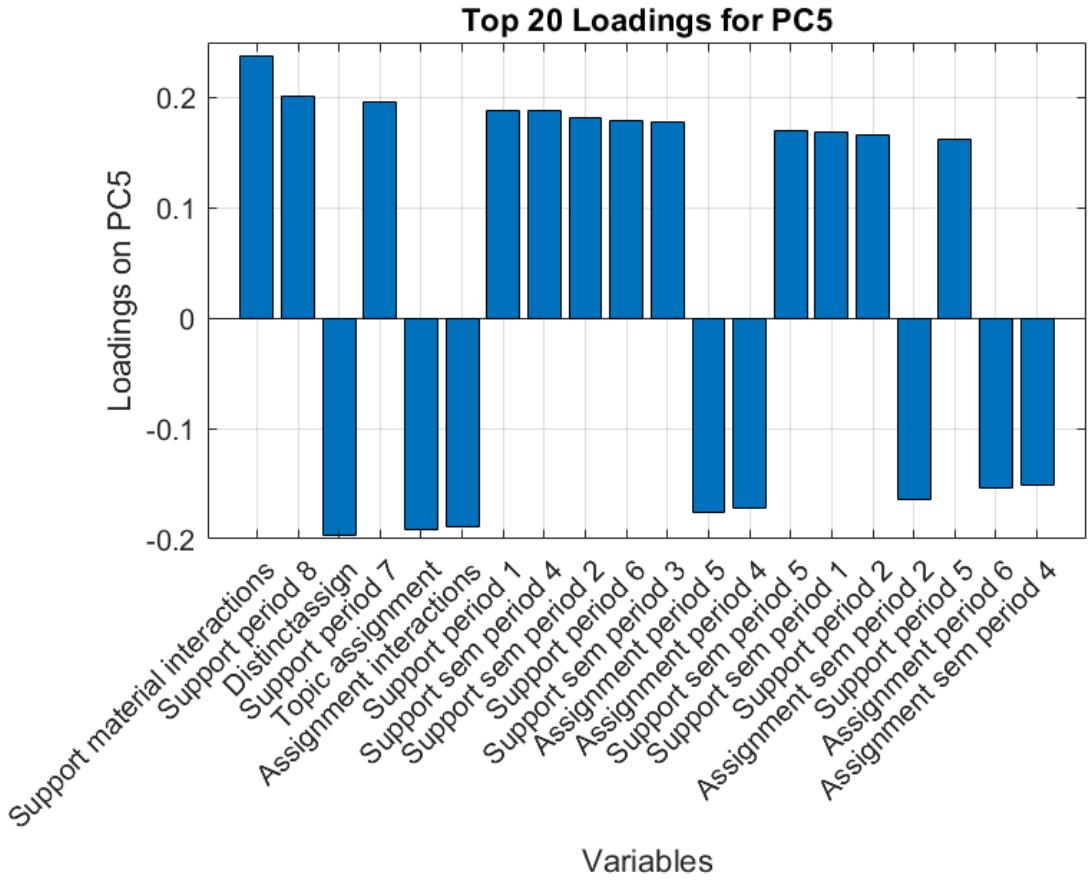

Figure 30 - Top 20 loading attributes for principal component five

In Figure 31, principal component six mainly concerns social interactions, with mostly negative loadings for social interactions during different periods. This component seems to reflect the social aspect of LMS usage, capturing how social interactions vary across different times and periods of the semester. Higher scores may indicate less social engagement.

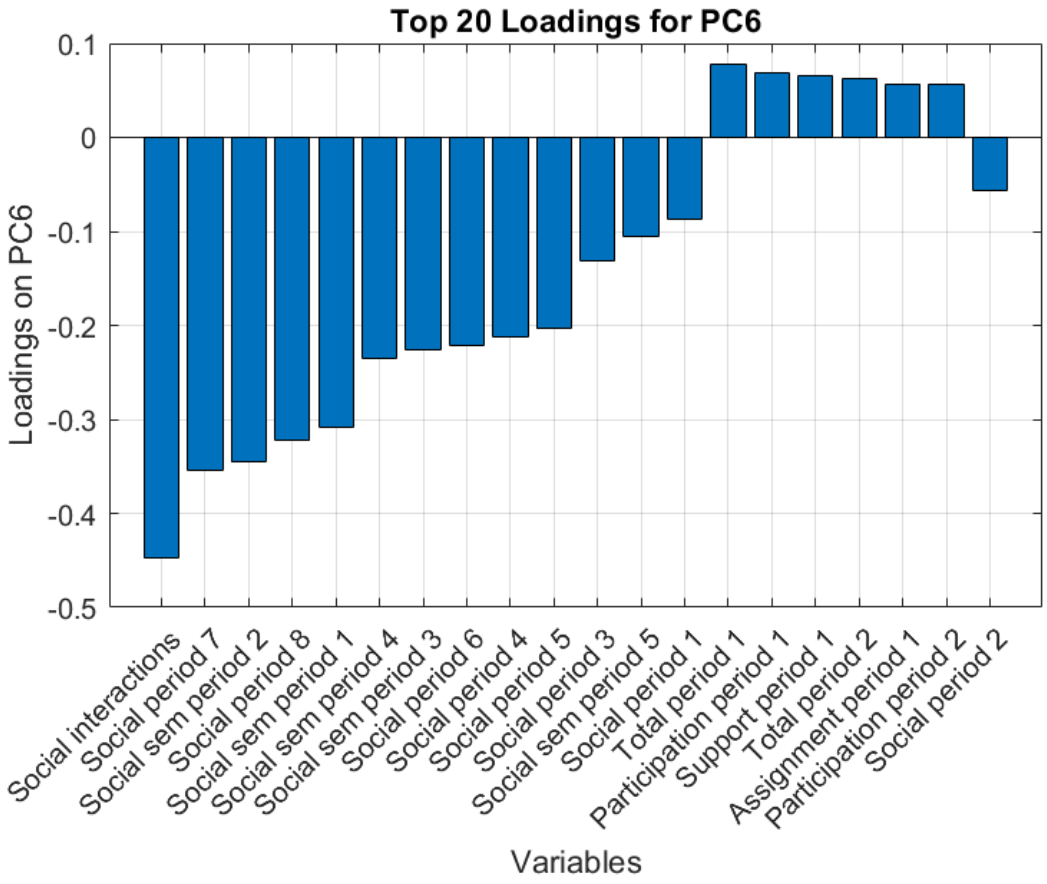

Figure 31 - Top 20 loading attributes for principal component six

In Figure 32, principal component seven shows a shift from early engagement to support material use, with negative loadings for early Total period and Participation period, and positive loadings for support material in middle periods. The component may indicate a shift in student focus from early participation and general engagement to later reliance on support materials, reflecting a change in learning strategy or focus as the semester progresses.

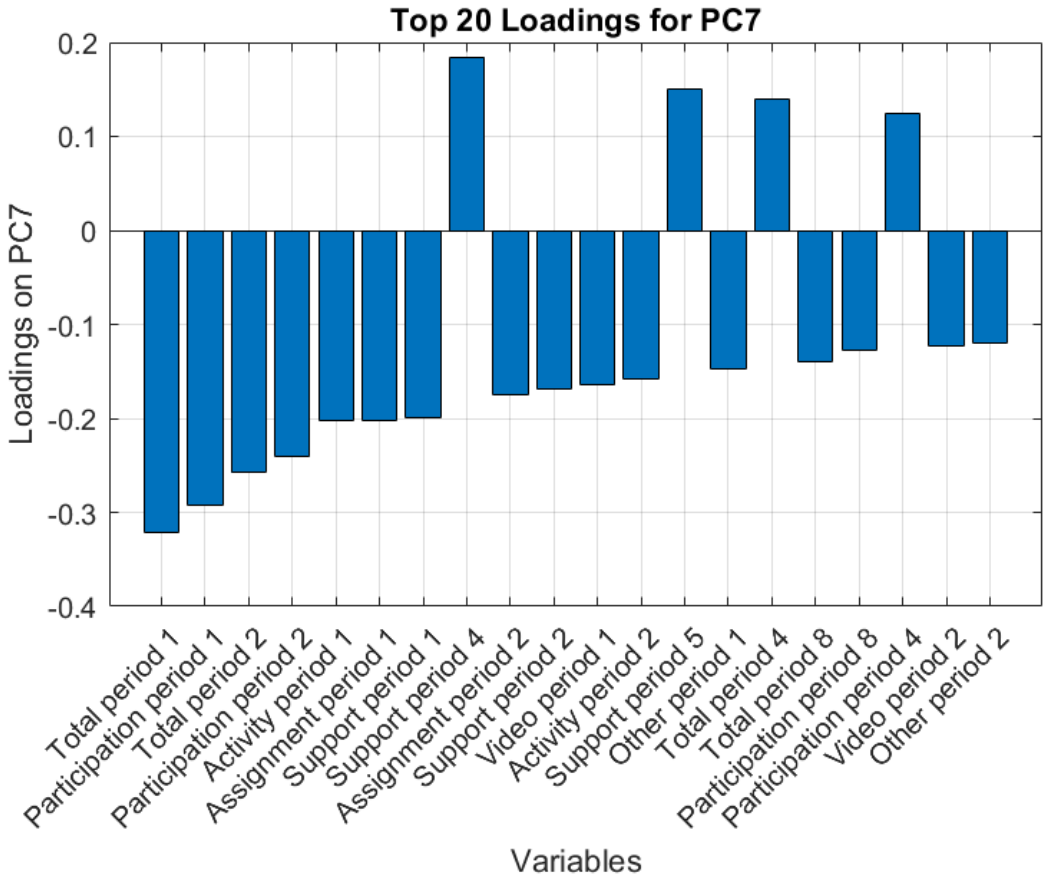

Figure 32 - Top 20 loading attributes for principal component seven

In Figure 33, principal component eight shows a temporal shift in focus and engagement, with positive loadings for total period 3 and participation period 3, and negative for support period 8. This component might represent a shift in student focus and engagement, with early emphasis on general and participation interactions and a move away from support materials later in the semester.

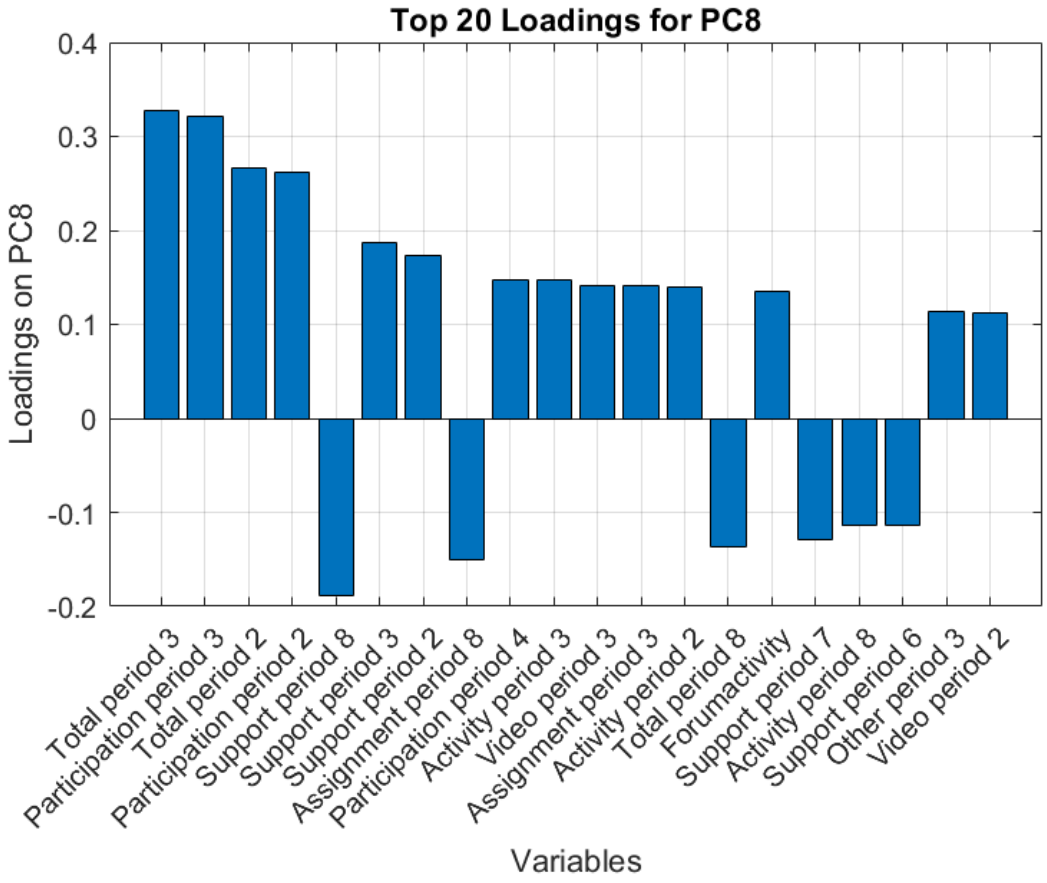

Figure 33 - Top 20 loading attributes for principal component eight

In Figure 34, principal component nine shows a balance between participation and support, with negative loadings on participation during later periods, positive loadings on support material in mid time periods. Principal component nine seems to capture a balance between participation and support material usage, possibly indicating different phases or strategies in the learning process.

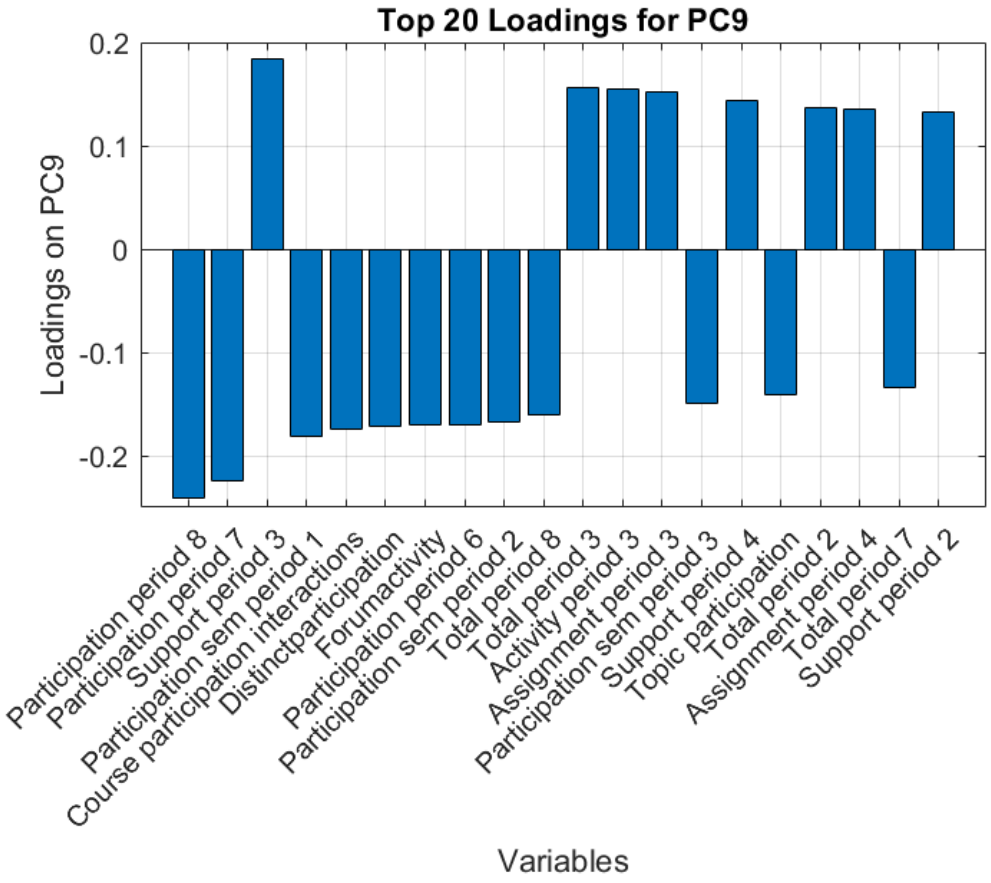

Figure 34 - Top 20 loading attributes for principal component nine

The last principal component that will be analysed, principal component ten (Figure 35), shows a comparison between general engagement and topic composition. With negative loadings on topic support and topic participation, and positive on participation percent and support percent. This component might reflect a balance between general engagement in the LMS and the construction of certain topics. Negative loadings on specific topic compositions contrasted with positive loadings on general participation percentages suggest a distinction between broad engagement and specific compositions of topics.

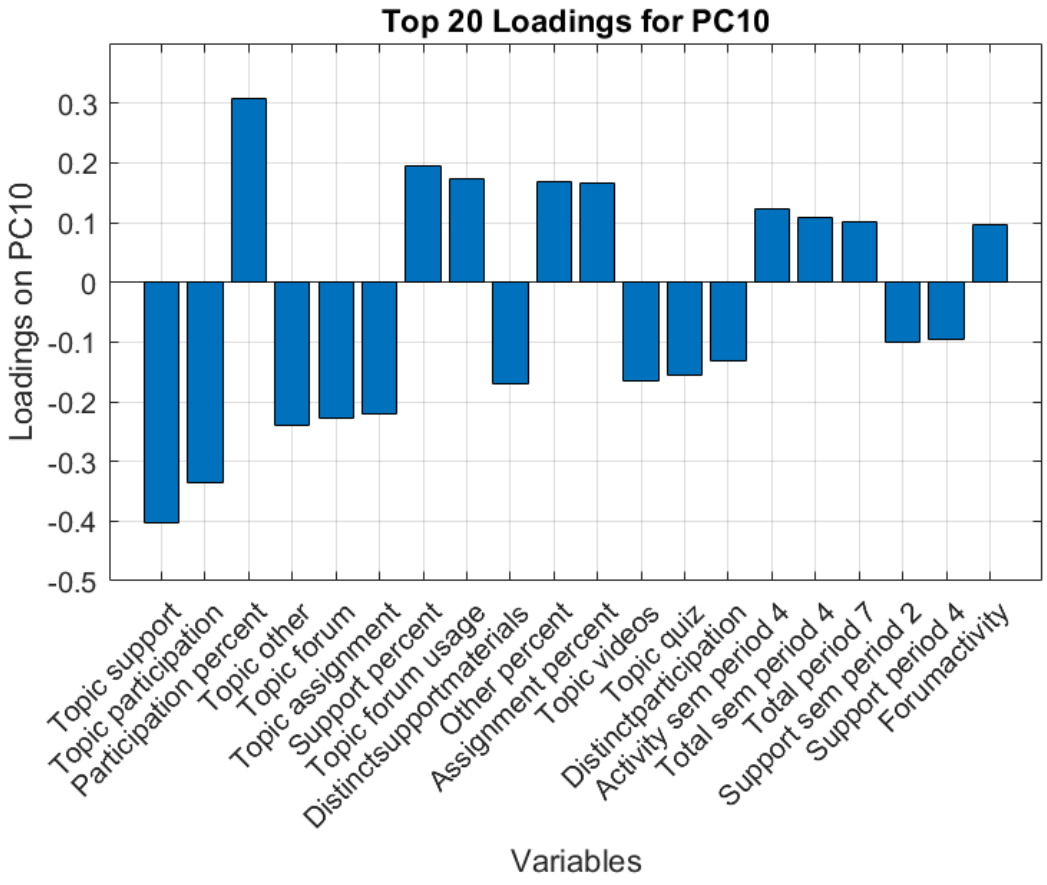

Figure 35 - Top 20 loading attributes for principal component ten

4.3.3. Comparing multiple principal components

Multiple biplots were generated to visualise the combinations of the first four principal components of the dataset, aiming to elucidate the contributions of the dataset's attributes to these components. The analysis of biplots revealed two noteworthy features: the angle between groupings of attributes and the length of their vectors. Specifically, the angle between the vectors of any two attributes indicates their correlation strength and direction: smaller angles suggest a strong positive relationship, 90-degree angles denote no linear relationship, and 180-degree angles signal a strong negative relationship.

Furthermore, the length of the vectors, depicted by blue lines extending from the origin to the attribute point, signifies the proportion of variance each attribute contributes. This visualisation serves as an initial gauge of attribute relatedness. Subsequent analyses will delve into the pairings of principal components, with attributes identified by numerical indexes (based on their dataset position) and colour-coded groupings according to their utilisation in the LMS, such as Enrolment information (Black), and so forth. The association between indexes displayed on biplots and dataset labels is detailed in Table 31.

Table 31 - Biplot attribute index / label mappings

#	Attribute label	#	Attribute label	#	Attribute label	#	Attribute label
1	**Days active**						
3	Topic videos	4	Topic quiz	9	Topic other		
10	Video percent						
19	Distinct Lecture	20	Distinct Quiz	21	Distinct Assign	25	Distinct Other Interaction Materials
17	Avg. sec. between actions	18	Interactions	26	Assignment interactions	27	Video interactions
28	Support material interactions	29	Activity interactions	31	Course participation interactions	32	Other interactions
36	Total period 4	37	Total period 5	38	Total period 6	39	Total period 7
40	Total period 8	47	Other period 1	61	Other period 3	63	Video period 4
65	Activity period 4	68	Other period 4	69	Assignment period 5	70	Video period 5
72	Activity period 5	74	Participation period 5	75	Other period 5	76	Assignment period 6
77	Video period 6	79	Activity period 6	81	Participation period 6	82	Other period 6
84	Video period 7	88	Participation period 7	89	Other period 7	91	Video period 8
96	Other period 8						
97	Total sem period 1	98	Total sem period 2	99	Total sem period 3	100	Total sem period 4
101	Total sem period 5	103	Video sem period 1	107	Participation sem period 1	108	Other sem period 1
109	Assignment sem period 2	110	Video sem period 2	114	Participation sem period 2	115	Other sem period 2
117	Video sem period 3	122	Other sem period 3	123	Assignment sem period 4	124	Video sem period 4
126	Activity sem period 4	128	Participation sem period 4	129	Other sem period 4	131	Video sem period 5
133	Activity sem period 5	136	Other sem period 5				

In Figure 36, a distinct clustering of attributes is observable, each demonstrating strong positive relationships and similar contributions to variance. Despite the diverse mix of attribute types within each clustering, certain pairs, such as Cluster 1 (comprising attributes 3, 10, 19, 27, 63, 70, 77, 84, 91, 103, 110, 117, 124, and 131) and Cluster 2 (comprising attributes 20, 29, 72, and 133), as well as Cluster 4 (comprising attributes 18, 36, 37, 38, 39, 40, 99, and 100) and Cluster 5 (comprising attributes 1, 28, 31, 74, 81, 88, 97, 98, 107, 114, and 128), exhibit closely related attributes as indicated by their minimal angular difference.

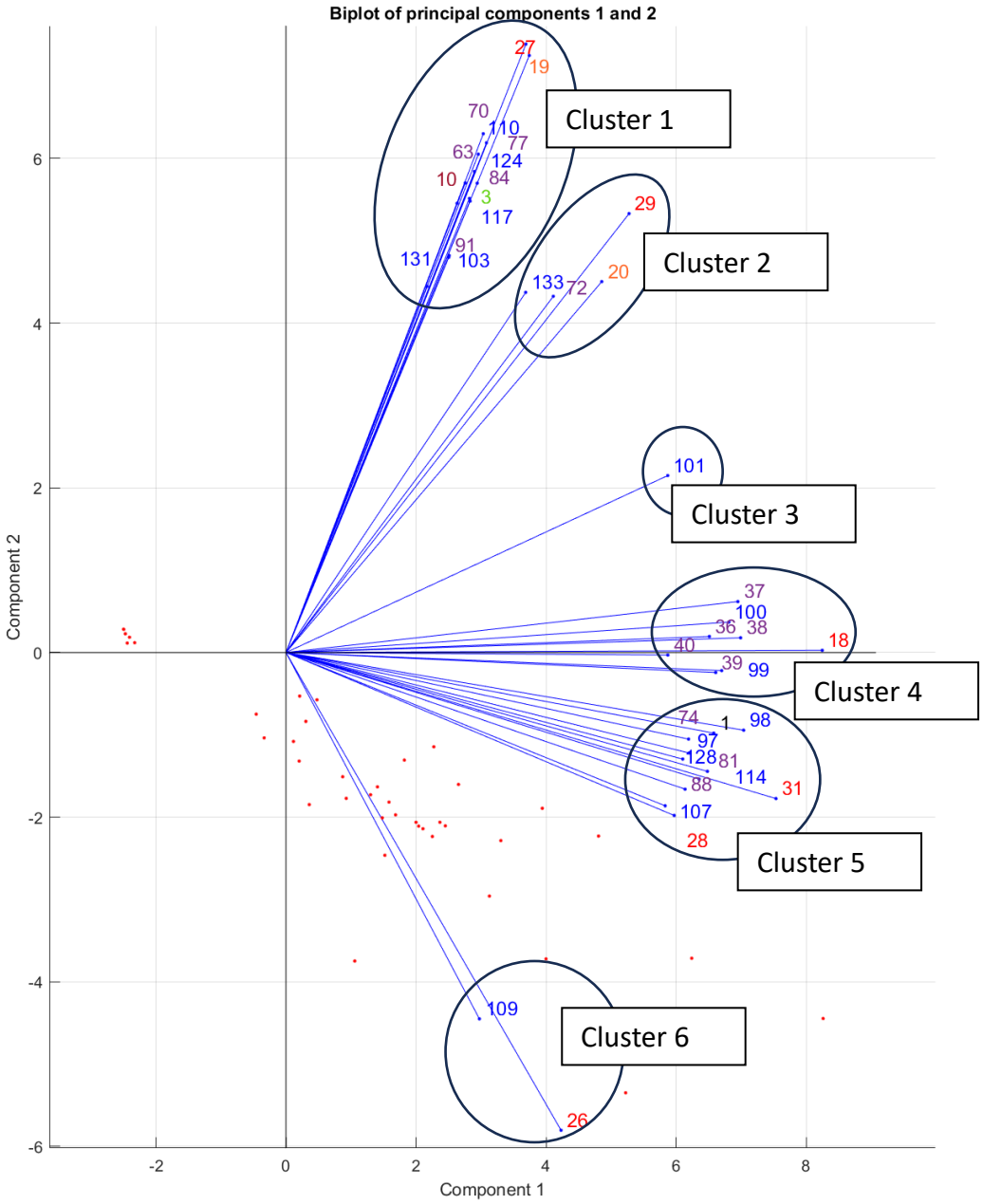

Figure 36 - Biplot of first and second principal components

Cluster 3, with a single attribute (101), and Cluster 6, comprising attributes 26 and 109, serve as smaller yet significant clusters. They indicate mild to no relationships with other clusters, enriching our understanding of the dataset's dimensional structure.

Cluster 1 likely represents aspects of video-based content engagement and activity across different periods, suggesting consistent engagement patterns. Cluster 2's attributes relate to interactions with diverse content types, implying a multifaceted approach to learning where video content consumption complements other forms of engagement.

The single attribute in Cluster 3 represents a broad measure of engagement during a specific semester period, capturing elements common to both video consumption and diverse content interaction.

Cluster 4 encompasses overall participation metrics and engagement across several time periods. Its low angle of difference with Cluster 5 indicates consistent engagement levels across different times and interactions. Cluster 5 relates to the timing of interactions and participation in various activities, highlighting the interconnectedness of temporal engagement patterns with overall participation metrics.

Finally, Cluster 6 focuses on assignment-related interactions, suggesting a specific facet of academic engagement. Its slight negative relationship with Cluster 1 and Cluster 2 might indicate that assignment engagement varies independently from video and quiz interactions, possibly reflecting different learning styles or course demands.

Comparing all clusters, Cluster 1 and Cluster 2 are closely related, suggesting a complementary relationship between video content engagement and interactive activities. Cluster 3 acts as a bridge between detailed participation metrics (Cluster 4) and temporal interaction patterns (Cluster 5), while Cluster 6's unique nature underscores the potential independence of assignment engagement from other forms of content interaction.

For component one and three, shown in Figure 37, five clusters of attributes are identified. Clusters 1 and 2 are positively related but diverge on the amount of variance explained. Cluster 3 and Cluster 4 are closely related, positioned at approximately 90 degrees to Clusters 1 and 2, indicating a different dimension of engagement focused more on the breadth and consistency of participation over time. Cluster 4 acts as a bridge between Cluster 3 and Cluster 5, with Cluster 5 being relatively closely related to Cluster 4.

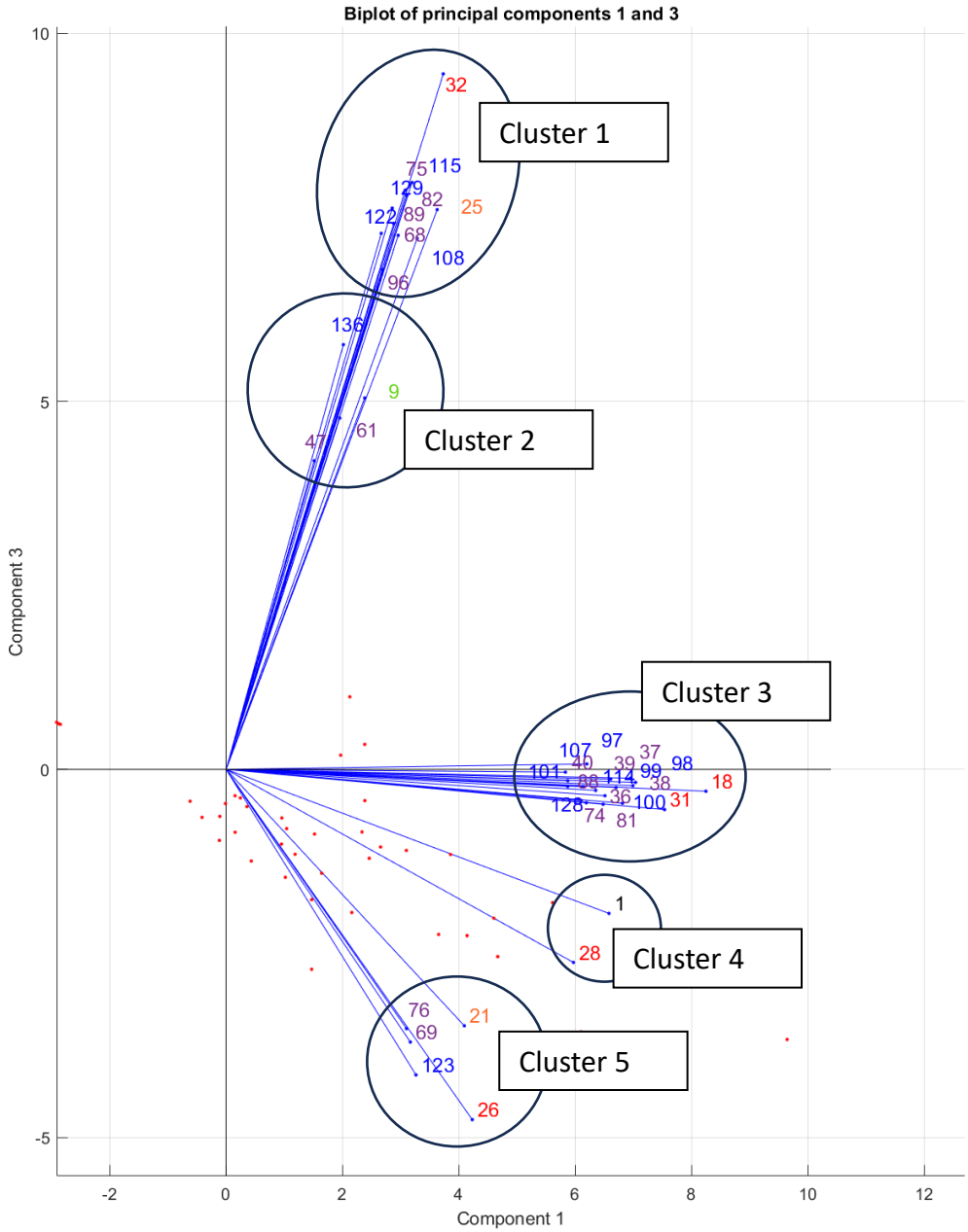

Figure 37 - Biplot of first and third principal components

Cluster 1 captures a wide range of interactions and engagements across different periods, suggesting a multifaceted engagement strategy by students. The variance in how much each attribute contributes points to differences in how each type of interaction influences overall engagement.

Cluster 2 focuses on overall participation metrics and engagement across several time periods. Its relationship with Cluster 4, positioned at approximately 90 degrees, underscores a distinct aspect of engagement measured by participation and temporal patterns.

With only two attributes, Cluster 4 represents core aspects of engagement, such as days active and support material interactions. Its role as a bridge to Cluster 5 suggests these core engagement metrics are fundamental to understanding both broad participation patterns and more specific academic activities, like assignments.

Finally, Cluster 5 focuses on assignment interactions and related activities. The relative closeness of Cluster 5 to Cluster 4 indicates that core engagement metrics are relevant to understanding assignment-related activities. Cluster 4's position as a bridge suggests that assignment engagement incorporates unique aspects of student interaction not fully captured by broader engagement metrics.

Clusters 1 and 2 share a positive relationship, highlighting the complementary nature of engaging with diverse and specific content types over various periods. Clusters 3 and 4, positioned 90 degrees to Clusters 1 and 2, emphasise a different engagement dimension focusing on participation breadth and temporal patterns. Cluster 4's foundational role in connecting broad participation patterns with specific academic activities, like assignments, is crucial for a comprehensive understanding of student engagement.

For component one and four, shown in Figure 38, five clusters of attributes are evident, each with relatively similar levels of variance explained. There is an almost 90-degree angle (slightly larger) between Cluster 1 (attributes 10, 17, 19, 27, 63, 70, 77, 84, 91, 110, 124, 117, and 131) and Cluster 5 (attributes 4, 20, 29, 65, 72, 79, and 126), suggesting no relationship to a mild negative relationship.

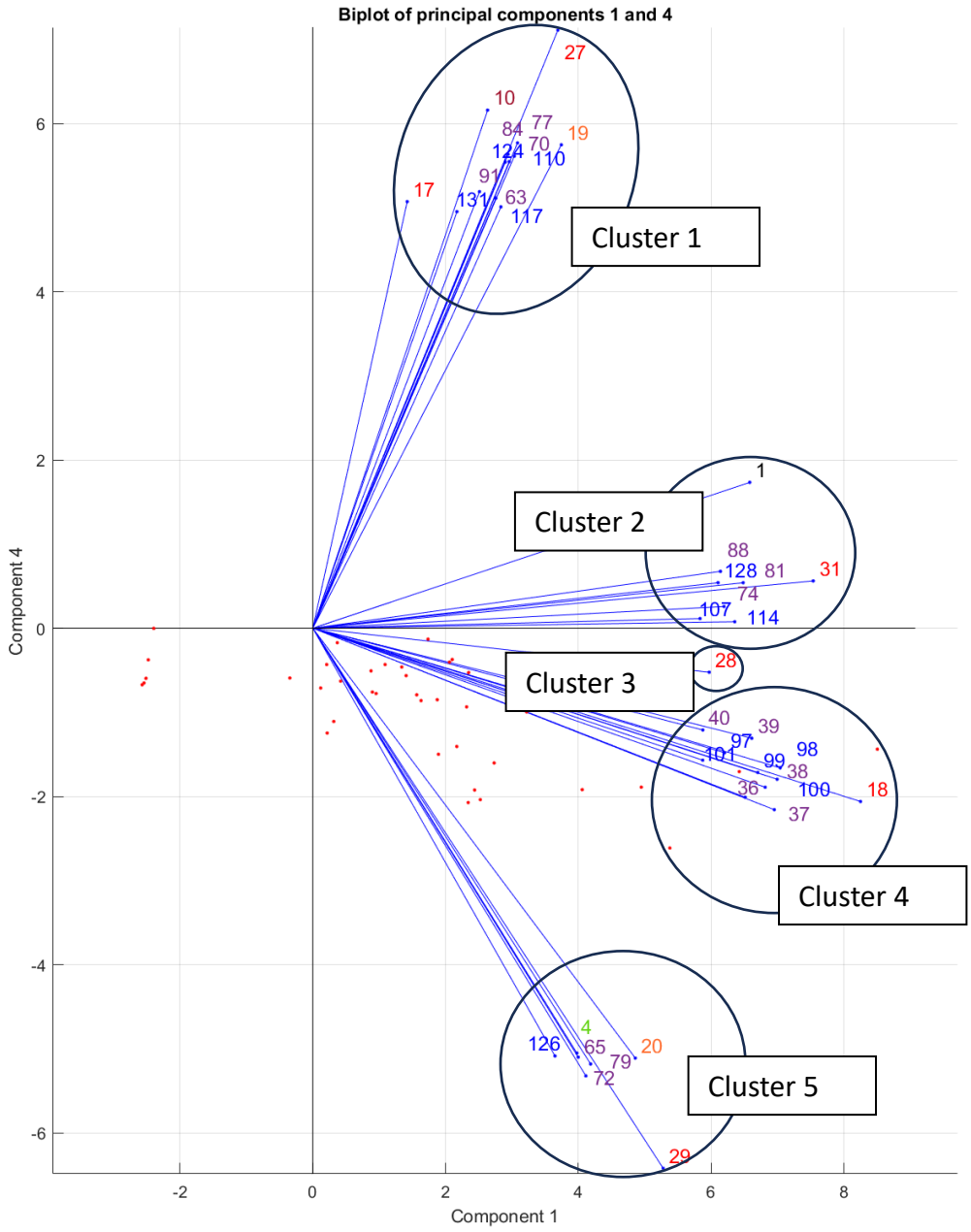

Figure 38 - Biplot of first and fourth principal components

Positioned between these clusters, Cluster 2 (attributes 1, 31, 74, 81, 88, 107, 114, and 128) and Cluster 4 (attributes 18, 36, 37, 38, 39, 40, 97, 98, 99, 100, and 101) show a reasonably strong relationship, with Cluster 2 having a closer angle to Cluster 1, and Cluster 4 closer to Cluster 5. Lastly, Cluster 3 (attribute 28), is located centrally among all clusters.

Cluster 1 primarily involves attributes related to video content engagement, such as the percentage of videos watched, average seconds between actions (potentially indicating engagement intensity), and video interactions across various periods.

The relatively similar levels of variance explained suggest these aspects of video content engagement are equally significant. The angle slightly larger than 90 degrees to Cluster 5 indicates a relationship ranging from non-existent to mildly negative, suggesting that video engagement metrics operate independently or in contrast to the activities represented by Cluster 5.

Cluster 2 represents core aspects of engagement, including days active and various forms of participation interactions. Its closer angle to Cluster 1 suggests a positive relationship with video content engagement, indicating that students who are active on more days and engage in various participation activities are likely also engaged with video content.

Cluster 3, positioned centrally among all clusters, signifies engagement with support materials. Its central positioning suggests it may bridge different forms of engagement, being potentially relevant to both content consumption (videos) and interactive or assignment activities. This highlights the importance of support materials in linking various aspects of the learning experience.

Cluster 4 captures overall participation metrics and engagement across several time periods, indicating a broad measure of student engagement over time. Its proximity to Cluster 5 suggests that this broader engagement closely relates to interactive and assignment activities more than to video content engagement, possibly indicating that consistent participation correlates with engagement in assignments and interactive activities.

Cluster 5 focuses on interactive activities, such as quizzes, assignment interactions, and various types of active engagement. The slight negative relationship with Cluster 1 suggests engagement with interactive and assignment activities may vary inversely with video content engagement, highlighting different engagement patterns or preferences among students.

The slight negative relationship between Clusters 1 and 5 suggests differing student engagement patterns, with some students preferring video content while others lean towards interactive and assignment activities. The strong relationship between Clusters 2 and 4, along with their positioning relative to Clusters 1 and 5, indicates a spectrum of engagement from content consumption to interactive participation. Core engagement metrics and overall participation are key components of this spectrum. Cluster 3's central positioning emphasises its crucial role in understanding how different forms of engagement are interconnected.

For component two and three, shown in Figure 39, there is a wider alignment of clusters, almost forming a full 360-degree pattern. Cluster 1 (attributes 25, 32, 68, 75, 82, 89, 96, 108, 115, 122, and 129) and Cluster 2 (attributes 9, 47, 61, and 136) are aligned, with Cluster 1 showing longer vectors, indicating significantly more variance explained. Positioned almost 90 degrees to Clusters 1 and 2, Cluster 3 (attributes 20, 29, 72, and 133) suggests no relationship to Clusters 1 and 2.

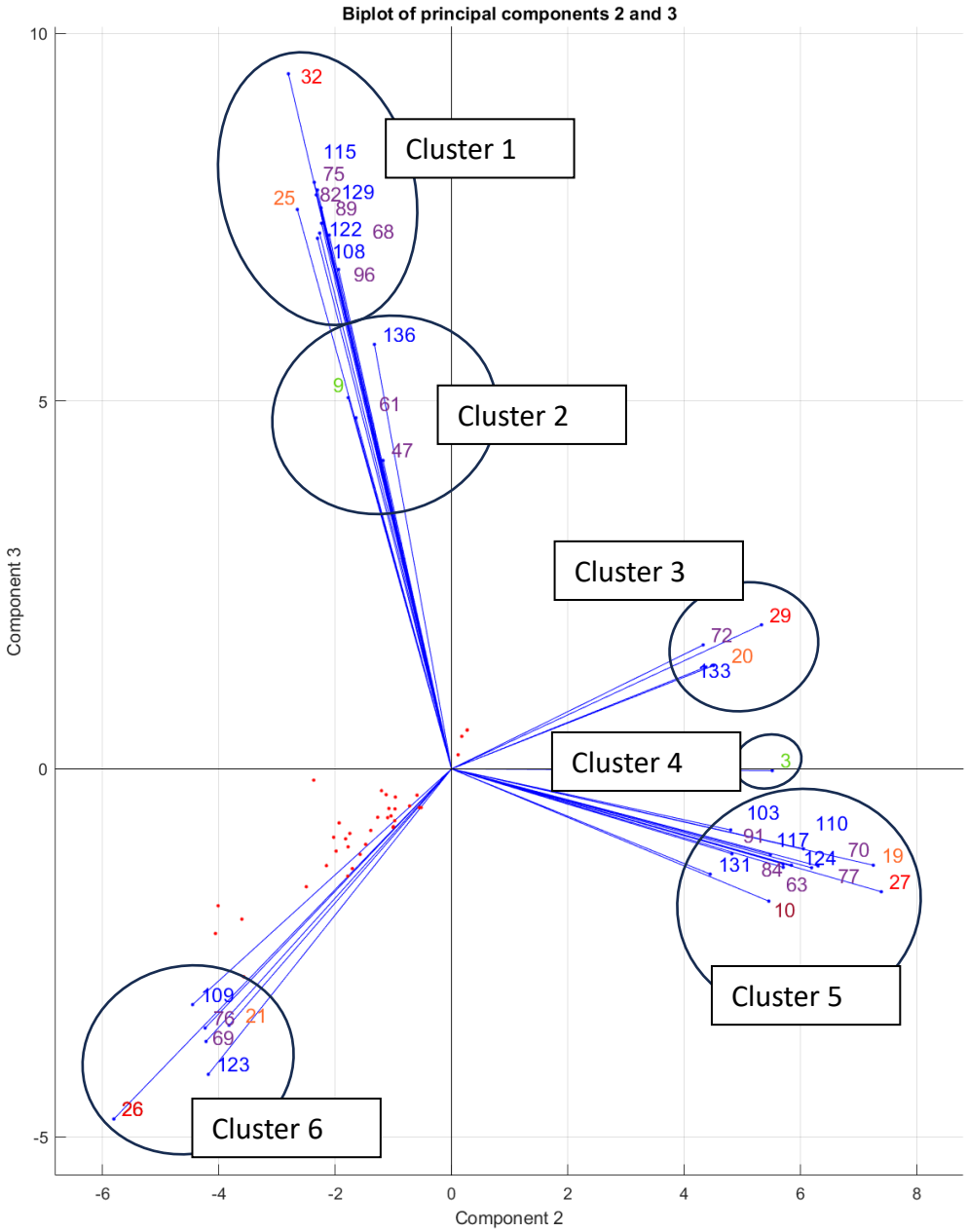

Figure 39 - Biplot of second and third principal components

Moving clockwise, Cluster 4 (attribute 3) is directly between Cluster 3 and Cluster 5 (attributes 10, 19, 27, 62, 70, 77, 84, 91, 103, 110, 117, 124, and 131), suggesting Cluster 4 has a slightly negative relationship with Clusters 1 and 2 but is closely related to Clusters 3 and 5. Continuing clockwise, the final Cluster 6 (attributes 21, 26, 69, 76, 109, and 123) is almost 180 degrees away from Cluster 3, and just over 90 degrees from Cluster 5 and Clusters 1 and 2, suggesting a very slight negative relationship with Clusters 5, 1, and 2, and a strong negative relationship with Cluster 3.

Cluster 1, with its longer vectors, indicates a significant explanation of variance, capturing a wide range of interactions and engagements across different periods. The alignment with Cluster 2 suggests complementary engagement patterns, with Cluster 2 representing more focused areas of student activity.

Cluster 3, positioned almost 90 degrees to Clusters 1 and 2, signifies no relationship with these clusters, emphasising a distinct dimension of engagement related to interactive and assignment activities. This separation highlights the unique role of these activities in the learning experience, distinct from the engagement patterns of Clusters 1 and 2.

Cluster 4, situated directly between Clusters 3 and 5 and represented by a single attribute related to topic videos, plays a pivotal role in bridging interactive/assignment activities with broader content engagement. Its slightly negative relationship with Clusters 1 and 2, alongside closer ties with Clusters 3 and 5, highlights the nuanced role of video content in the broader context of student engagement.

Cluster 5, positioned near Cluster 4 and demonstrating close relationships, underscores the importance of video content engagement across various periods. Its angle just over 90 degrees from Clusters 1 and 2 indicates a shift from diverse and specific content engagement towards more focused engagement with video content.

Cluster 6, almost 180 degrees away from Cluster 3, shows a very slight negative relationship with Clusters 5, 1, and 2, highlighting a strong negative relationship with interactive and assignment activities (Cluster 3). This positioning suggests that assignment engagement represents a distinct, possibly more solitary, or reflective form of engagement, contrasting sharply with the more interactive or diverse forms of engagement represented by other clusters.

The alignment of Clusters 1 and 2, with differences in variance explained, underscores the varying importance of diverse interactions and specific content engagement. The sequence from interactive activities to video content engagement for Clusters 3, 4, and 5 illustrates a continuum of engagement types, with topic videos serving as a crucial link. Cluster 6's positioning indicates a strong negative relationship with Cluster 3 and slight negative relationships with other clusters, suggesting that assignment engagement occupies a unique niche within the spectrum of student engagement.

For component two and four, shown in Figure 40, the pattern differs from the previous pairing, showcasing a full 180-degree relationship between all clusters rather than a 360-degree arrangement. This is evidenced by the strong negative relationship between Cluster 1 (attributes 26 and 109) and Cluster 5 (attributes 4, 20, 29, 65, 72, 79, 126, and 133), with an almost 180-degree angle between them.

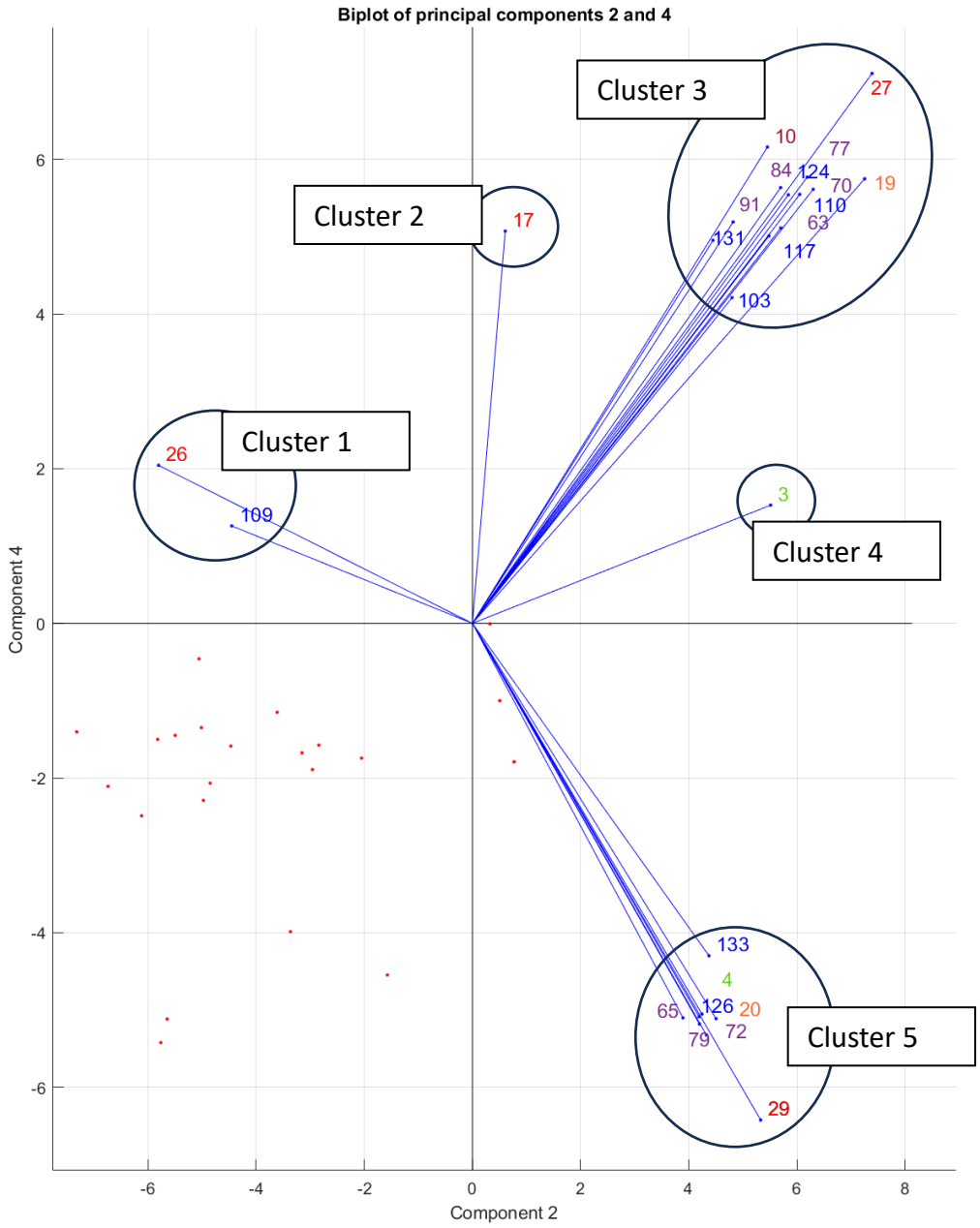

Figure 40 - Biplot of second and fourth principal components

Cluster 1 and Cluster 3 (attributes 10, 19, 27, 63, 70, 77, 84, 91, 103, 110, 117, 124, 131) share an almost 90-degree angle, indicating no relationship. Between these two, Cluster 2 (attribute 17), closer to Cluster 3, shows a more significant relation to both than the relationship between Clusters 1 and 3 themselves. Moving clockwise, Cluster 4 (attribute 3) finds itself in a similar position, closer to Cluster 3 but situated between Cluster 3 and Cluster 5. This suggests that Cluster 4 is reasonably related to Cluster 3 but maintains a 90-degree angle from Cluster 5, indicating no relation. Conversely, Cluster 5 exhibits a slightly more than 90-degree angle from Cluster 3, suggesting a slight negative relationship.

Cluster 1, represented by only two attributes, focuses on specific aspects of assignment engagement. The almost 180-degree angle in its relationship with Cluster 5 suggests that the forms of engagement represented by these clusters are diametrically opposite, indicating that engagement behaviours captured by Cluster 1 markedly differ from those involved in interactive and group activities.

Cluster 2, positioned closer to Cluster 3 but situated between Clusters 1 and 3, likely reflects a measure of engagement intensity or pacing. Its proximity to Cluster 3 suggests a stronger relation to video content engagement, yet it retains relevance to both Clusters 1 and 3, acting as a potential bridge in understanding how engagement intensity correlates with content consumption patterns.

Cluster 3 signifies extensive engagement with video content across various periods. The lack of a direct relationship with Cluster 1, as indicated by the almost 90-degree angle, shows that video content engagement functions independently from the specific assignment engagement identified in Cluster 1.

Cluster 4, focused on topic videos, aligns more closely with Cluster 3, emphasising the theme of video content engagement. Its position between Clusters 3 and 5, and the 90-degree angle from Cluster 5, signifies no direct relationship with Cluster 5's interactive activities, highlighting that topic video engagement, while part of the broader video content engagement category, is distinct from the interactive and collaborative activities in Cluster 5.

Cluster 5 encompasses attributes related to quizzes, assignment interactions, and various active engagements, embodying interactive and group-based learning activities. Its significant separation from Cluster 1 underscores the fundamental differences in engagement, highlighting the contrast between solitary assignment work and collaborative or interactive experiences.

The strong negative relationship between Clusters 1 and 5 emphasises the contrasting engagement behaviours between solitary assignment tasks and collaborative or interactive activities. The absence of a direct relationship between Cluster 3 and Cluster 1 stresses the distinct nature of video content engagement from both assignment and interactive activities. Cluster 2's positioning suggests it may provide insights into engagement pacing across different content consumption types, while Cluster 4's relationship with Cluster 3 underlines the nuanced role of topic-specific video engagement within the broader context of video content interaction.

Finally, for component three and four, as shown in Figure 41, a 360-degree relationship is again evident across all attribute clusters. Starting with Cluster 1 (attributes 21, 26, 69, 76, and 123) and Cluster 2 (attributes 10, 17, 19, 27, 63, 70, 77, 84, 91, 110, 117, 124, and 131), they exhibit a relatively strong relationship. Clusters 3 (attributes 9, 47, 61, and 136) and 4 (attributes 25, 32, 68, 75, 82, 89, 96, 108, 115, 122, 129) form an even stronger bond, with Cluster 3 displaying slightly shorter vectors than Cluster 4, which suggests a significant variance explanation.

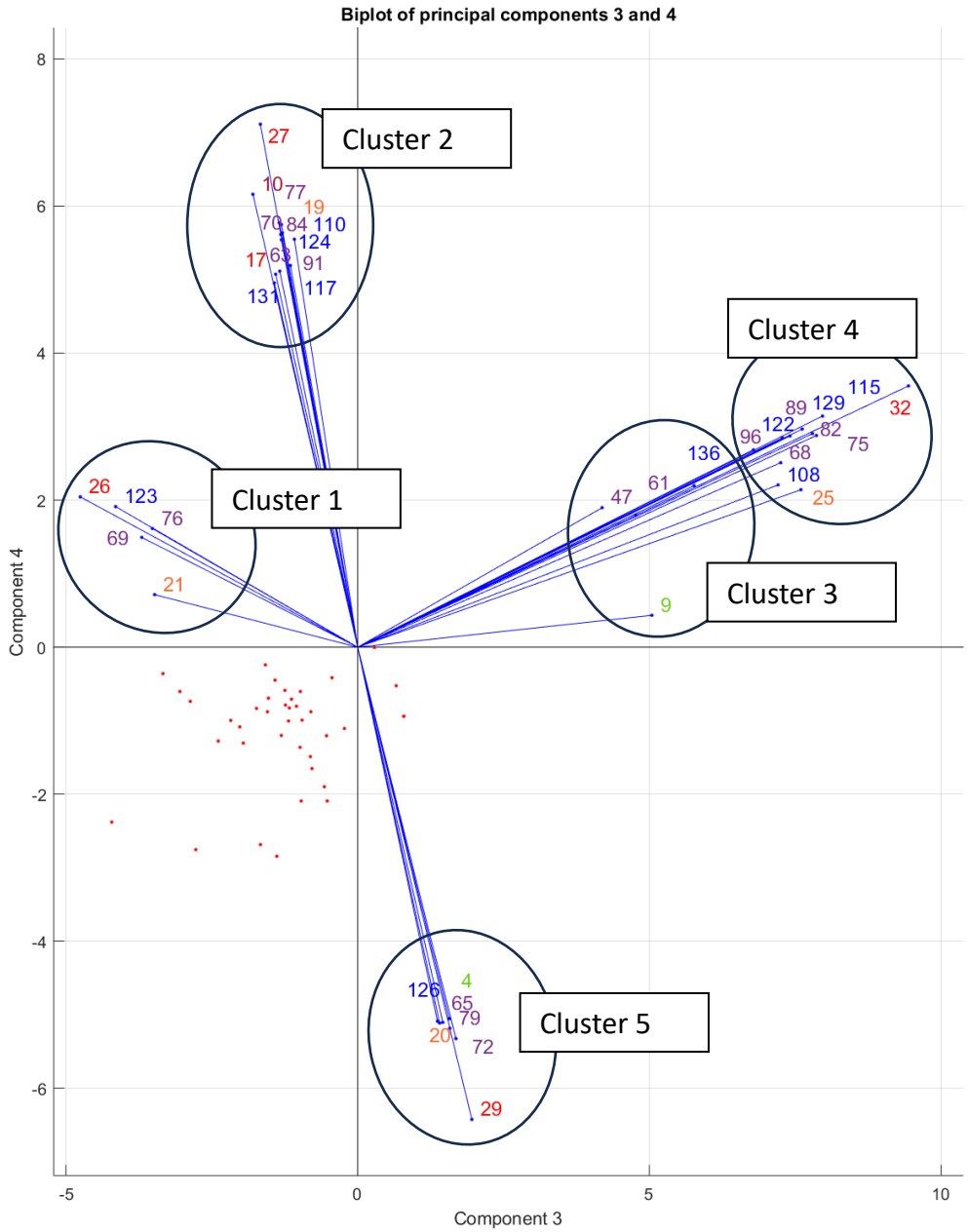

Figure 41 - Biplot of third and fourth principal components

Lastly, Cluster 5 (attributes 4, 20, 29, 65, 72, 79, and 126) is positioned perfectly 180 degrees from Cluster 2, and slightly over 90 degrees from Cluster 4, establishing it as strongly negatively related to Cluster 2, not related to, or very mildly negatively related to Clusters 3 and 4, and moderately negatively related to Cluster 1.

Cluster 1, focusing on assignment-related interactions, suggests a dimension of engagement centred around completing and interacting with assignments. Its positioning relative to other clusters highlights distinct patterns of engagement that contrast with other activity types, especially the content engagement represented by Cluster 2.

Cluster 2, showing extensive engagement with video content, has a strong relationship with Cluster 1, indicating complementary patterns between assignment engagement and video content engagement. This implies that students involved in assignments are also likely to engage actively with video content.

Cluster 3, highlighting specific content types and engagement during particular periods, offers a more focused engagement dimension potentially indicative of targeted learning or review sessions. The slightly shorter vector length compared to Cluster 4, yet maintaining a strong relationship, indicates that these specific engagements contribute to a broader spectrum of engagement types.

Cluster 4, with its wide range of interactions and engagements across different periods, has longer vectors that suggest a significant contribution to the dataset's variance. The strong relationship with Cluster 3 underscores the diverse and specific engagement types' interconnectedness, highlighting the complexity of student interaction patterns.

Cluster 5, related to interactive and group-based learning activities, is positioned perfectly 180 degrees from Cluster 2, suggesting a strong negative relationship with video content engagement. This positioning indicates that the engagement behaviours of Cluster 5 are almost entirely opposite to those of Cluster 2, accentuating the dichotomy between interactive activities and content consumption. Its relationship with Clusters 3 and 4 suggests varying degrees of connectivity, indicating different engagement dimensions.

Clusters 1 and 2's relatively strong relationship underscores complementary patterns between assignment and video content engagement. The strong connection between Clusters 3 and 4 emphasises a spectrum of engagement from specific content interactions to broader, diverse interactions. Cluster 5's positioning as strongly negatively related to Cluster 2 and moderately negatively related to Cluster 1 indicates a distinct form of engagement focused on interactive and collaborative activities, significantly differing from content consumption and assignment engagement patterns.

4.3.4. PCA results summary

Variance explained

The initial analysis, depicted in scree and cumulative variance plots, identified a significant drop in variance explanation after the initial few components, with the first ten components showing a progressive increase in cumulative variance explained. This indicates diminishing returns on additional components beyond this range for explanation purposes. To achieve higher levels of cumulative variance explanation (90%, 95%, 99%), a substantially greater number of components are required (65, 82, and 101 components, respectively).

Visualisation of principal components

The top loading attributes for the first ten principal components are analysed to discern patterns of LMS engagement. These components reveal diverse aspects of student interaction with the LMS, ranging from overall engagement, contrasts between video and assignment interactions, to the use of support materials and social interactions.

Principal Component One: Highlights general engagement across time, with significant variance in semester and daily activities.

Principal Component Two: Contrasts video engagement with assignment interactions, suggesting a preference for video-based learning.

Principal Component Three: Focuses on 'Other' interactions, inversely related to these activities.

Principal Component Four: Balances video and active participation, indicating a mix of passive and active learning behaviours.

Principal Component Five: Compares support material usage against assignment engagement, reflecting different learning strategies.

Principal Component Six: Centres on social interactions within the LMS.

Principal Component Seven to Ten: Further dissect student engagement patterns, revealing shifts in focus from early engagement to support material use, temporal shifts in engagement focus, and balances between participation and support, as well as general engagement the composition of topics.

Comparison of Multiple Principal Components:

Biplots generated to compare the first four principal components highlight the relationships between different attribute clusters, with emphasis on the angles and vector lengths indicating correlation strengths and variance contributions. These visual analyses reveal patterns of relatedness and contrast among various forms of LMS engagement, such as video content engagement, interactive activities, and assignment interactions.

The PCA analysis within this study provides a comprehensive overview of student engagement patterns with the LMS. By examining the variance explained and the visualisation of principal components, it's evident that LMS engagement is multifaceted, encompassing a wide range of activities from content consumption to interactive participation and social interactions.

The initial components reveal significant aspects of engagement, such as a preference for video content and a balance between active and passive learning behaviours. Subsequent components illustrate nuanced patterns, including shifts in engagement focus and the importance of support materials.

Through the comparison of multiple principal components, the analysis underscores the complexity of learning behaviours and preferences, highlighting the need for diverse strategies to support student engagement within the LMS. This nuanced understanding of engagement patterns is crucial for tailoring educational content and interventions to meet varied student needs and preferences effectively.

Reduction of number of attributes

One of the pivotal objectives of the experiments outlined in this section is to scrutinise the necessity and efficacy of dimensionality reduction, aiming to preserve as much valuable information as possible from the dataset, which will be pivotal in subsequent analyses aimed at enhancing model performance. This research is particularly crucial given the expansive dimensionality of the dataset, as highlighted in the summary of PCA results. The research is primarily focused on uncovering intricate patterns within the data, where every detail could potentially contribute to a deeper understanding of the underlying dynamics.

The challenge of attribute reduction lies not only in determining the optimal number of components required to explain a significantly large proportion of the dataset's variance but also in managing the overall attribute participation across a relatively limited number of principal components. The cumulative variance plot, as demonstrated in Figure 25, explains this challenge by revealing that the initial ten components account for merely 54% of the variance. Yet, intriguingly, the top-loading attributes participating in these components span a total of 125 out of the 130 attributes in the dataset. This finding underscores the dataset's complex structure, where a vast majority of attributes play a role in the variance explained by the principal components, indicating a high level of interconnectedness and complexity within the data.

Given the primary aim to explore patterns without losing any detail and the availability of sufficient computing hardware, there is a compelling argument to proceed with the analysis despite the challenges associated with dimensionality reduction. The advanced computing resources at our disposal allow for the handling of the dataset's comprehensive dimensionality, thereby enabling a thorough investigation that does not compromise on the granularity of the data. This approach aligns with the research's goal to delve into the data's intricacies, ensuring that no critical patterns or details are overlooked in the pursuit of understanding the phenomena under study. Thus, while dimensionality reduction poses its set of challenges, the intent to uncover and analyse patterns in their entirety, bolstered by adequate computational support, guides the decision to embrace the dataset's complexity and proceed with the analysis.

In addressing the research question (RQ1.4), "Is dimensionality reduction necessary to accurately capture the essential aspects of LMS use, and what impact does this reduction have on the performance of predictive models?", the analysis suggests that dimensionality reduction may not be imperative for our purposes. The complexity of the dataset, highlighted by the PCA results, indicates that each principal component, along with its unique combination of attributes, captures distinct facets of LMS interactions. These facets include varied aspects of LMS usage and intersections of interaction types, each contributing uniquely to the comprehensive understanding of user engagement within the system.

The findings from the PCA analysis reveal a nuanced landscape of LMS interactions, where even seemingly minor attributes play a significant role in depicting the overall engagement patterns. Given the expansive dimensionality of the dataset, reducing the number of attributes poses a substantial risk of omitting critical variance and, consequently, essential insights into LMS use. With the first ten components accounting for just over half of the variance and the majority of attributes participating in these components, it becomes evident that a significant amount of information might be lost in the process of dimensionality reduction.

Moreover, the availability of sufficient computing hardware mitigates the potential computational challenges posed by the dataset's expansive dimensionality, enabling a thorough investigation without the need for compromise on data granularity. Considering this, the safest course of action is to retain all attributes for subsequent analyses. This approach not only ensures that no vital aspects of LMS interactions are overlooked but also maximises the potential for predictive models to leverage the full spectrum of available data, thereby enhancing their performance and accuracy.

Therefore, in response to the research question, the comprehensive exploration of LMS usage patterns and their impact on model performance is best served by proceeding with the full set of attributes. This strategy aligns with the overarching goal of the research to uncover detailed patterns of LMS use without sacrificing the richness of the dataset, providing a strong foundation for subsequent experiments and analyses.

4.4. Experiment 1 results

The performance of each selected algorithm on the dataset, including the college attribute, was compared. The results as shown in Table 32, reveals that while there are minor differences across algorithms in performance metrics, no single algorithm outshines the others significantly. However, Decision Stump, Random Tree, and Naïve Bayes show lower results in Kappa and Matthew's Correlation metrics. REPTree, J48, and Simple CART, among the tree-based algorithms, perform similarly, with ensemble algorithms offering marginally better outcomes at the cost of interpretability. The comparison highlights the balance between algorithm complexity, interpretability, and performance, with simpler models like REPTree and Simple CART providing smaller, more interpretable models compared to ensemble methods like Random Forest and Rotation Forest, which yield larger models but potentially higher accuracy.

4.4.1. Predicting grade (including college attribute)

Table 32 - Algorithm comparison (predict grade including college attribute)

	Accuracy	Weighted Avg. KAPPA	Weighted Avg. MCC	Weighted Avg. AU ROC
REPTree	0.38	0.20	0.20	0.68
DStump	0.30	0.08	-	0.56
RandomTree	0.31	0.12	0.12	0.56
J48	0.37	0.20	0.19	0.66
J48 (RT Prune)	0.38	0.20	0.19	0.67
NaiveBayes	0.17	0.06	0.06	0.58
SimpleCART	0.38	0.20	0.20	0.68
RandomForest	0.41	0.23	0.23	0.71
RotationForest	0.40	0.21	0.21	0.70
NBTree	0.29	0.15	0.14	0.63
AdaBoost (RT)	0.38	0.20	0.20	0.68

In terms of tree size and model complexity, simpler algorithms like REPTree and Simple CART yielded smaller model sizes, while ensemble algorithms like Random Forest and Rotation Forest significantly increased the model size. This is evident from Table 33, which illustrates the disparity among algorithms. Tree-based algorithms like REPTree and J48 offered reasonably sized models with similar performance in grade prediction.

Table 33 - Tree size comparison (predict grade including college attribute)

	Tree Size	Model Size
REPTree	787.72	193,526.88
DStump	N/A	9,271.00
RandomTree	N/A	7,416,602.84
J48	5,637.56	1,487,410.72
J48 (RT Prune)	1,993.08	536,511.12
NaiveBayes	N/A	61,332.00
SimpleCART	850.12	171,390,643.60
RandomForest	N/A	490,144,973.28
RotationForest	N/A	4,541,505.88
NBTree	1.00	111,039,167.20
AdaBoost (RT)	N/A	197,522.32

4.4.2. Predicting grade (excluding college attribute)

Excluding the college attribute from the model for predicting grades (Table 34), shows modest improvements in most performance metrics, suggesting that this attribute may not significantly contribute to the overall performance of the model. This adjustment in the modelling approach leads to mixed results in Accuracy (improvements in RandomForrest, but others remaining stable), and a slight increase in Kappa, MCC, and AU ROC. The changes in model size for algorithms like Random Forest and Simple CART, alongside the stability observed in REPTree and J48 models, indicate how the exclusion of the college attribute affects model complexity and performance.

Table 34 - Algorithm comparison (predict grade excluding college attribute)

	Accuracy	**Weighted Avg. KAPPA**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.40	0.22	0.22	0.70
DStump	0.30	0.08	-	0.56
RandomTree	0.32	0.14	0.13	0.57
J48	0.38	0.22	0.21	0.68
J48 (RT Prune)	0.39	0.22	0.21	0.68
NaiveBayes	0.18	0.06	0.06	0.59
SimpleCART	0.40	0.23	0.22	0.69
RandomForest	0.42	0.25	0.25	0.72
RotationForest	0.41	0.23	0.23	0.71
NBTree	0.30	0.16	0.14	0.63
AdaBoost (RT)	0.32	0.22	0.22	0.70

Comparing Table 33 and Table 35 reveals significant differences in model size due to the exclusion of the college attribute. Notable changes were observed in Random Forest (model size increased) and Simple CART (model size decreased). Both REPTree and J48 showed relative stability in model size.

Table 35 - Tree size comparison (predict grade excluding college attribute)

	Tree Size	Model Size
REPTree	793.00	195,009.72
DStump	N/A	9,482.00
RandomTree	N/A	7,409,991.20
J48	5,561.24	1,468,230.04
J48 (RT Prune)	2,060.12	554,814.92
NaiveBayes	N/A	62,099.00
SimpleCART	962.04	170,520,085.48
RandomForest	N/A	492,437,264.08
RotationForest	N/A	4,749,426.72
NBTree	1.00	111,869,719.40
AdaBoost (RT)	N/A	193,269.96

4.4.3. Summary of performance and size metrics

The summary of performance and size metrics from Experiment 1 reveals nuanced insights into algorithm effectiveness for grade prediction with and without the college attribute. Including the college attribute shows no definitive algorithmic superiority but highlights a trade-off between interpretability and performance, particularly with tree-based versus ensemble methods. Excluding the college attribute modestly enhances some performance metrics (such as KAPPA and MCC) while not affecting others (such as Accuracy), suggesting it potentially has limited predictive value. Model size analysis indicates a significant impact on complexity, with simpler algorithms providing more interpretable models, while ensemble methods offer higher accuracy at increased model sizes.

4.4.4. Decision tree (REPTree)

For the first decision tree, predicting student grades shown in Figure 42, the tree depth was reduced to the top 5 levels to ensure viewability and interpretability. The path through the tree appears to be influenced by a variety of attributes, including Days in Topic, Forum Participation, Days Active, and specific performance metrics like Topic Quiz Scores and Assignment Interactions. The overall accuracy of paths through the decision tree are not particularly high, except for Fail paths, that have significantly higher accuracy. However, compared to chance for each grade (20%), each is significantly better, and offer insight into the key features required to achieve certain grades.

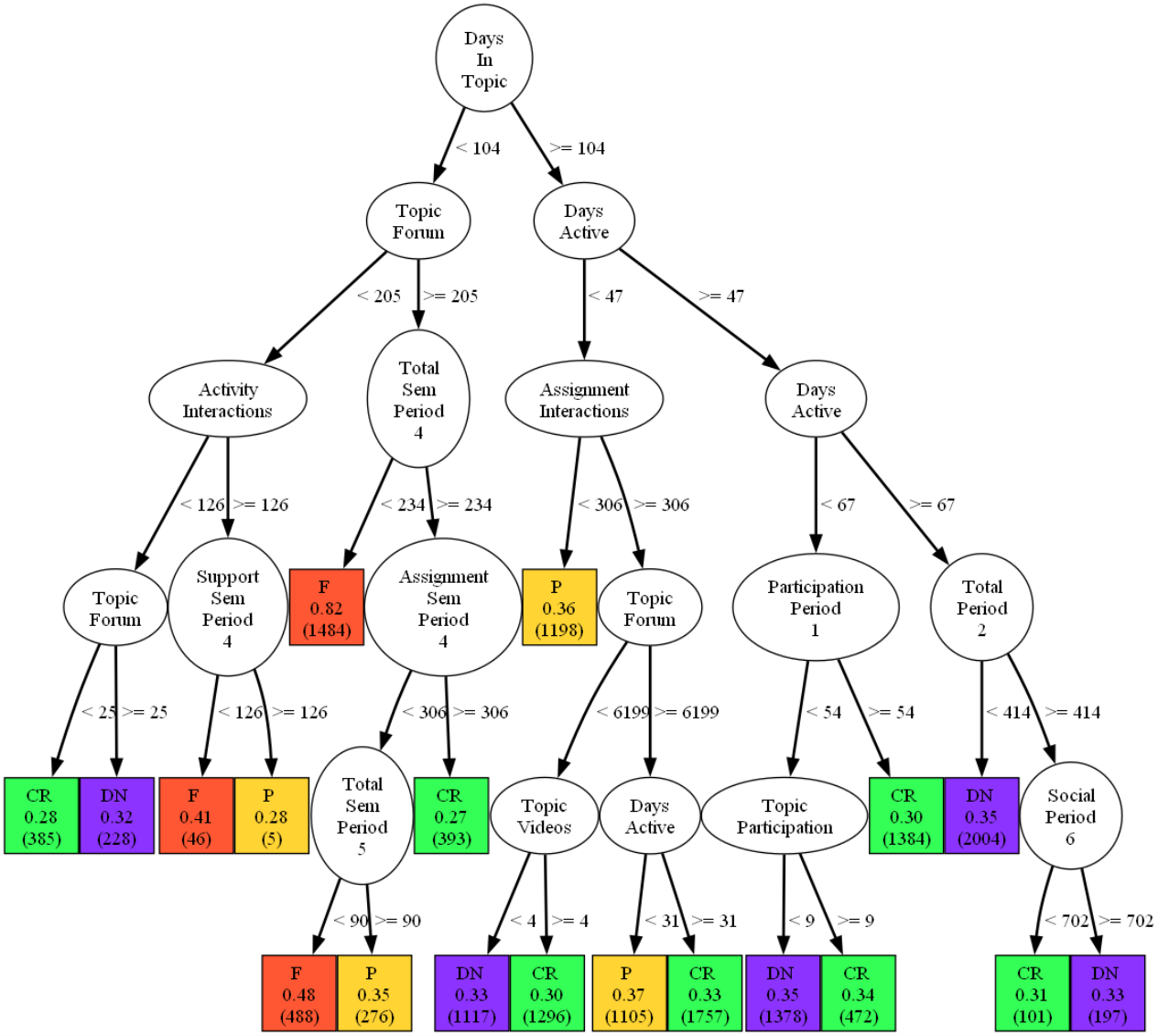

Figure 42 - Pruned grade decision tree (reduced to 5 levels)

High Distinction (HD) and Distinction (DN) grades are associated with high engagement and achievement levels, such as extensive LMS activity, targeted forum participation, and successful quiz and assignment completion. Credit (CR) and Pass (P) grades also reflect engagement but with variations in intensity and focus, while a Fail (F) grade is linked to low engagement across these metrics. The attributes underscore the complexity of academic success, emphasising the balance between engagement, resource utilisation, and academic performance within the LMS environment.

The focus on engagement metrics, alongside academic performance indicators, underscores the multifaceted nature of predicting student grades, emphasising that both participation and academic diligence are crucial for success. The emphasis on active engagement and academic diligence is nuanced by understanding that specific interaction patterns, such as extensive forum participation and consistent LMS activity, are key indicators for higher achievement levels like High Distinction and Distinction.

4.4.5. Experiment 1 results summary

This section consolidates findings from Experiment 1, addressing RQ1.1 and RQ1.2 by evaluating predictive factors of academic performance within the LMS. It highlights the nuanced effectiveness of various algorithms in grade prediction, underscoring the critical role of engagement metrics and academic indicators. The analysis reveals that excluding the college attribute slightly improves model performance, suggesting specific behaviours and engagement patterns are more indicative of academic success than college affiliation, such as Days in Topic, Forum Participation, and Assignment Interactions, can influence grade outcomes. These insights contribute to a deeper understanding of student learning behaviours and their impact on academic outcomes, laying a foundation for targeted interventions and educational strategies.

4.5. Experiment 2 results

4.5.1. Predicting college (including grade attribute)

The results from predicting college affiliation with the inclusion of the grade attribute (Table 36), indicates that algorithms such as SimpleCART, NBTree, and AdaBoost exhibit superior overall performance. These algorithms achieve high scores across all evaluated metrics, underscoring their effectiveness in handling complex predictive tasks. Conversely, Decision Stump and NaiveBayes show limitations in their predictive capabilities, evidenced by their lower performance metrics.

Table 36 - Algorithm comparison (predict college including grade attribute)

	Accuracy	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.80	0.76	0.76	0.96
DStump	0.31	0.15	-	0.60
RandomTree	0.65	0.57	0.57	0.79
J48	0.93	0.92	0.92	0.99
J48 (RT Prune)	0.90	0.88	0.88	0.99
NaiveBayes	0.32	0.17	0.19	0.70
SimpleCART	0.96	0.95	0.95	0.99
RandomForest	0.90	0.87	0.88	0.99
RotationForest	0.87	0.84	0.84	0.98
NBTree	0.96	0.95	0.95	0.99
AdaBoost (RT)	0.99	0.99	0.99	1.00

In terms of model complexity and training efficiency, as shown in Table 37 SimpleCART and NBTree demonstrate a significant increase in model size, suggesting a more complex model structure that, while yielding high accuracy, may demand greater computational resources. On the other hand, algorithms like REPTree and J48 manage to strike a balance, offering reasonable predictive performance without excessively large model sizes or prolonged training times. This balance is crucial for applications where model interpretability and operational efficiency are important considerations alongside predictive accuracy.

Table 37 - Tree size comparison (predict college including grade attribute)

	Train Time	Tree Size	Model Size
REPTree	14.21	1,379.32	327,375.56
DStump	4.12	N/A	9,486.00
RandomTree	1.84	N/A	4,410,176.20
J48	40.65	1,972.68	527,485.40
J48 (RT Prune)	24.14	1,653.48	447,154.28
NaiveBayes	1.86	N/A	62,103.00
SimpleCART	623.67	4,138.92	146,605,712.60
RandomForest	30.58	N/A	304,457,802.04
RotationForest	106.98	N/A	6,216,911.88
NBTree	11,170.80	3,120.68	314,729,093.32
AdaBoost (RT)	290.68	N/A	6,778,813.04

4.5.2. Predicting college (excluding grade attribute)

The exclusion of the grade attribute in predicting college affiliation (Table 38) results in minor improvements across some algorithms while others remain the mostly unchanged, as evidenced by the slight increases in Accuracy, Kappa, MCC, and AU ROC scores. This enhancement suggests that the grade attribute may not be as critical for distinguishing between colleges as initially thought, potentially due to the overarching patterns of engagement and academic behaviours that transcend individual grades.

Table 38 - Algorithm comparison (predict college excluding grade attribute)

	Accuracy	Weighted Avg. Kappa	Weighted Avg. MCC	Weighted Avg. AU ROC
REPTree	0.88	0.85	0.85	0.98
DStump	0.31	0.15	-	0.60
RandomTree	0.67	0.59	0.60	0.80
J48	0.93	0.92	0.92	0.99
J48 (RT Prune)	0.90	0.88	0.88	0.99
NaiveBayes	0.31	0.17	0.19	0.69
SimpleCART	0.96	0.95	0.95	0.99
RandomForest	0.91	0.89	0.89	0.99
RotationForest	0.87	0.84	0.84	0.98
NBTree	0.96	0.95	0.95	0.99
AdaBoost (RT)	1.00	1.00	1.00	1.00

The algorithms' performance trends remained relatively consistent with their outcomes when including the grade attribute (Table 36), affirming their inherent strengths and weaknesses in classification tasks. High Performers such as SimpleCART, NBTree, and AdaBoost maintained exceptional performance levels, indicating their robustness irrespective of the grade attribute's presence. These algorithms demonstrated superior predictive accuracy, with AdaBoost achieving near-perfect metrics, underscoring its effectiveness in complex classification scenarios.

Moderate Performers, including REPTree, RandomTree, J48, and their variants, along with RandomForest and RotationForest, showcased a broad range of effectiveness. Their performance underscores the importance of algorithm selection based on the specific characteristics and requirements of the classification task at hand.

Interestingly, the model size dynamics shifted with the exclusion of the grade attribute (Table 39). While SimpleCART and RandomForest experienced adjustments in model size, suggesting a complex relationship between feature selection and model complexity, algorithms like REPTree and J48 continued to offer a balanced approach. They represent an efficient compromise between training time, model size, and performance, making them suitable for scenarios where computational resources or interpretability are key considerations.

Table 39 - Tree size comparison (predict college excluding grade attribute)

	Train Time	Tree Size	Model Size
REPTree	14.48	1298.44	308,103.76
DStump	3.96	N/A	9,273.00
RandomTree	1.95	N/A	4,145,907.40
J48	39.09	1969.80	526,459.60
J48 (RT Prune)	23.04	1628.52	440,269.28
NaiveBayes	1.86	N/A	61,334.00
SimpleCART	182.56	4437.56	146,005,456.44
RandomForest	14.05	N/A	288,793,795.80
RotationForest	51.86	N/A	6,183,772.44
NBTree	11,880.71	3123.00	312,265,298.56
AdaBoost (RT)	303.33	N/A	6,149,421.08

4.5.3. Summary of performance and size metrics

This section synthesises the outcomes from Experiment 2, focusing on algorithmic efficacy in discerning college affiliations with the inclusion and exclusion of grade attributes. High-performing algorithms like SimpleCART, NBTree, and AdaBoost (RT) demonstrated notable predictive accuracy, albeit with increased model complexity, suggesting a trade-off between accuracy and interpretability. Middle-tier algorithms like REPTree and J48 maintained a balance, offering substantial predictive capabilities without the computational demand of more complex models. Lower-performing algorithms, particularly Decision Stump and NaiveBayes, highlighted the challenge in using simplistic models for nuanced tasks such as predicting college affiliations.

The experiment highlights the relationship between model size, complexity, and performance. While high accuracy is achievable, it often requires more extensive models, as seen with SimpleCART and NBTree. Conversely, REPTree and J48 illustrate that efficiency doesn't necessarily come at the cost of performance, maintaining moderate model sizes and training times. These insights point to the importance of selecting the right algorithm based on specific needs for predictive performance and model manageability, directly engaging with RQ1.1 and RQ1.2 by exploring how different algorithmic approaches can illuminate student behavioural patterns across various colleges.

4.5.4. Decision tree (REPTree)

As shown in Figure 43, predicting which college students are from based on LMS data involves analysing engagement metrics like Topic Videos and Participation, Forum Interactions, and Support Material Interactions. These attributes suggest that different colleges may have distinct learning environments and expectations, as indicated by students' engagement patterns. Accuracy is generally very good, with some paths achieving 100% accuracy (predicting S&E, and NHS for example). Compared to the grade predictions of Experiment 1, these results are significantly more accurate.

College decision tree

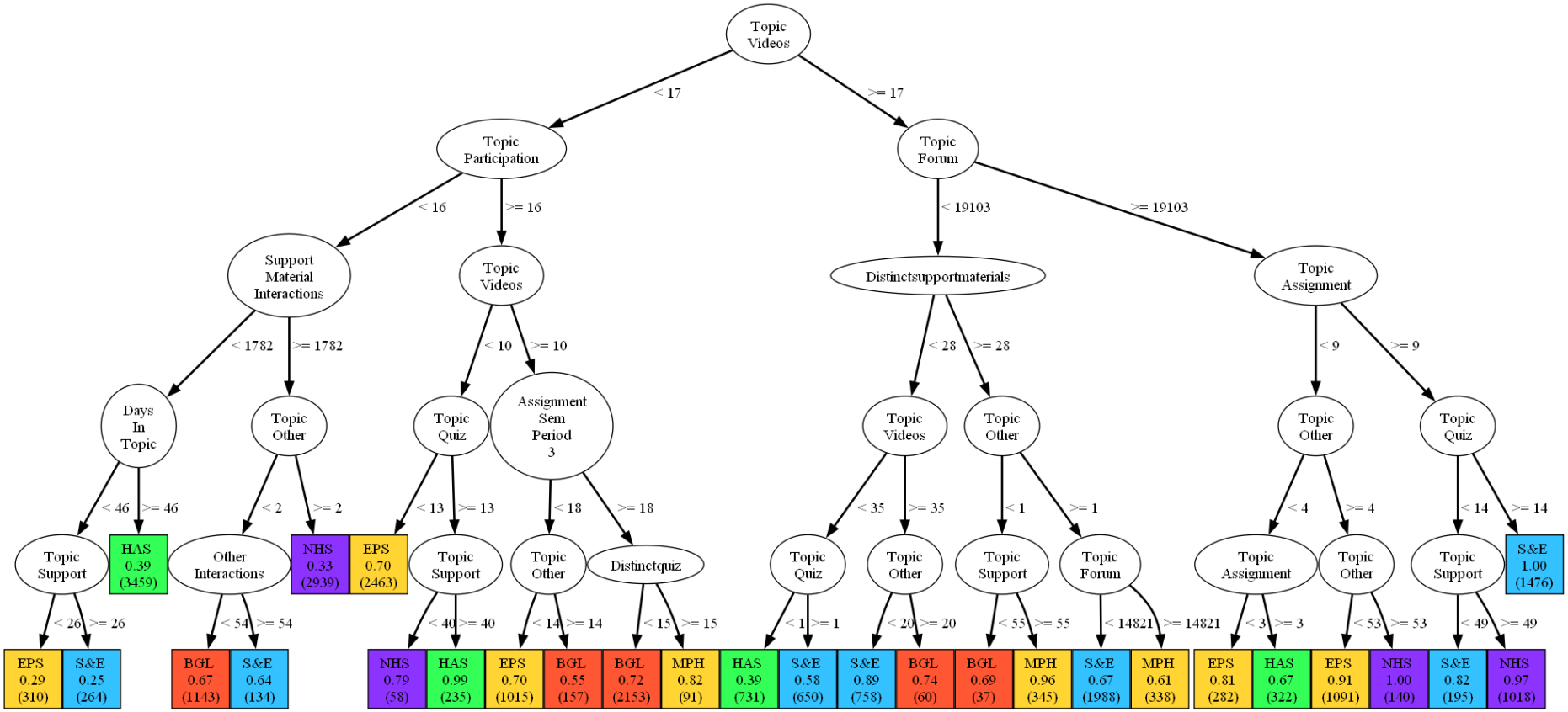

Figure 43 - Pruned college decision tree (reduced to 5 levels)

4.5.5. Experiment 2 results summary

The results of Experiment 2, which aimed to discern college affiliations with and without the grade attribute, highlight the ability of the chosen machine learning algorithms to identify distinct educational and engagement patterns across colleges. High-performing algorithms such as SimpleCART, NBTree, and AdaBoost exhibited robust predictive accuracy, suggesting that LMS behaviour is a significant predictor of college affiliation. This finding directly addresses Research Question RQ1.3, illustrating that specific patterns of LMS interaction can indicate a student's college, despite the complexity of educational data. The nuanced improvements in algorithm performance, observed when excluding the grade attribute, further emphasise the importance of focusing on behavioural patterns rather than solely on academic outcomes.

4.6. Experiment 3 results: Predictive analytical models for E-Learning by discipline

4.6.1. College of Business, Government, and Law

The performance metrics for the College of Business, Government, and Law (BGL) (Table 40) reveal that RandomForest demonstrates the highest effectiveness in predicting grades for particular college affiliations, with superior Accuracy, Weighted Avg. Kappa, and AU ROC scores. RotationForest and SimpleCART also show strong performance, closely followed by REPTree and J48 (RT Prune), indicating a good balance between accuracy and model interpretability. The lower performance by NaiveBayes and Decision Stump suggests challenges in capturing the complexity of behavioural patterns within BGL using simpler models.

Table 40 - Algorithm comparison (BGL)

BGL	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.39	0.18	0.18	0.67
DStump	0.35	0.10	-	0.56
RandomTree	0.32	0.12	0.12	0.56
J48	0.38	0.18	0.17	0.65
J48 (RT Prune)	0.39	0.18	0.18	0.65
NaiveBayes	0.24	0.09	0.08	0.60
SimpleCART	0.40	0.19	0.19	0.67
RandomForest	0.42	0.21	0.21	0.70
RotationForest	0.41	0.19	0.20	0.69
NBTree	0.33	0.18	0.14	0.64
AdaBoost (RT)	0.39	0.10	0.18	0.67

Comparison to Experiment 1

Comparing the performance metrics for BGL with Experiment 1, the results exhibit only marginal differences across most algorithms. Ensemble methods, particularly Random Forest and Rotation Forest, continue to show slightly higher Accuracy and Kappa scores.

However, simpler models like REPTree, J48, and SimpleCART maintain their performance levels, offering similar results between experiments. Algorithms such as Decision Stump, Random Tree, and Naive Bayes remain less effective in both scenarios, highlighting their limitations in capturing the nuanced patterns necessary for accurate grade prediction.

Overall, the comparison suggests that while ensemble methods provide marginally better results, suggesting that simpler models like REPTree and SimpleCART are useful as more interpretable models without sacrificing a significant level of performance.

Decision tree analysis

For BGL (Figure 44), overall interaction with participation content in later weeks of the semester, forum activity, and overall engagement with the topic appear to be very important. This suggests a strong emphasis on both social and academic aspects of learning within the LMS. The decision tree suggests the importance of active learning and community engagement as predictors of academic success. Accuracy is mixed, with paths achieving up to 95% accuracy (again notably higher predicting Fail grades).

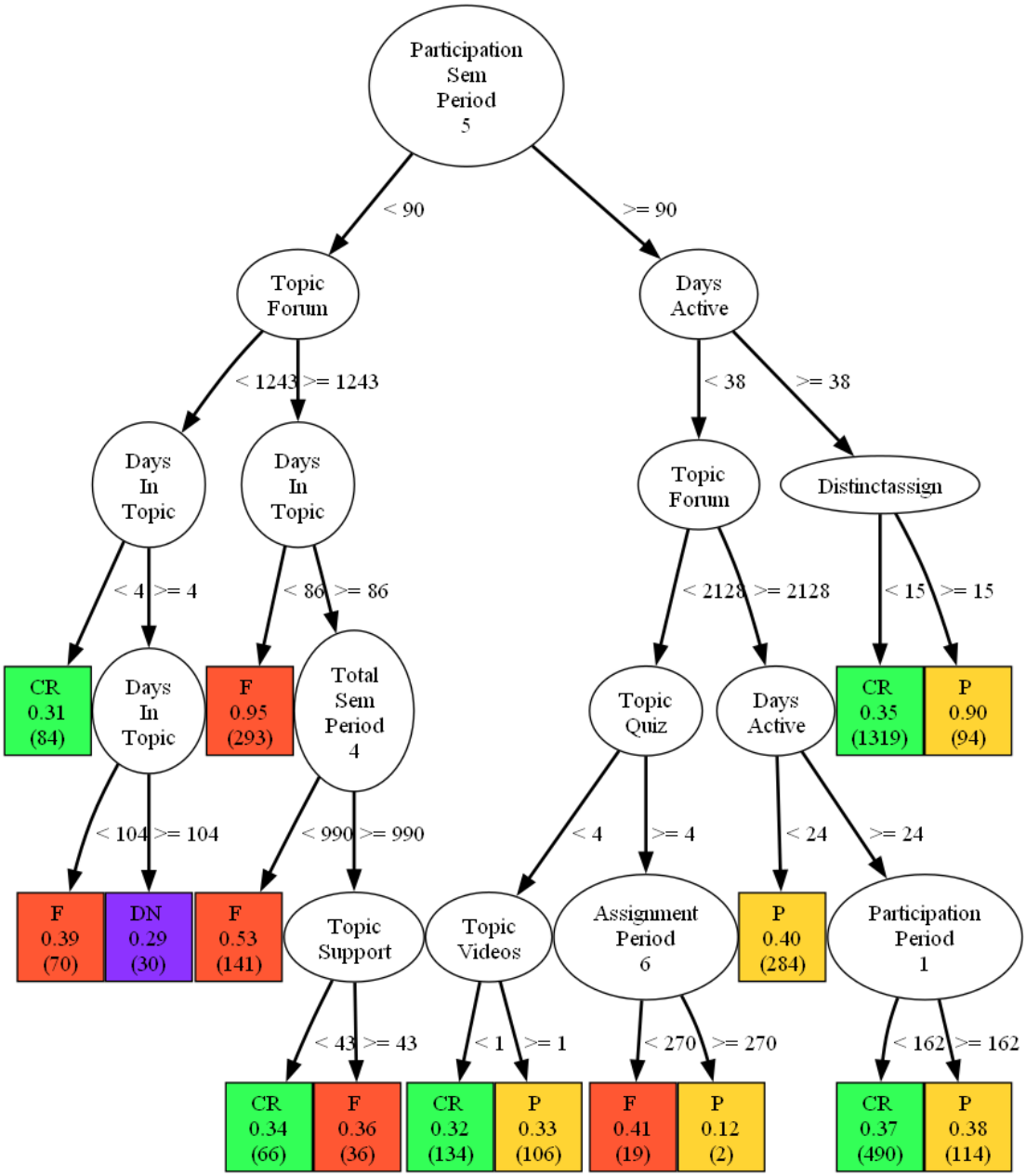

Figure 44 - Predictive analytical model for E-Learning in BGL: Pruned grade decision tree for BGL (reduced to 5 levels)

4.6.2. College of Education, Psychology, and Social Work

The performance metrics for the College of Education, Psychology, and Social Work (EPS) shown in Table 41, indicate strong algorithm effectiveness, with Weighted Avg. Kappa greater than 0.3 for most algorithms, suggesting that close to a third of the time student outcomes can be predicted just from their use of FLO. Performance wise, RandomForest leads in predictive accuracy, followed closely by SimpleCART and RotationForest. Other algorithms, such as REPTree, J48, and J48 (RT Prune), also perform well, demonstrating a good balance between accuracy and model simplicity. The lower performance by NaiveBayes and Decision Stump highlights their limitations in capturing the nuanced behavioural patterns associated with EPS.

Table 41 - Algorithm comparison (EPS)

EPS	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.46	0.30	0.29	0.73
DStump	0.33	0.11	-	0.60
RandomTree	0.37	0.20	0.19	0.59
J48	0.46	0.31	0.29	0.73
J48 (RT Prune)	0.47	0.31	0.30	0.73
NaiveBayes	0.22	0.10	0.09	0.61
SimpleCART	0.47	0.32	0.31	0.74
RandomForest	0.49	0.33	0.33	0.76
RotationForest	0.48	0.31	0.31	0.76
NBTree	0.36	0.30	0.22	0.69
AdaBoost (RT)	0.46	0.11	0.29	0.73

Comparison to Experiment 1

EPS exhibits improved performance metrics across most algorithms compared to Experiment 1, particularly in Accuracy and Weighted Avg. Kappa. RandomForest and SimpleCART show exceptional performance, with notable improvements over their Experiment 1 results. For instance, SimpleCART demonstrates a measurable improvement in Accuracy, while J48 (RT Prune) achieves a significant increase in Weighted Avg. Kappa.

Decision tree (REPTree)

For EPS (Figure 45), Assignments, and general engagement with the LMS are shown to be most important. Overall accuracy is mixed, but generally high, with several paths achieving above 90% accuracy, and the Fail path continuing the trend of being very accurate.

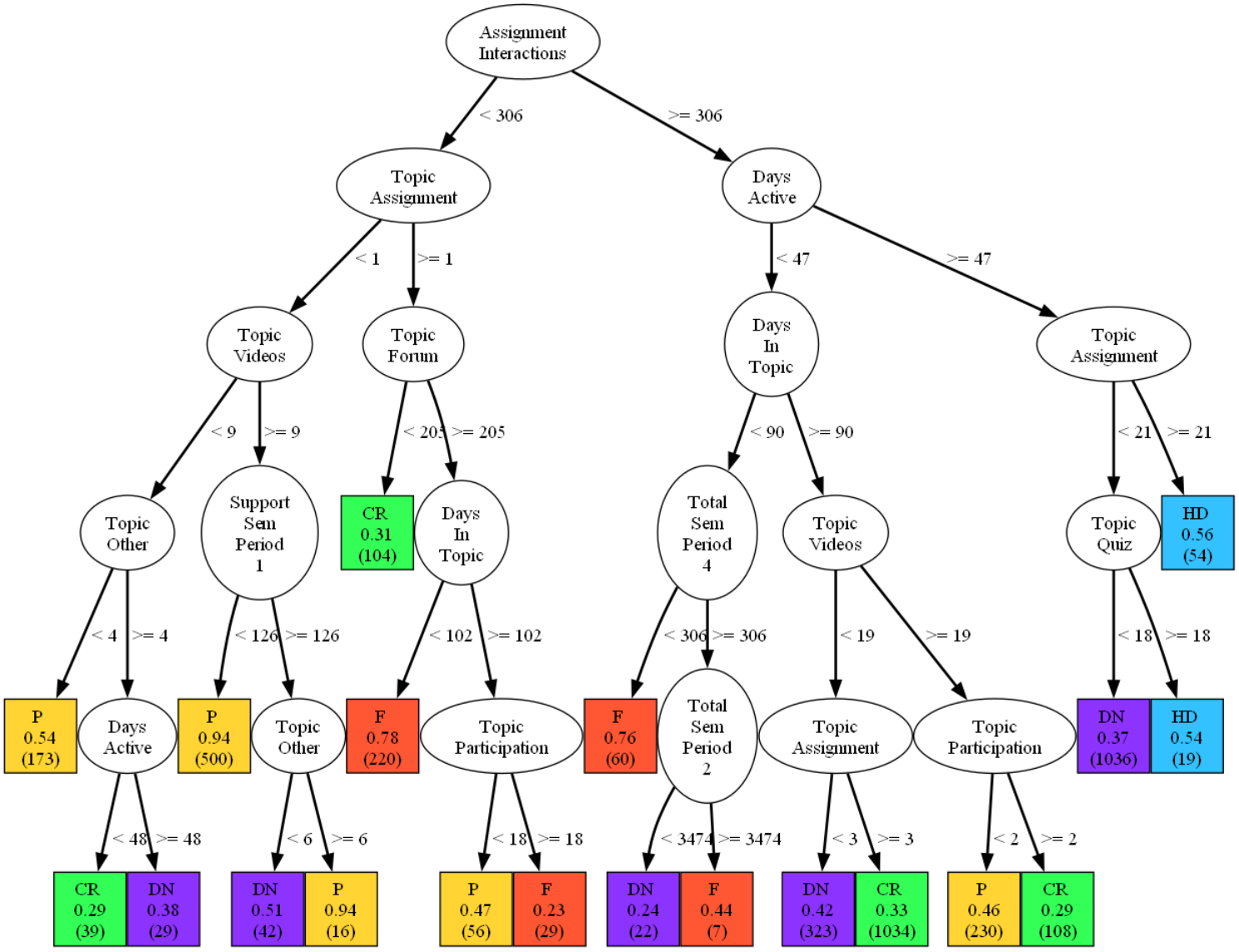

Figure 45 - Predictive analytical model for E-Learning in EPS: Pruned grade decision tree for EPS (reduced to 5 levels)

4.6.3. College of Humanities, Arts, and Social Sciences

The performance metrics for College of Humanities, Arts, and Social Sciences (HAS) shown in Table 42, suggest that RandomForest and RotationForest demonstrate the highest predictive performance, with Accuracy, Weighted Avg. Kappa, and AU ROC slightly outperforming other algorithms. Other algorithms, such as REPTree, J48, and SimpleCART, perform similarly to their outcomes in previous tests.

Table 42 - Algorithm comparison (HAS)

HAS	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.39	0.18	0.18	0.67
DStump	0.33	0.11	-	0.57
RandomTree	0.31	0.12	0.11	0.55
J48	0.37	0.17	0.17	0.65
J48 (RT Prune)	0.38	0.18	-	0.65
NaiveBayes	0.23	0.08	0.07	0.59
SimpleCART	0.39	0.19	-	0.66
RandomForest	0.40	0.20	0.19	0.69
RotationForest	0.40	0.19	-	0.69
NBTree	0.30	0.18	0.13	0.62
AdaBoost (RT)	0.39	0.11	0.19	0.67

Comparison to Experiment 1

Comparing these results to Experiment 1, HAS's performance metrics are slightly lower in certain aspects but remain competitive. Decision Stump and NaiveBayes show improvements in their Accuracy and Kappa scores, with Decision Stump improving by +3.44 in Accuracy and +0.03 in Kappa, and NaiveBayes improving by +5.46 in Accuracy and +0.02 in Weighted Avg. Kappa.

Decision tree (REPTree)

For HAS (Figure 46), general interactions in the later part of the semester, along with assignments and general engagement with the topic appear to be most important. Overall accuracy is lower than the previous colleges, with again, a notable exception of the Fail paths. Most paths are still better than chance, except for one DN path that is shown to be 19% accurate.

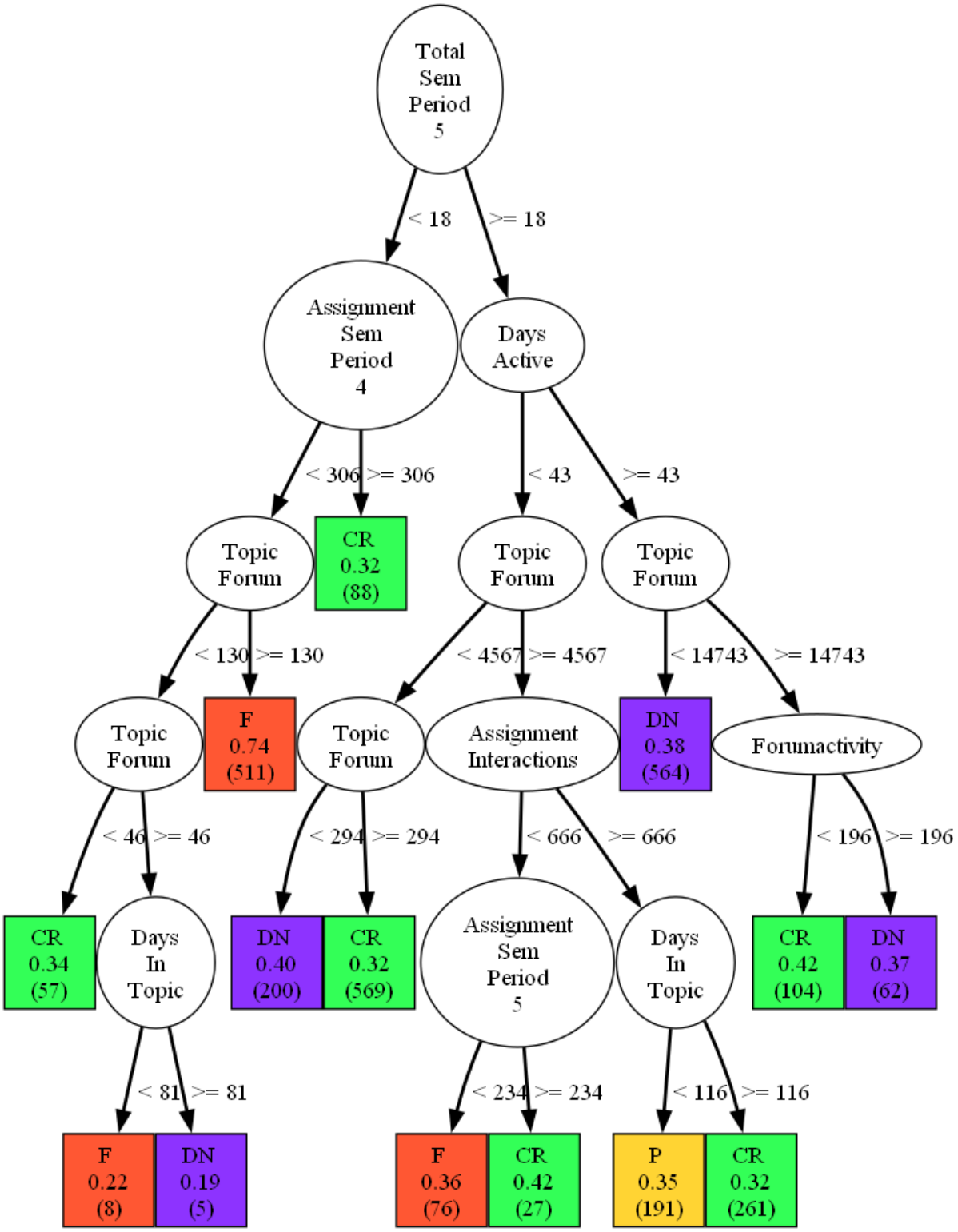

Figure 46 - Predictive analytical model for E-Learning in HAS: Pruned grade decision tree for HAS (reduced to 5 levels)

4.6.4. College of Medicine, and Public Health

For the College of Medicine and Public Health (MPH), the results as shown in Table 43, suggest that RandomForest stands out with the highest performance metrics, indicating strong predictive capability for this college's affiliation. Again, other algorithms performing as expected, and no unusual occurrences.

Table 43 - Algorithm comparison (MPH)

MPH	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.42	0.24	0.23	0.69
DStump	0.34	0.13	-	0.59
RandomTree	0.35	0.19	0.17	0.59
J48	0.40	0.24	0.23	0.69
J48 (RT Prune)	0.41	0.23	0.22	0.68
NaiveBayes	0.27	0.16	0.15	0.64
SimpleCART	0.42	0.25	0.24	0.69
RandomForest	0.45	0.29	0.29	0.74
RotationForest	0.43	0.23	0.24	0.72
NBTree	0.35	0.24	0.20	0.68
AdaBoost (RT)	0.42	0.13	0.24	0.69

Comparison to Experiment 1

Comparing to Experiment 1, the predictive accuracy for MPH shows significant increases across all algorithms and metrics. The largest increases are in Accuracy and Weighted Avg. Kappa, with improvements of +4.87 and +0.06, respectively. These results suggest that the distinct patterns of LMS interaction within MPH are well-defined and recognisable by machine learning models, enabling more effective prediction.

Decision tree (REPTree)

MPH (Figure 47), shows a unique decision tree, with the most important aspects of the tree being focused on topic composition, specifically regarding support materials and forum posts. Accuracy is slightly mixed, but generally very high, with one Fail path receiving 100% accuracy.

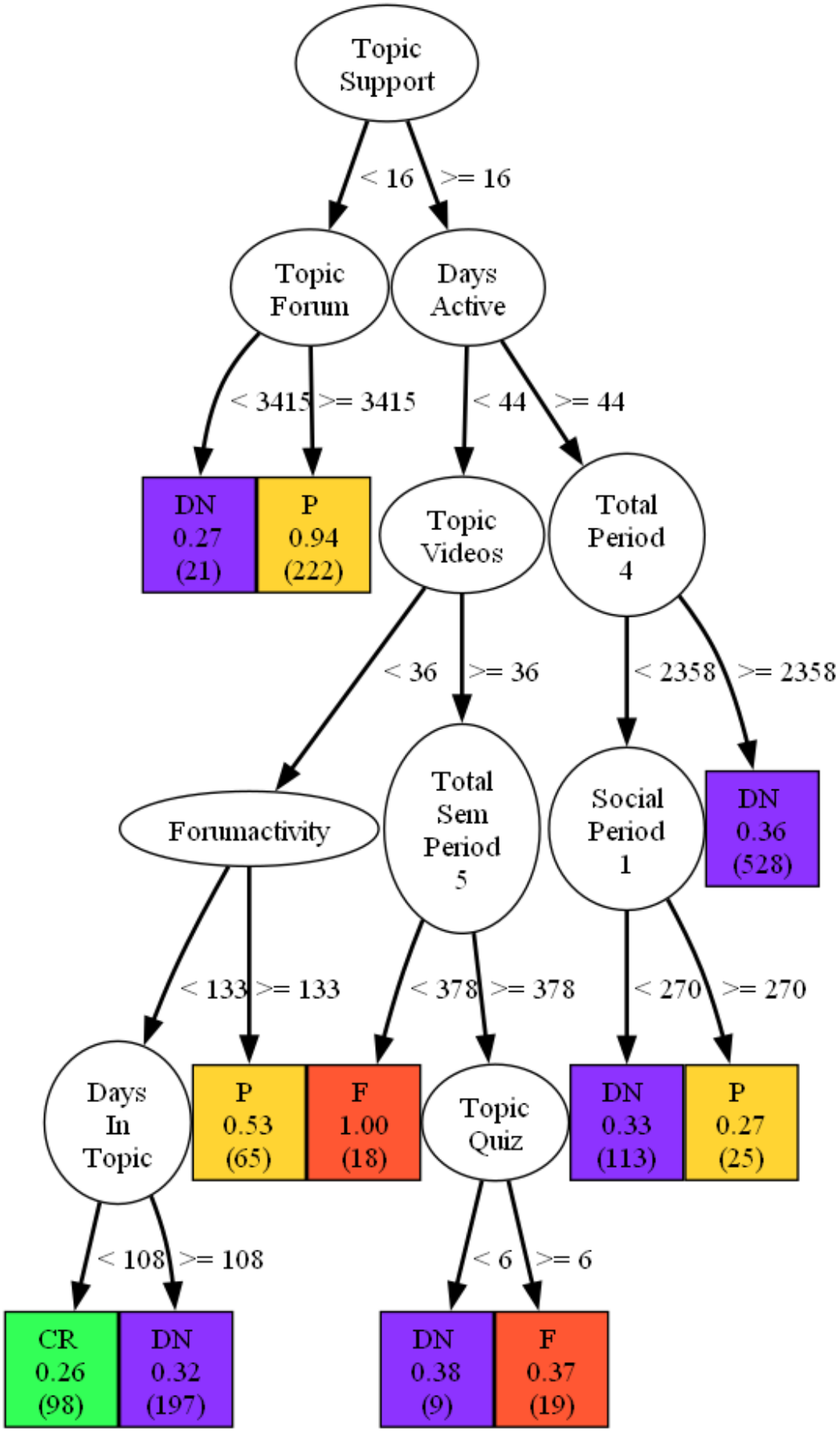

Figure 47 - Predictive analytical model for E-Learning in MPH: Pruned grade decision tree for MPH (reduced to 5 levels)

4.6.5. College of Nursing, and Health Sciences

The College of Nursing and Health Sciences (NHS) results shown in Table 44, sees RandomForest leading in predictive performance, with other algorithms performing as expected.

Table 44 - Algorithm comparison (NHS)

NHS	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.40	0.17	0.17	0.65
DStump	0.32	-	-	0.53
RandomTree	0.33	0.12	0.11	0.56
J48	0.40	0.19	0.18	0.65
J48 (RT Prune)	0.40	0.18	0.18	0.65
NaiveBayes	0.19	0.07	0.07	0.57
SimpleCART	0.41	0.18	0.18	0.65
RandomForest	0.44	0.22	0.23	0.70
RotationForest	0.42	0.18	0.19	0.69
NBTree	0.32	0.17	0.14	0.63
AdaBoost (RT)	0.40	0	0.18	0.66

Comparison to Experiment 1

Comparing this with Experiment 1, the performance metrics for NHS are improved across most metrics, with mild increases in Accuracy and Weighted Avg. Kappa. However, there were slight decreases in Weighted Avg. MCC and AU ROC. No specific algorithm demonstrated significant outperformance, but overall performance improvements were observed.

Decision tree (REPTree)

NHS (Figure 48), shows that general engagement with the topic, especially with social interaction materials during the period of 9am to 11:59am, and the number of forum posts within the topic are critical for this college. Overall accuracy is mixed, with several paths achieving above 90% accuracy (again, notably one of the Fail paths). While remaining better than chance predictors for the lower accuracy paths.

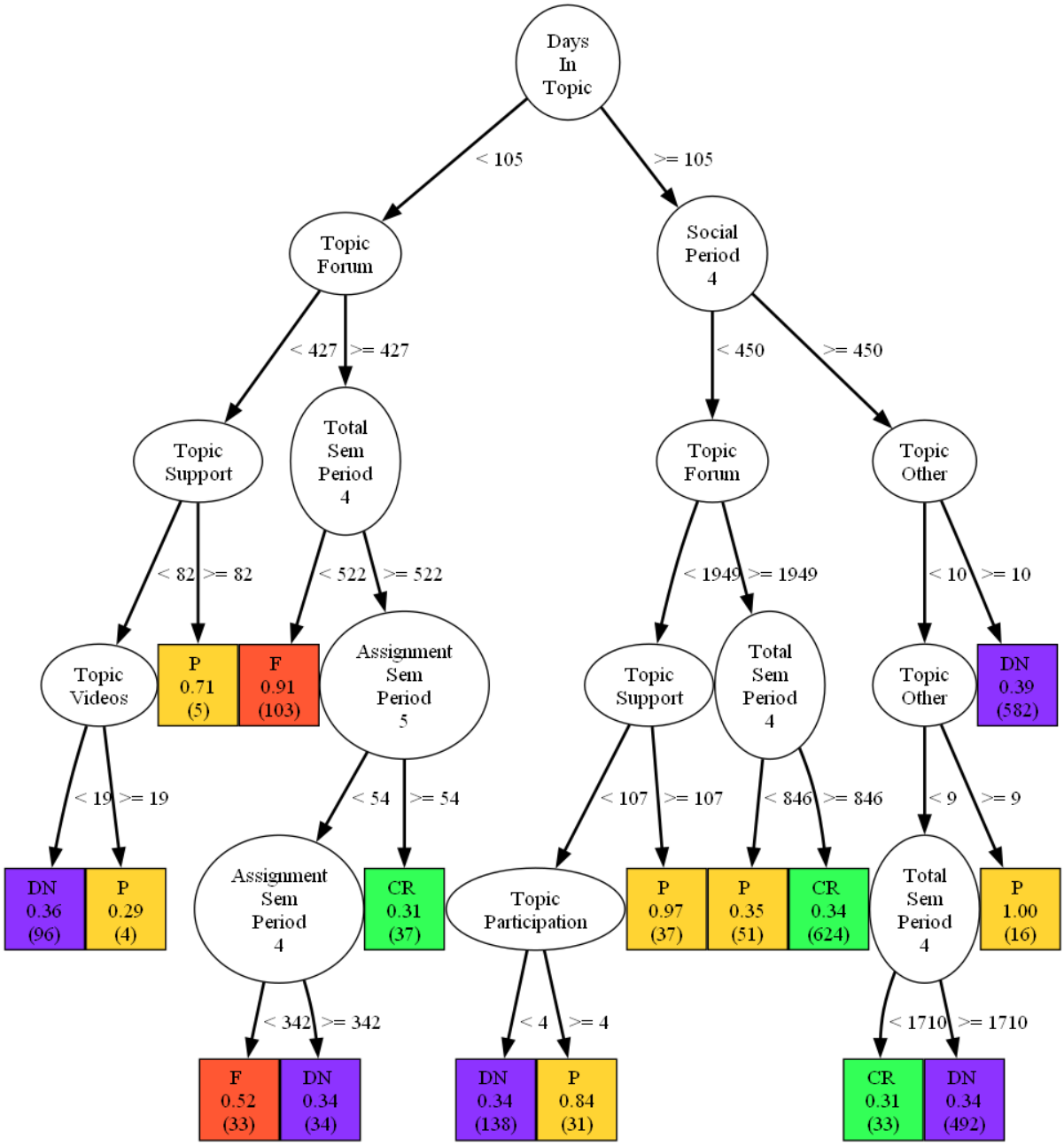

Figure 48 - Predictive analytical model for E-Learning in NHS: Pruned grade decision tree for NHS (reduced to 5 levels)

4.6.6. College of Science and Engineering

For the College of Science and Engineering (S&E), results shown in Table 45, indicate that RandomForest outperforms other algorithms with the highest Accuracy and AU ROC scores. SimpleCART and RotationForest also show good performance, indicating their effectiveness in capturing the unique characteristics of S&E student engagement with the LMS.

Table 45 - Algorithm comparison (S&E)

S&E	**Accuracy**	**Weighted Avg. Kappa**	**Weighted Avg. MCC**	**Weighted Avg. AU ROC**
REPTree	0.35	0.18	0.19	0.67
DStump	0.30	0.12	-	0.58
RandomTree	0.31	0.14	0.13	0.57
J48	0.35	0.19	0.19	0.67
J48 (RT Prune)	0.36	0.20	0.19	0.67
NaiveBayes	0.26	0.09	0.09	0.60
SimpleCART	0.37	0.22	0.21	0.69
RandomForest	0.40	0.25	0.25	0.72
RotationForest	0.37	0.21	0.21	0.70
NBTree	0.32	0.18	0.15	0.64
AdaBoost (RT)	0.35	0.12	0.19	0.67

Comparison to Experiment 1

Comparing the S&E results to Experiment 1, performance was mixed, while RandomForest maintained a strong performance in both scenarios, most metrics showed slight decreased. The notable exception to this was NaieveBayes showing +8.61, and +0.03, for Accuracy, and Kappa respectively. Conversely, REPTree and AdaBoost showed significant reductions in Weighted Average Kappa (-0.02, and -0.08 respectively), potentially indicating that some complexities within S&E are not being effectively addressed by these algorithms.

Decision tree (REPTree)

S&E (Figure 49), shows that a focus on engagement with the LMS, as well as interactions with participation content during the exam period is crucial for this college. Accuracy is moderate, with the notably large Fail path as other colleges showed, being the highest.

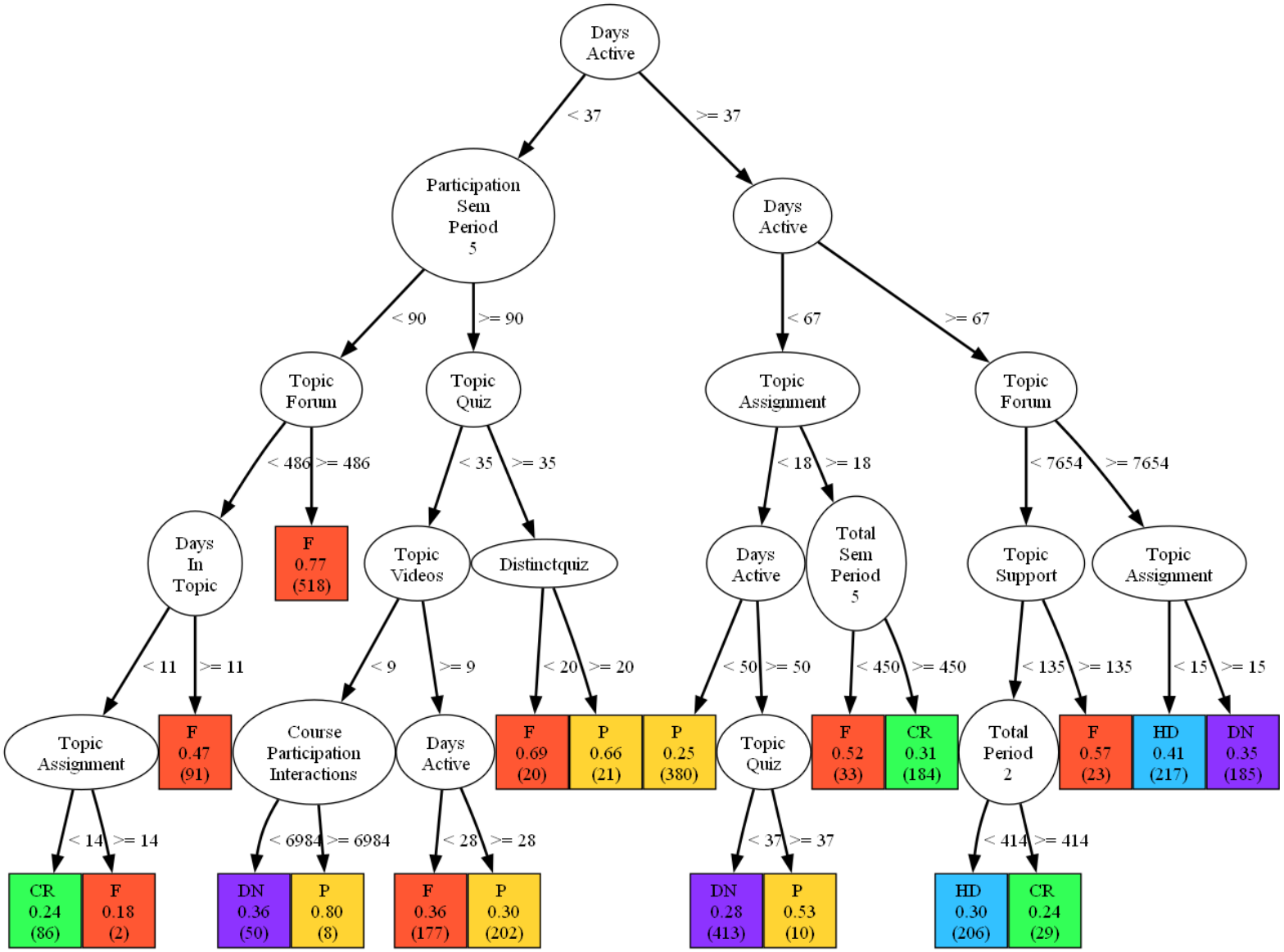

Figure 49 - Predictive analytical model for E-Learning in S&E: Pruned grade decision tree for S&E (reduced to 5 levels)

4.6.7. Experiment 3 results summary

College of Business, Government, and Law

For BGL, the algorithms that performed the best were RandomForest and RotationForest, showing high effectiveness with superior Accuracy, Kappa, and AU ROC scores. Additionally, SimpleCART also demonstrated strong performance, while NaiveBayes and Decision Stump showed poorer performance. BGL shows an emphasis on social and academic aspects of learning, with participation content in later weeks, forum activity, and engagement being important.

College of Education, Psychology, and Social Work

For EPS, both RandomForest and SimpleCART were the best-performing algorithms, with significant increases in Accuracy and Kappa compared to Experiment 1. Regarding attributes, assignments and general engagement with the LMS are emphasised as being most important.

College of Humanities, Arts, and Social Sciences

For HAS, the algorithms that performed the best were RandomForest and RotationForest, with metrics comparable to previous tests and only slight decreases in some areas. However, there were notable increases in Accuracy for Decision Stump (+3.44) and NaiveBayes (+5.46) compared to Experiment 1. The importance of interactions in later parts of the semester, assignments, and general engagement with the topic were shown to be critical for this college.

College of Medicine, and Public Health

For MPH, the top-performing algorithm was RandomForest, with increases across all algorithms compared to Experiment 1, indicating strong predictive capability. Notably, RandomForest achieved a +4.87 improvement in Accuracy and +0.06 in Kappa. The focus on topic composition, particularly support materials and forum posts, was identified as important for prediction.

College of Nursing, and Health Sciences

For NHS, the top-performing algorithm was RandomForest, with overall mild increases in Accuracy and Kappa across all algorithms compared to Experiment 1. For attribute importance, general engagement with the topic, particularly with social interaction materials during specific times and the number of forum posts, was found to be critical.

College of Science and Engineering

Finally, for S&E, the top-performing algorithm was again RandomForest, showing strong performance in Accuracy and AU ROC. However, results were mixed compared to Experiment 1, with notable improvements for NaiveBayes (+8.61 in Accuracy and +0.03 in Kappa) and decreases for REPTree (-0.02 in Kappa) and AdaBoost (-0.08 in Kappa). Engagement with LMS content and participation interactions during the exam period were identified as key for predictive accuracy.

Overall

Regarding algorithms, RandomForest and RotationForest consistently showed high performance across colleges, indicating their robustness in predicting college affiliation based on LMS data. Additionally, the attribute importance analysis across colleges highlights the critical role of engagement and interaction with LMS content as predictors of academic success. Tailoring models to specific contexts and understanding attribute importance can enhance predictive accuracy.

4.7. Chapter summary

The chapter on results begins with a comprehensive statistical examination of student engagement and performance across different colleges, employing a wide variety of statistical techniques to unearth patterns and correlations. Each statistical test identified unique facets of student data, ranging from grade distributions and attendance trends to the dynamics of interactions within the Learning Management System (LMS), and contrasts these findings among various colleges.

An important discovery from analysing grade distributions was the existence of significant disparities in academic outcomes among colleges. This indicates potential variations in teaching methods or grading standards across colleges, which may point to inequities in grading practices or levels of student achievement. Furthermore, the investigation into student attendance and activity within the LMS revealed notable differences in engagement levels, with some colleges showing either higher or lower engagement rates. This variation in LMS interaction and attendance hints at distinctive academic cultures or differences in how resources are allocated among colleges.

The analysis also demonstrated that student activity over the semester is indicative of academic performance disparities. Moreover, exploring how student interactions within the LMS change at different times of the day provided insights into customised online learning strategies to enhance engagement for individual colleges.

Investigating the usage of LMS components and their link to academic performance unveiled a multifaceted relationship between engagement, resource use, and academic success within the LMS, especially when comparing across colleges. This underscores the educational diversity and pedagogical uniqueness inherent to each college.

The results section progresses to discuss the utilisation of Principal Component Analysis (PCA), which not only aided in addressing research question RQ1.4 but also served as an effective strategy for justifying the necessity for the large student dataset not being broken into a smaller subset of dimensions for analysis.

The concluding segments of the chapter delve into three main machine learning experiments, further reinforcing the themes identified through statistical and PCA analyses. Experiment 1 grouped all colleges to forecast grades, Experiment 2 aimed to predict college affiliation, and Experiment 3 leveraged insights from the previous experiments to deepen the understanding of inter-college differences and refine grade predictions on a college-specific basis. This structured approach reinforces the initial findings and underscores the complex educational landscapes across colleges.

5. Discussion

5.1. Chapter overview

This chapter will be divided into eight sections, which are structured as follows:

The first section will be an analysis of the common features present between colleges. This will utilise results gathered from Section 4.2 Exploratory data analysis and Section 4.3 Principal Component Analysis (PCA). This will outline which common attributes are best used for the classification of, and the description of the LMS interactions by students for all colleges.

Next, will be sections devoted to each of the six colleges, discussing unique or heavily weighted attributes identified. Colleges will be discussed in the following order: 1 - Business, Government, and Law, 2 - Education, Psychology, and Social Work, 3 - Humanities, Arts, and Social Sciences, 4 - Medicine, and Public Health, 5 - Nursing, and Health Sciences, and finally, 6 - Science and Engineering. These sections will utilise results from Section 4.2 Exploratory data analysis, 4.5 Experiment 2 results, and 4.6 Experiment 3 results. Specific attributes favoured by each college will be discussed in their section, as well as discussion of explanations such as differences in topic structure, delivery methods, student individual differences, and student activity will be discussed.

Finally, an overall perspective of the results will be discussed, with a focus on what the results indicate for universities, and colleges within universities.

5.2. Introduction

This study has made an important contribution of a predictive analytical model for E-Learning across disciplines. Unlike past literature focused on a single discipline; for example E-Learning in Mathematics (Borba et al. 2016), Biology (DiCarlo 2009), Engineering (Kolmos, Hadgraft & Holgaard 2016), or Computer Science (Papastergiou 2009), this study makes a theoretical contribution by analysing the nuances across a wide variety of disciplines. It examined the relationship between LMS usage and student performance across various academic disciplines by employing an analytical approach, systematically processing large datasets of student LMS interaction logs through advanced machine learning techniques, specifically decision-tree-based algorithms. These techniques were chosen for their interpretability and relatively comparative performance with more complex techniques, such as ensemble and black-box-style algorithms.

The research sought to isolate and understand the influence of LMS features and user engagement on academic outcomes, in the form of student grades. Notably, the study diverged from the more common approach of overall student performance prediction within a single discipline, focusing instead on discipline-specific insights across multiple disciplines, thereby addressing a notable gap in the literature.

The study's methodology involved quantitative analysis, allowing for the processing of large-scale student data to extract meaningful patterns and rules that could inform educational practice and policy. This approach positions the research at the forefront of E-Learning personalisation, contributing a novel perspective on the interplay between digital learning environments and pedagogical effectiveness.

The study meticulously analysed LMS usage patterns and student performance, uncovering discipline-specific insights. It highlighted significant variations in LMS feature interaction and academic outcomes across disciplines, indicating that a tailored approach to LMS design and pedagogical strategies is crucial for positive student outcomes. The findings advocate for a nuanced understanding of LMS utility in education, emphasising the need for personalised learning experiences aligned with individual disciplinary requirements.

It is worth noting that there may be personalised approached being implemented, but this is not able to be shown in the data. Therefore, the data-driven approach of this research is best used to identify patterns, and optimal/sub-optimal outcomes. This can be used to show what an optimal customisation process might look like as a guide for educators. Which can be measured against current implementations.

5.3. Common features across colleges

The research has identified marked variations in student activity within the LMS (in both type and volume of interaction) categorised by college. This would suggest that there are significant associations between colleges and levels of student engagement with the LMS. However, certain commonalities persist, such as general utilisation of the LMS across all colleges, although to varying extents and in different manners, hinting at a level of universal recognition of the importance of an LMS in the educational process. These results are in line with those of Davidoff & Jayusi (2024), who found distinct differences between different grouping of disciplines (education, social sciences, exact sciences, business administration, and engineering). While their results were self-reported by students via questionnaires, they do show different usages of the LMS as was shown in this research.

This general usage pattern can be seen in Figure 42 of Section 4.4.4, which shows the (reduced to five levels) decision tree for predicting grades. This showed several aspects of the attributes that lead to grade prediction across colleges.

The first aspect identified was that of general interactions with the LMS, this can be seen in attributes such as Days Active and Days in Topic. These features are also shown in the first principal component shown in Figure 26 of Section 4.3, and represents aspects such as general engagement across time.

The second aspect was that of general topic composition, as seen in attributes such as Topic Forum, Topic Participation, and Topic Videos. This aspect is reflected in the tenth principal component (Figure 35), which contrasts overall interaction with the LMS and of topic composition.

The third aspect was that of specific interaction types, as seen in attributes such as Activity Interactions and Assignment Interactions. This is also reflected in the second principal component (Figure 27), and fifth principal component (Figure 30), both of which represent assignment interactions, and other aspects of the LMS that load against them, such as videos, or support materials.

The fourth aspect was that of bulk interactions during specific time of day periods, as seen in attributes such as Total Period 2, and bulk interactions in specific periods as seen in attributes such as Total Sem Period 4 and Total Sem Period 5. This is mostly represented in principal component one (Figure 26), and eight (Figure 33), with their temporal focus.

Finally, there was the aspect of the intersections of attribute types, combining multiple aspects of the attributes, as seen in attributes such as Assignment Sem Period 4, Participation Period 1, Social Period 6 and Support Sem Period 4, combining temporal, as well as interaction type. Like the previous aspects mentioned, this is also reflected in principal component 1 (Figure 26), and eight (Figure 33), due to the engagement over time attributes mentioned.

With regards to negative outcomes (Figure 50), key attributes found for identifying failing students included general LMS interaction attributes such Days in Topic, being the first and most important node, suggesting that having more than 104 days in topic is critical to not failing.

The Topic Forum size attribute suggests that topics with larger numbers of forum posts would require more interactions in the period weeks before the exams (Total Sem Period 4), and increased activity with assignments in said period, as well as overall activity during the exams (Total Sem Period 5) to not fail. While students enrolled in topics with lower numbers of forum posts require more interaction with ‘activity’ type modules (Activity Interactions), and more interactions with support materials weeks before the exams (Support Sem Period 4), to not fail.

The accuracy of these Fail paths, as was shown in Experiment 2, are of reasonable levels, with largely more than chance for two of the paths, and a greater than 80% accuracy for the remaining.

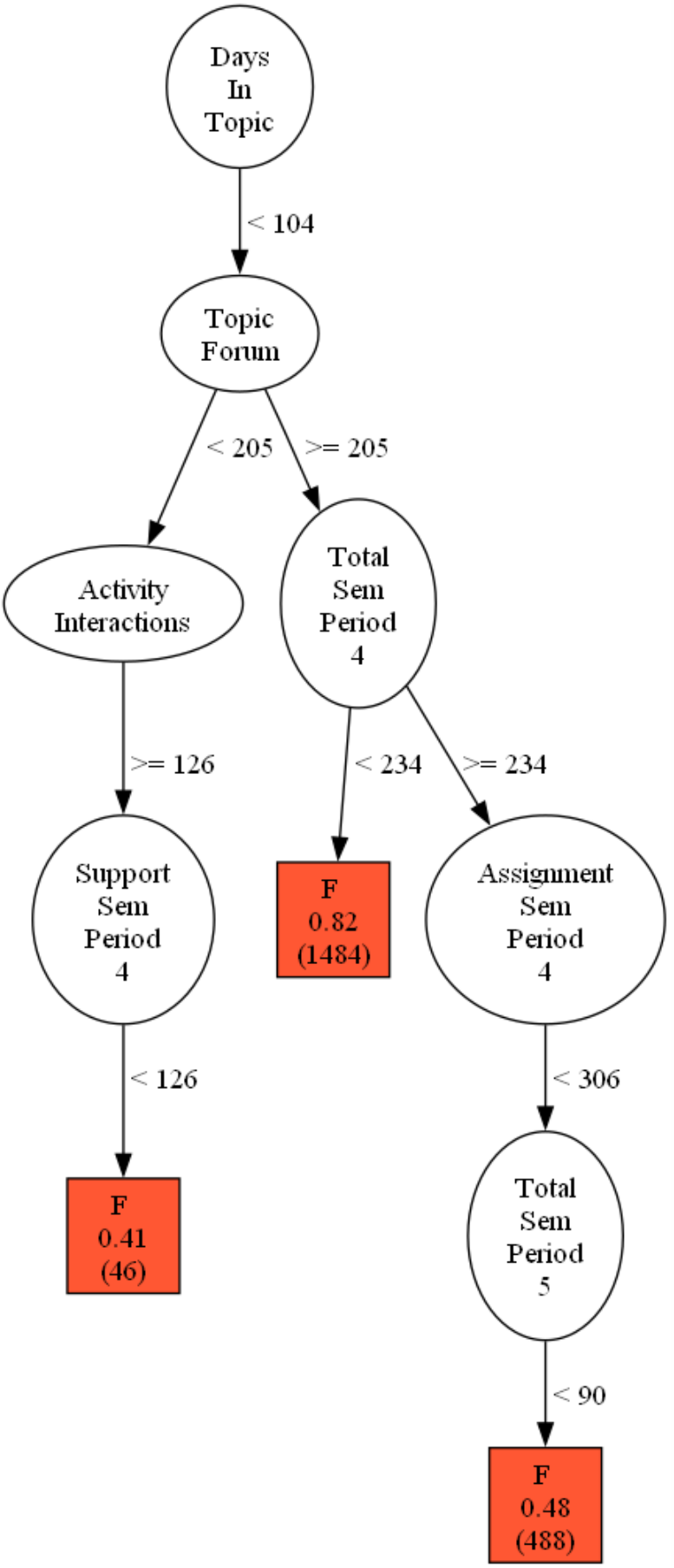

Figure 50 - Path for fail grades (reduced to 5 levels)

These general rules would suggest that across colleges, taking record of the total level of engagement with the topic, and especially at the end of the semester is important to identify at-risk students. In contrast, identifying high performing students is a more difficult task as shown in Figure 51, as the decision tree required the depth to be increased from five to seven, with the rest of the grade path trees only needing five levels to reach a grade outcome.

This complexity is evident in the required attributes for analysis. Specifically, attributes related to general LMS interactions; namely Days Active and Days in Topic, are of particular importance. Among these, Total Days in Topic emerges as a critical attribute, indicating a necessity for more than 104 days to achieve a High Distinction (HD), contrasting with Days Active which carries varying implications based on the scenario.

For topics abundant in support materials (Topic Support) and quizzes (Topic Quiz), HD students are characterised by fewer Days Active. These students also demonstrate minimal interaction with assignments yet engage significantly with over eight distinct participation components (Distinct Participation). As discussed by Bandura (2002), self-efficacy influences how learners allocate time, and what tasks are focused on defined by their perceived value. This would align with what was observed, with a minimal level of interaction, and selecting participation. Additionally, Felder & Silverman (1988) suggest that adaptive learners (such as HD students), are skilled at managing cognitive load and identifying tasks that can maximise learning efficiency.

In scenarios where students have a moderate number of Days Active, those who initially have fewer interactions with participation modules during Participation Period 1, but later access more assignment modules (Distinct Assign) and engage in a wide range of general LMS interactions (Other Interactions), stand a higher chance of achieving an HD. This likelihood increases if the topic features a substantial number of participation modules (Topic Participation). The research by Xie et al. (2019) suggest that in fact, high-performing students utilise personalisation to focus on content that would directly contribute to a positive outcome, potentially explaining the lower levels of engagement with abundant support materials.

This is also echoed by Avella et al. (2016), identifying that high performers often utilise LMS analytics to identify gaps in knowledge, targeting specific content areas. This would suggest that HD students in this case exhibit strategic engagement patterns with the topics that could be attributed to higher levels of self-efficacy.

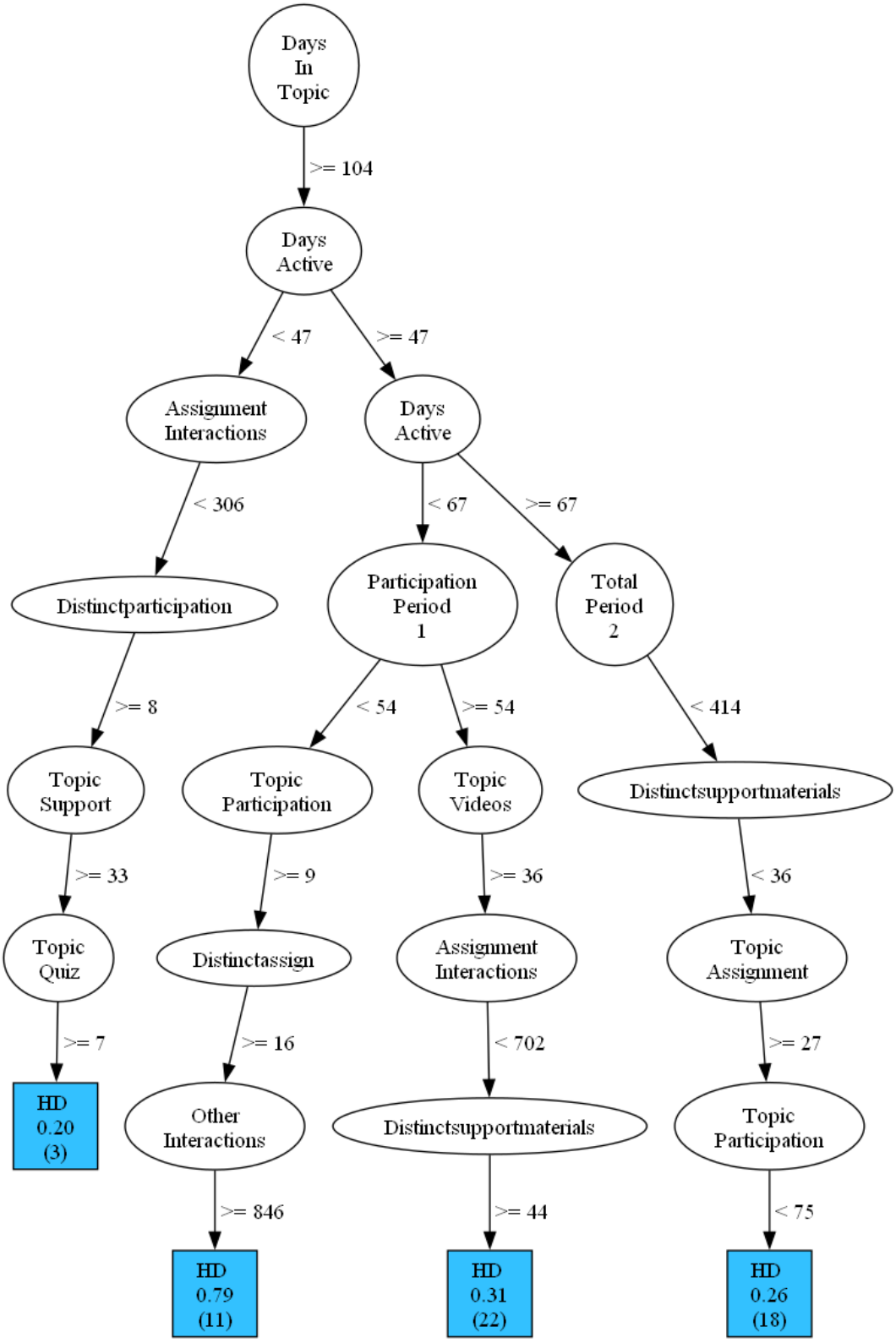

Figure 51 - Path for high distinction grades (reduced to 7 levels)

Alternatively, when students engage early with participation materials (Participation Period 1), the presence of topic videos becomes more important. Additionally, accessing a variety of unique support materials (Distinct Support Materials) and maintaining fewer repetitive interactions with assignments (Assignment Interactions) augments their potential for an HD.

Finally, students who are active over many days are more inclined towards achieving an HD if they limit their interactions early in the semester (Total Period 2), seek out unique support materials (Distinct Support Materials), and are enrolled in topics marked by a significant volume of assignments (Topic Assignment) with fewer participation modules (Topic Participation).

This suggests that all colleges have some common usage of general number of interactions with the LMS, and that that activity levels, and time enrolled is a factor with predicting student performance. In addition, the topic composition is notable in its inclusion, which will also be discussed in upcoming sections. This suggests that the composition of the topic itself is useful in predicting grades, in addition to or instead of student interactions. This finding is extremely useful for educators, as this is an aspect that can be controlled, as opposed to the behaviours of students. This feature will become important when discussing implications for educators, and recommendations in the subsequent chapter.

Overall accuracy of the paths shown are above chance, with one path reaching 79% accuracy. While the accuracy of this path is not as impressive as the Fail path, the HD path is still reasonably useful for in providing informative patterns in attribute usage and topic composition.

5.4. Discipline-specific analysis

Throughout this research, there has been considerable data to suggest that not only the activity of students between colleges differ (in the temporal aspect, as well as type and intensity of interaction), but that the topics themselves, as well as grade distribution differ significantly.

Considerable analysis was performed to show that each college is statistically significantly different in; grade distribution (Section 4.2.1), student attendance levels (Section 4.2.2), average student activity across the semester (Section 4.2.3), average student activity across daily time periods (Section 4.2.4), student utilisation of LMS components (Section 4.2.5), and overall topic composition (Section 4.2.6).

This data points to considerable variability in interaction patterns, LMS component usage, and topic content structure across colleges, suggesting that disciplines within each college have either tailored their approach to LMS utilisation to best suit their specific educational needs and pedagogical strategies, or have varied levels of expertise with LMS technologies, or requirements for said technologies within their course structure.

In addition to the statistical analysis, applying machine learning algorithms to the data confirmed differences between colleges in Experiment 2 (Section 4.5) and Experiment 3 (Section 4.6). In contrast to Experiment 1 (Section 4.4), these tests had excellent results for predicting college affiliation, suggesting that the differences between colleges are significant enough to be reliably used in classification.

The following sections will discuss these intra college differences, outlining unique or heavily weighted LMS attributes that contribute to college differentiation. Decision trees created in the following sections are extracted from larger decision trees (retaining college pathways, or grade pathways for colleges), due to size constraints and better interpretability.

5.4.1. Business, Government, and Law

The BGL decision tree path, as shown in Figure 52, highlights a strong emphasis on topic composition, with Topic Videos being particularly noteworthy, indicating a medium to large number of videos in BGL topics. Additional attributes of topic composition include Topic Forum (fewer than 19,103 posts), Topic Other (a small number of objects, unless accompanied by a large number of videos, or if there is low student engagement with assignments mid-semester), Topic Participation (more than 16 objects with a medium number of videos, or fewer when there is a low number of 'other' materials), and Topic Support (fewer than 55 objects).

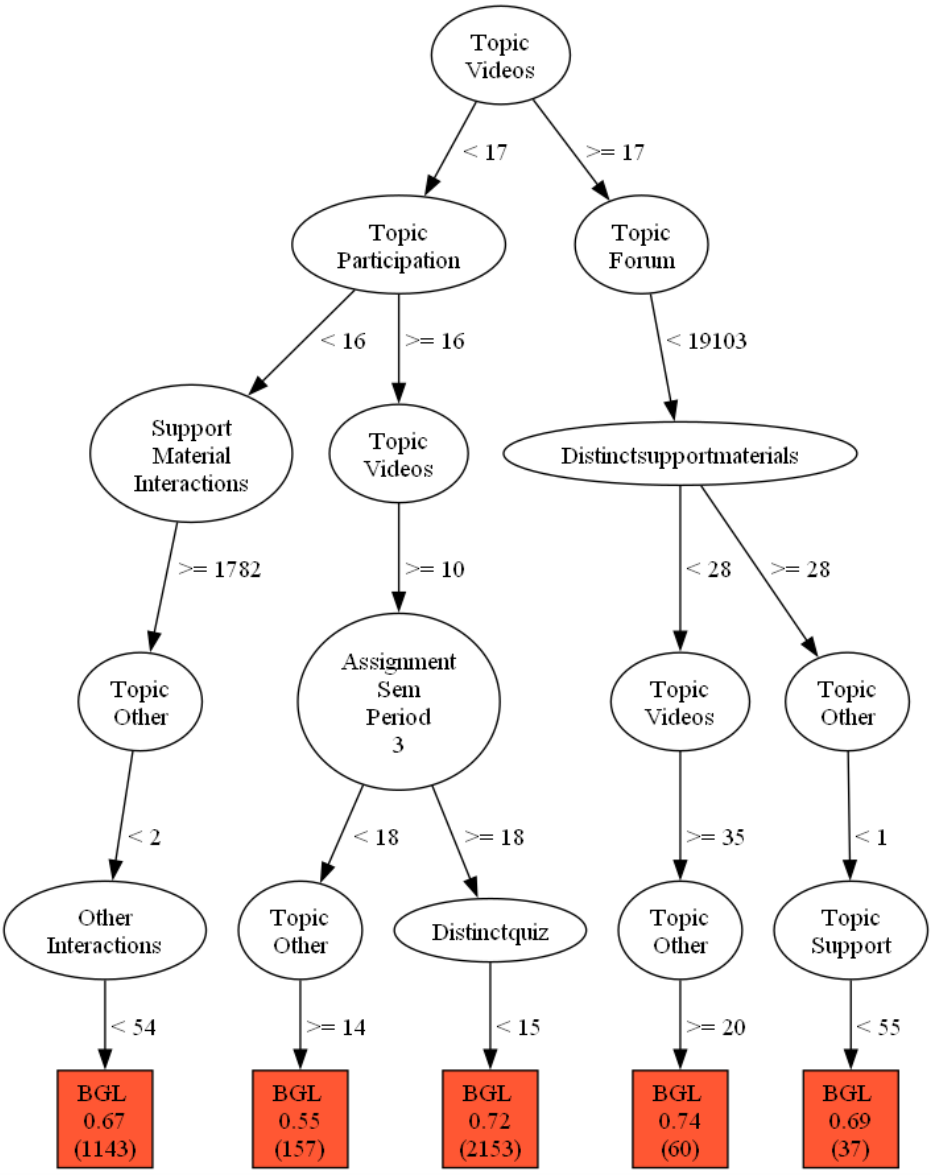

Figure 52 - Path for BGL college classification (reduced to depth of 5)

Other attributes implicated in the classification suggest a relatively low utilisation of materials on the LMS. For example, Distinct Support Materials (many objects viewed when 'Other' and support objects are low, and a low number of objects viewed when there is many videos), and overall, fewer than 15 Distinct Quiz objects viewed.

Accuracy of the paths are very high, with random chance being approximately 16% for predicting colleges. Showing as high as 74% accuracy, down to 55% accuracy is a positive result for the prediction of the BGL college, and the use of the topic composition attributes.

Identifying at-risk students within the BGL college involves a combination of topic, enrolment, and activity metrics, as shown in Figure 53. Interactions with participation objects towards the end of the semester play a significant role. Students with substantial activity at the end of the semester, who are not active throughout or have low numbers in enrolment attributes such as Days Active (fewer than 38 days) and Days in Topic, in addition to low Topic Forum posts and minimal evening assignment activity, suggest that low engagement is a strong predictor of a failing student.

Accuracy is mixed, but still relatively high, as mirrored by the results of Experiment 1 regarding the ease of predicting Fail grades.

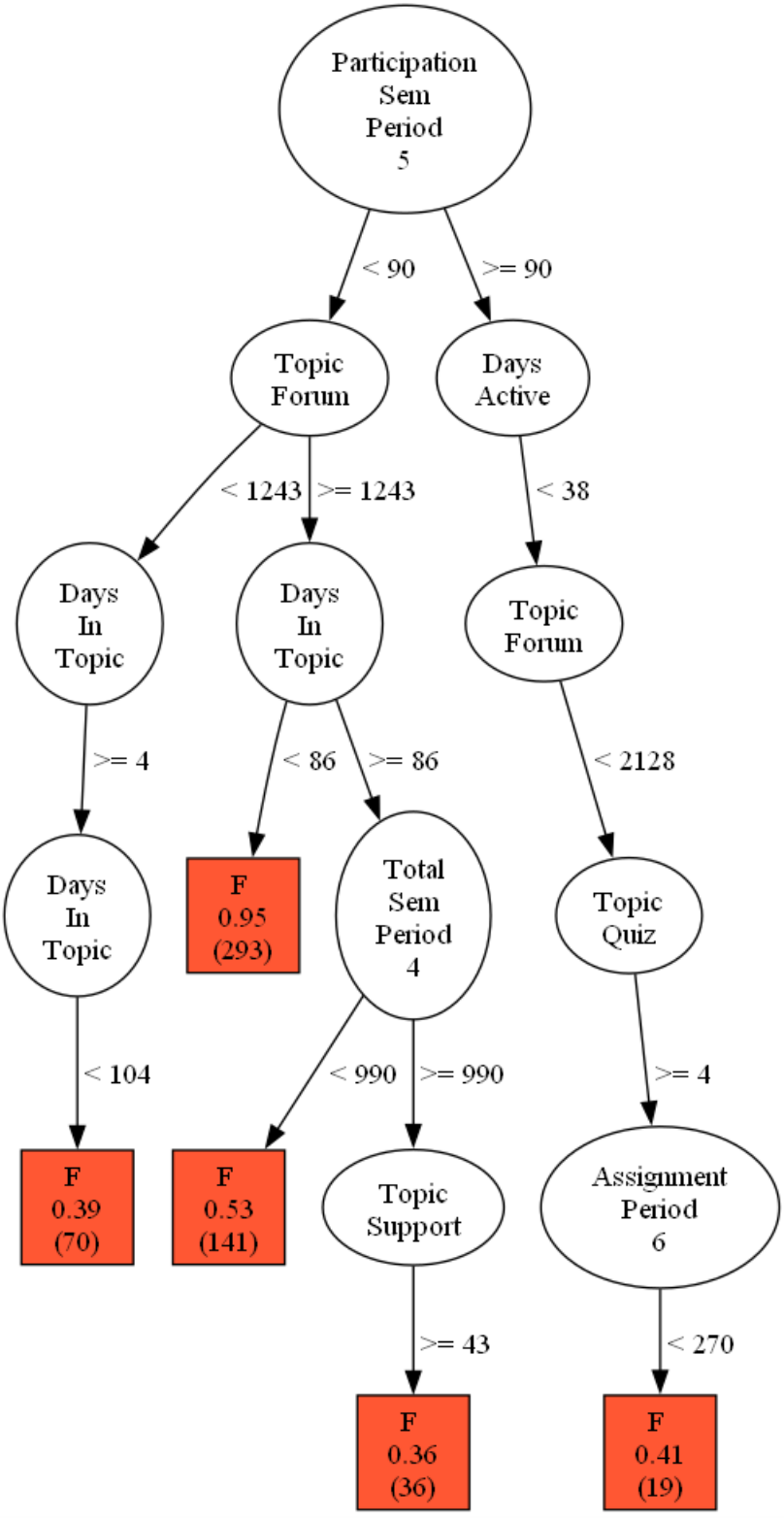

Figure 53 - BGL path for fail grades (reduced to 5 levels)

Classifying high achievers (HD students) for BGL is more straightforward to process, as indicated in Figure 54, yet requires at least ten levels to classify HD students accurately. This may suggest overfitting for HDs for the BGL College. Additionally, the accuracy shown for this path is reasonably better than chance (22% compared to 16%).

Concerning individual attributes, high values in Days Active, Participation Sem Period 5, Video Sem Period 3, and Video Sem Period 4 suggest that students engaging with videos and participation materials from mid-semester onwards, particularly in the last few weeks of the semester for participation objects, are more likely to achieve high distinctions.

As mentioned in Section 3.4 Exploratory data analysis, the BGL college demonstrated a higher number of Fail grades than expected and more students in the 'Low' activity, and fewer students in both ‘Medium’ and ‘High’ activity categories within the LMS. This observation aligns with the focus on Days Active and Days in Topic, and with a preference for activity later in the semester. Additionally, BGL along with EPS were shown to have a larger number of videos and assignments than all other colleges except S&E, which would appear to support what is shown in the HD outcome path.

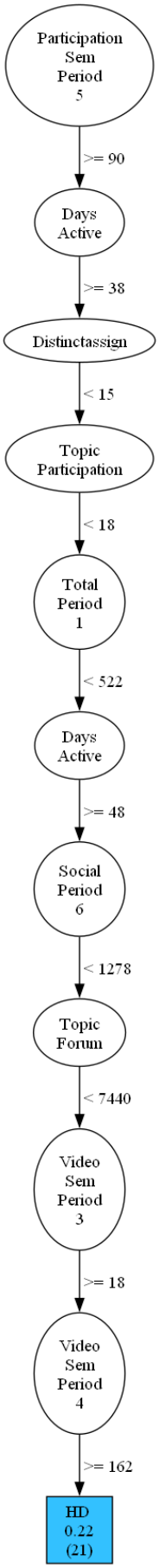

Figure 54 - BGL path for high distinctions (not reduced)

5.4.2. Education, Psychology, and Social Work

The EPS decision tree path as shown in Figure 55, primarily focuses on topic composition with only Assignment Sem Period 3, Days in Topic, and Support Material Interactions being mentioned outside of topic attributes. However, decision points are set for lower values, suggesting low activity and low time engaged with the LMS. Topic composition attributes such as Topic Videos along with Topic Participation both appear in positive decision points, whereas most other topic composition attributes are shown as negative decision points.

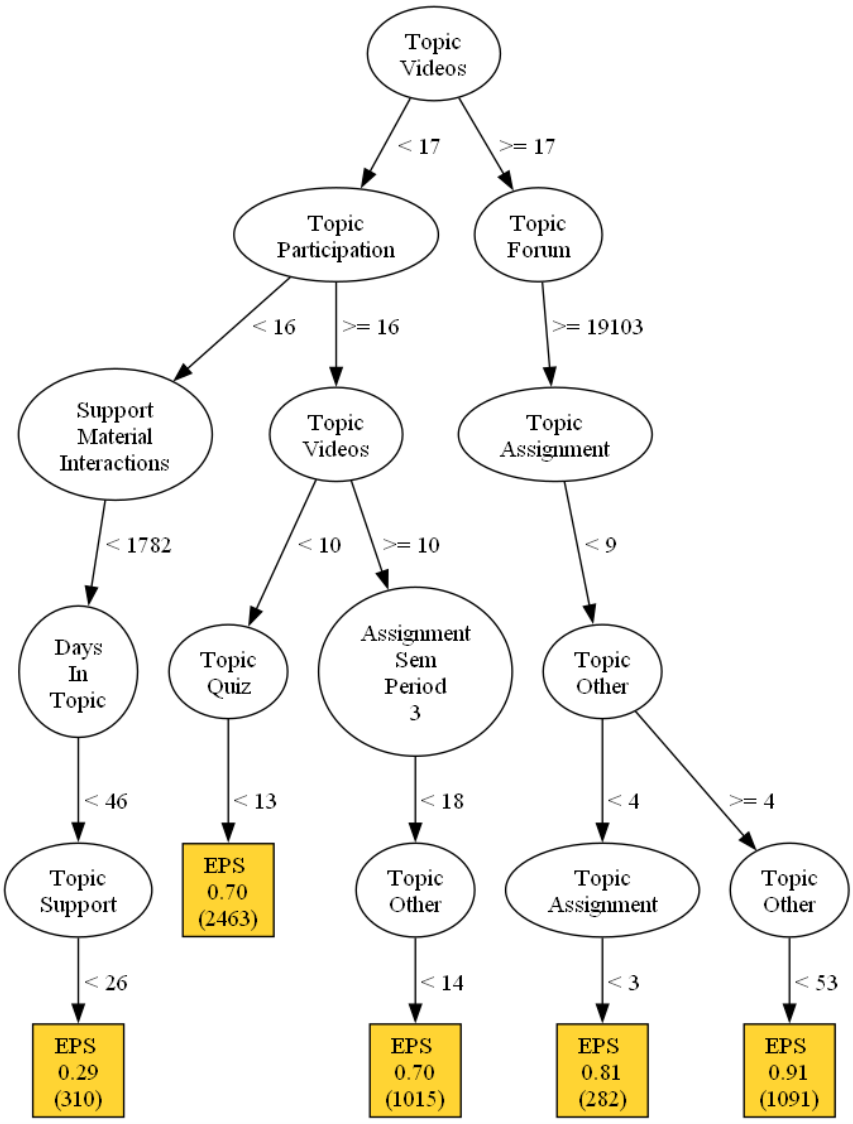

Figure 55 - Path for EPS college classification (reduced to depth of 5)

Regarding accuracy, each path is relatively high in accuracy, with only one being 29%, with the rest achieving greater than 70% accuracy. Suggesting topic composition is very much related to classifying affiliation with EPS.

As shown in Figure 56, Identifying at-risk students within the EPS college involves recognising those with limited activity, characterised by fewer than 47 days active in a topic. Additionally, a significant number of early semester interactions with assignments (Assignment Interactions during Total Sem Period 2) is also indicative.

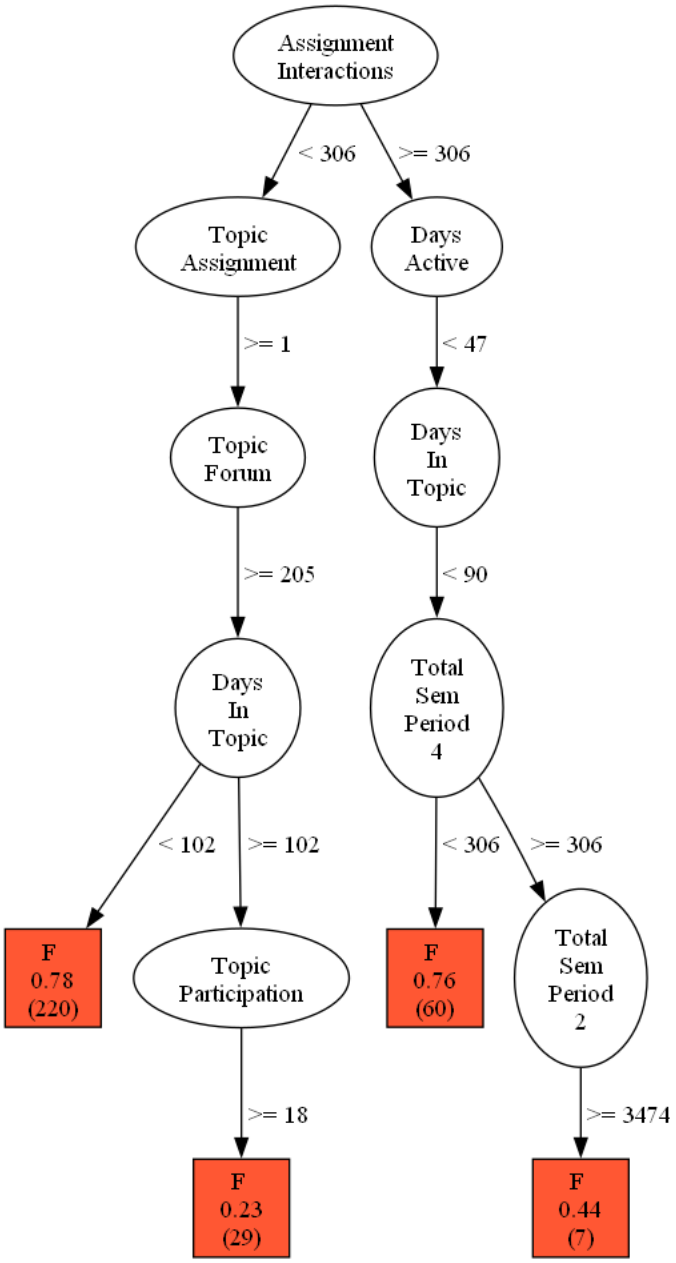

Figure 56 - EPS path for fail grades (reduced to 5 levels)

Conversely, students with more days spent on a topic, associated with topics that contain a relatively large number of participation objects (Topic Participation), are also noted. Accuracy is mixed for the paths shown, ranging from better than chance (23%), to moderate (44%), to relatively high above 70% accuracy.

As shown in Figure 57, classifying students with high distinctions tends to be simpler compared to identifying those at risk of failing. The decision tree primarily concentrates on topic composition. It identifies that students that have Assignment Interactions exceeding 306 and Days Active greater than 47, and that are enrolled in topics with a substantial number of Topic Assignments and Topic Quizzes, tend to achieve high distinctions.

Additionally, students with Days Active fewer than 47 who enrol early in a topic (indicated by Days in Topic) and engage with topics that have a smaller forum presence (Topic Forum) yet have larger quantities of topic materials (including Topic Participation, Topic Support, Topic Videos) are also more likely to earn high distinctions.

Accuracy of these HD path is higher than most, suggesting EPS is easier to predict in this regard, with each path being greater than 50%.

In EPS, the decision trees appear to indicate a correlation between engagement with the topic, as well as LMS components such as video content support content, and quizzes, and overall student performance. As described in Section 3.4 Exploratory data analysis, EPS was found to have a higher number of fail grades than expected, as well as similar activity levels as BGL (with more ‘Low’ activity students than expected, and fewer ‘Medium’ and ‘High’ activity students than expected). Additionally, EPS was shown to have a higher number of assignments, support material, forum activity, and participation materials. This supports the suggestion of increased activity levels, especially with the more limited number of quizzes and videos that are often associated with these topics.

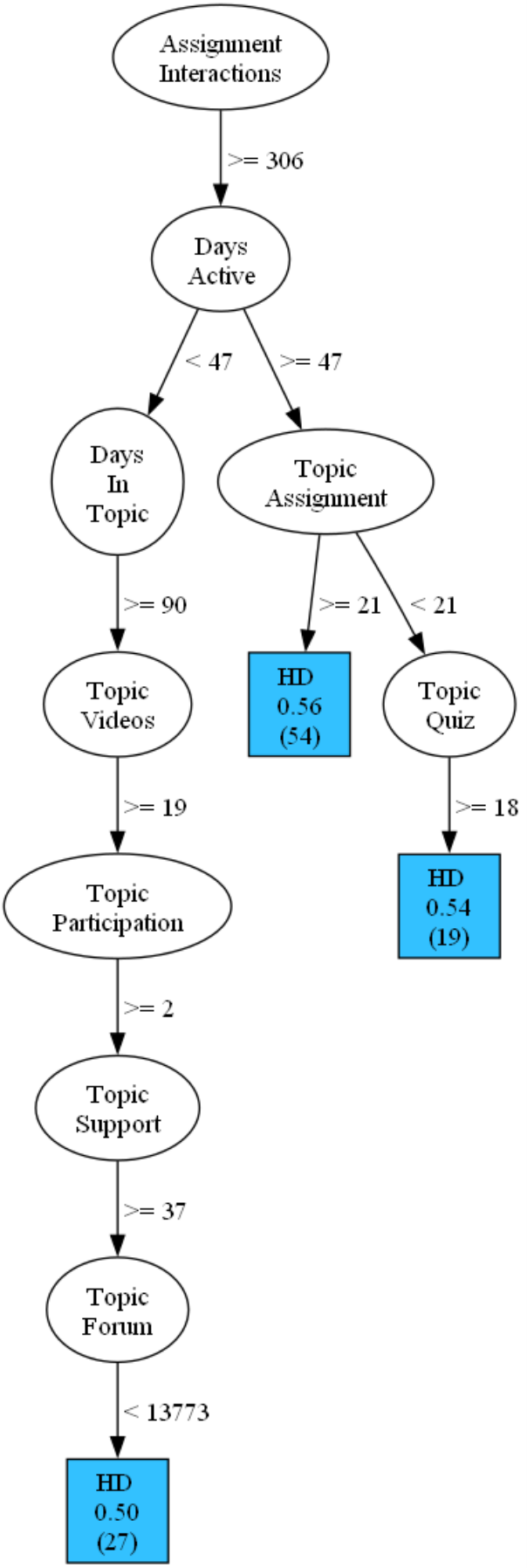

Figure 57 - EPS path for high distinctions (reduced to 7 levels)

5.4.3. Humanities, Arts, and Social Sciences

The HAS decision tree path as shown in Figure 58, primarily focuses on topic construction attributes, particularly Topic Videos, Topic Participation, and Topic Forum. Additionally, non-topic composition such as Days in Topic, Distinct Support Materials, and Support Material Interactions were shown to contribute to the HAS path.

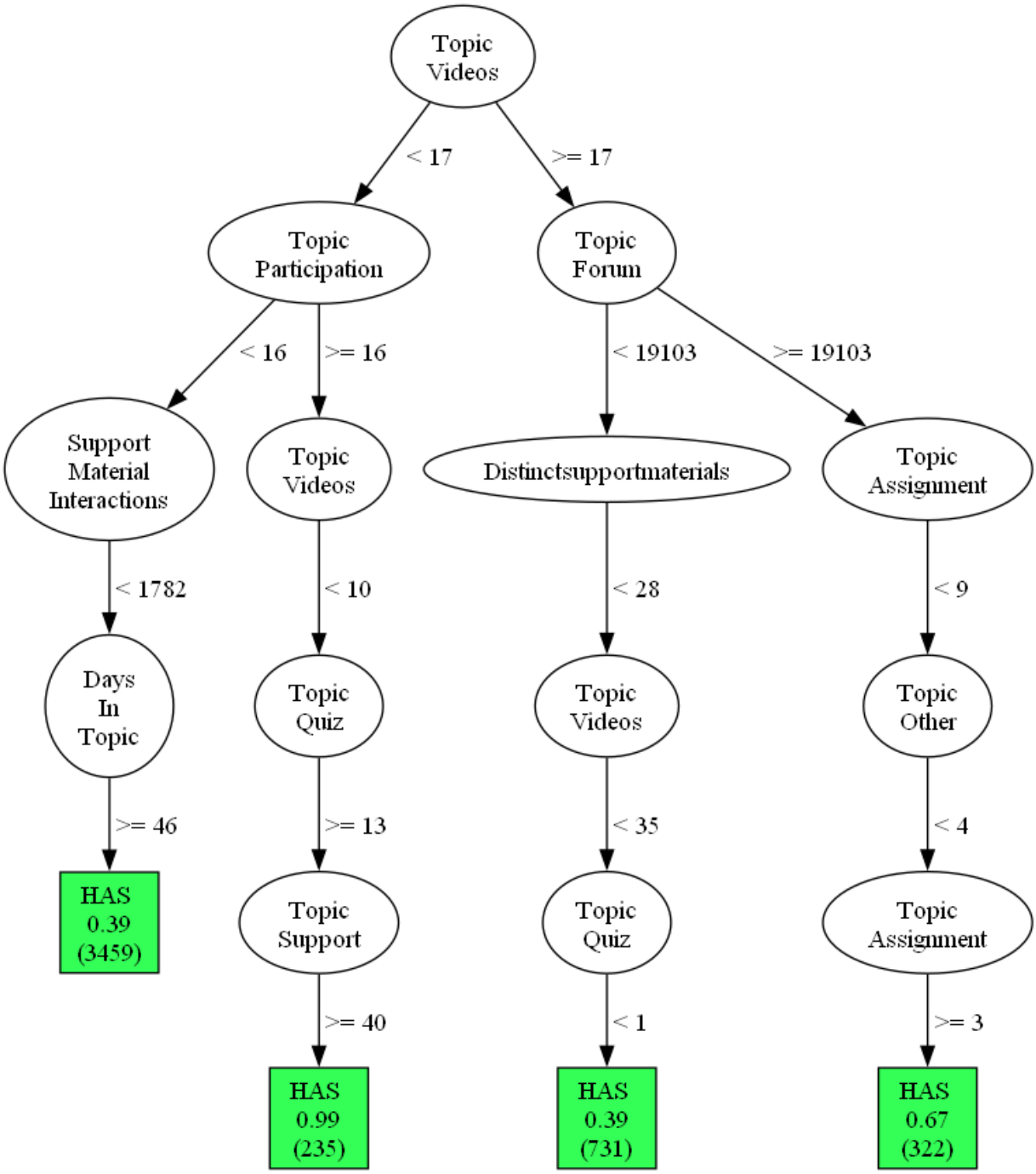

Figure 58 - Path for HAS college classification (reduced to depth of 5)

Accuracy is mixed, with extremely high accuracy for some paths (99%), while others being as low as 39%. However, compared to chance, this is still very good, and would suggest topic composition is relatively reliable in classifying HAS affiliation.

For the prediction of at-risk students (Figure 59), attributes such as total Assignment Interactions, and assignment interactions in later periods of the semester (Assignment Sem Period 4 and Assignment Sem Period 5), appear heavily, as well as lower Days Active, and Days in Topic. Additionally, topic composition attributes involving the size of the forum also appear, suggesting medium to large number of forums posts are a predictor.

Accuracy of the Fail path is mixed, however, has a highly accurate path (74%), in addition to the low (22%), and moderate (36%) paths.

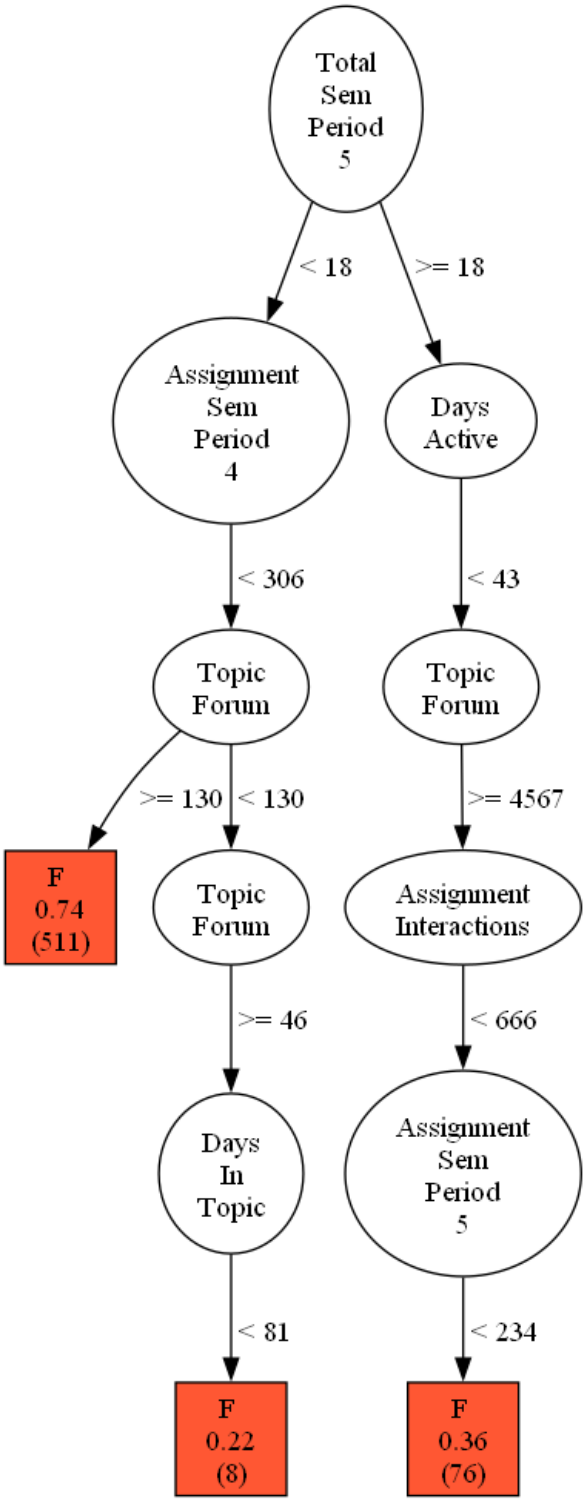

Figure 59 - HAS path for fail grades (reduced to 5 levels)

Classifying high distinction students (Figure 60), utilise similar attributes, but in different contexts. For the topic composition, having a smaller forum size for the topic appears to be predictive of having HD students, while for non-topic related attributes, having larger Days Active and more use of support materials, and social objects early in the morning appear to be predictive of HD students.

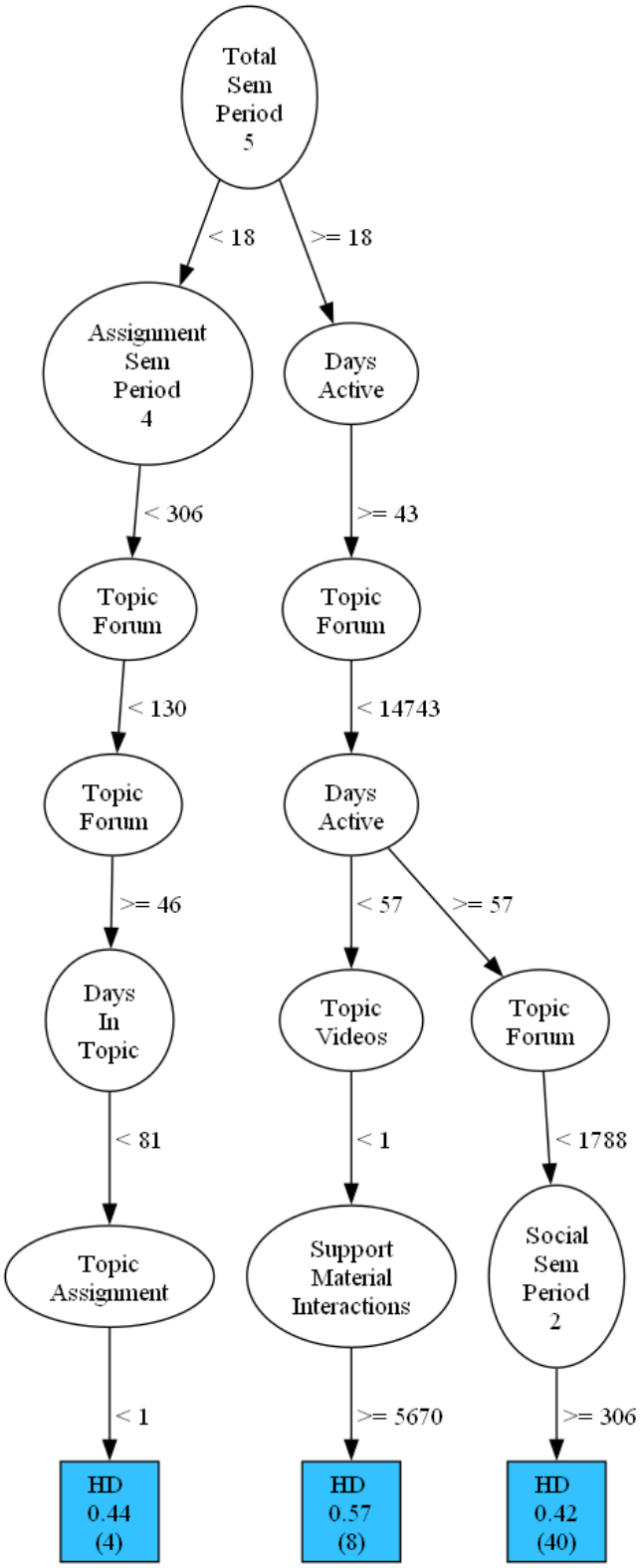

Figure 60 - HAS path for high distinctions (reduced to 7 levels)

Path accuracy is moderate for predicting HDs for HAS, with all being significantly more than chance, and as high as 57%.

For HAS, both engagement with the LMS as well as assignment and social interactions appear to be highly related with positive and negative outcomes. As described in Section 3.4 Exploratory data analysis, HAS was found to have less Fail grades than expected, but more students with ‘Low’ levels of activity, and fewer students with ‘Medium’ and ‘High’ levels of activity. EPS was also found to have more participation and ‘other’ activities; however, these are not specifically mentioned in the decision trees.

5.4.4. Medicine, and Public Health

The MPH path as shown in Figure 61, has a focus on topic composition attributes such as Topic Videos, Topic Participation, and Topic Forum. Suggesting larger forums, and larger number of Topic Participation and Topic Support. Additionally, non-topic attributes such as Distinct support and quiz objects suggest students who interact with large percentages of topic content.

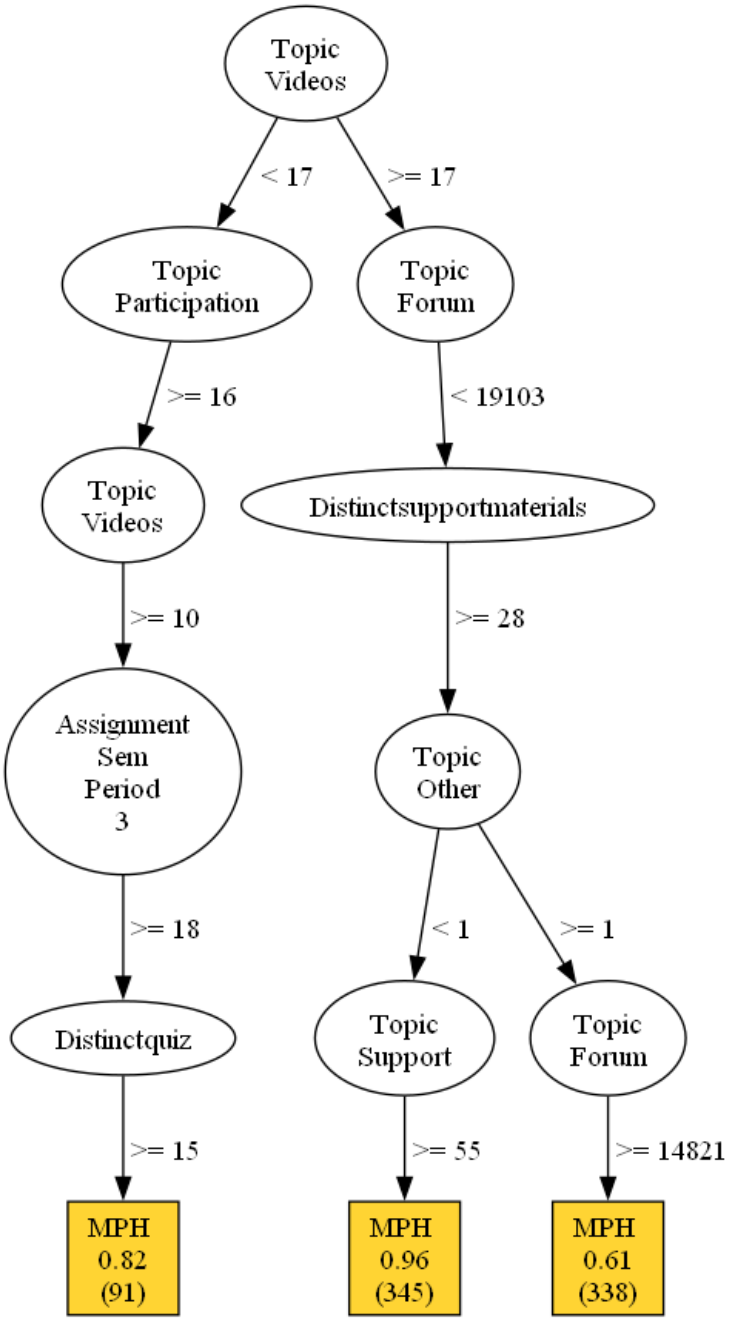

Figure 61 - Path for MPH college classification (reduced to depth of 5)

Path accuracy for MPH is relatively good, with the lowest being 61%, and the highest being 96%, suggesting the topic composition is relatively reliable in predicting MPH affiliation.

For MPH (Figure 62), predicting at-risk students is relatively straightforward compared to other colleges. A combination of specific topic components such as Topic Support, Topic Quiz, and Topic Videos, along with a lack of engagement (fewer Days Active), and a reduced number of interactions during the exam period, appear to be indicative of potential failure.

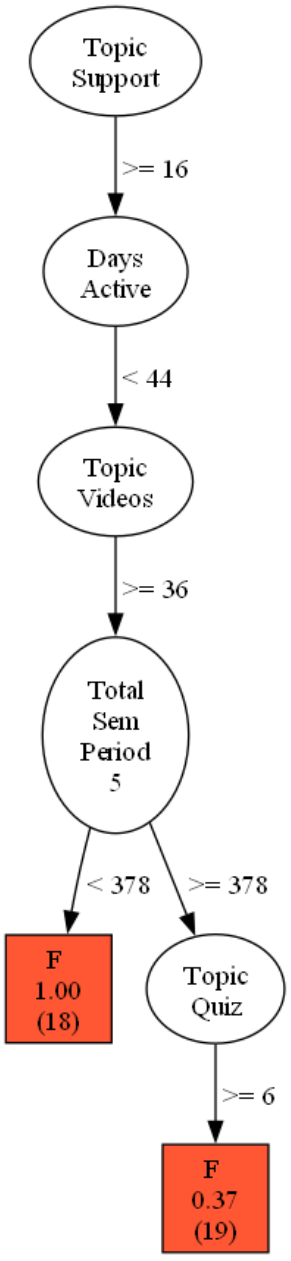

Figure 62 - MPH path for fail grades (reduced to 5 levels)

Accuracy is also very good, with 100% accuracy on one path, and only reducing to 37% for students with high number of interactions during the exam period.

Predicting high distinctions is also relatively simple (Figure 63). With more than 44 Days Active, a large number of interactions during the early semester (Total Period 2), and a large number of activity interactions during the weeks before the exam (Activity Sem Period 4) appearing to predict HDs. This single path may suggest that there is overfitting occurring for HDs for the MPH College.

Accuracy is well above chance, at 49%, suggesting that this path is moderately accurate in identifying HD students for MPH.

For topic composition, larger numbers of support materials, lower numbers of videos, are suggestive of predicting high distinctions. As described in Section 3.4 Exploratory data analysis, MPH showed less Fail grades than expected, and more HD grades than expected. Along with more students in the 'Medium' and 'High' activity categories within the LMS and fewer students with ‘Low’ activity levels. Additionally, MPH along with NHS were shown to have more quizzes than other colleges, however in the case of MPH, larger numbers of quizzes were shown to be related to the fail path as shown in Figure 62.

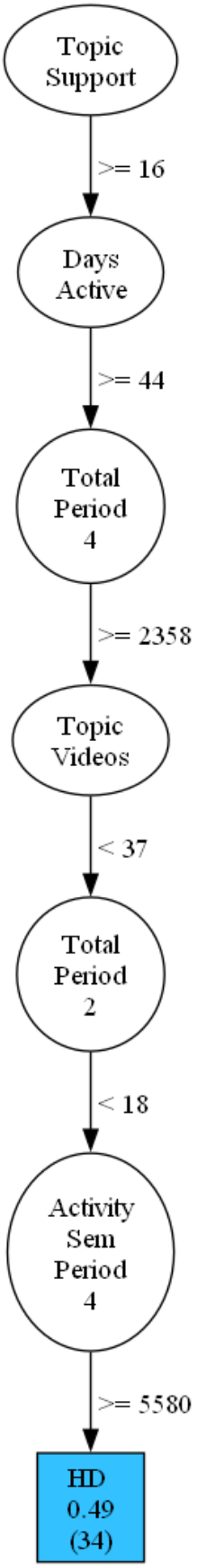

Figure 63 - MPH path for high distinctions (reduced to 7 levels)

5.4.5. Nursing, and Health Sciences

The NHS decision tree path as shown in Figure 64, primarily emphasises topic composition, with Support Material Interactions (exceeding 1,782 interactions) being the sole student interaction attribute represented. A larger number of forum posts and a lower number of videos, along with a higher number of quizzes, are also seen as predictors for NHS outcomes.

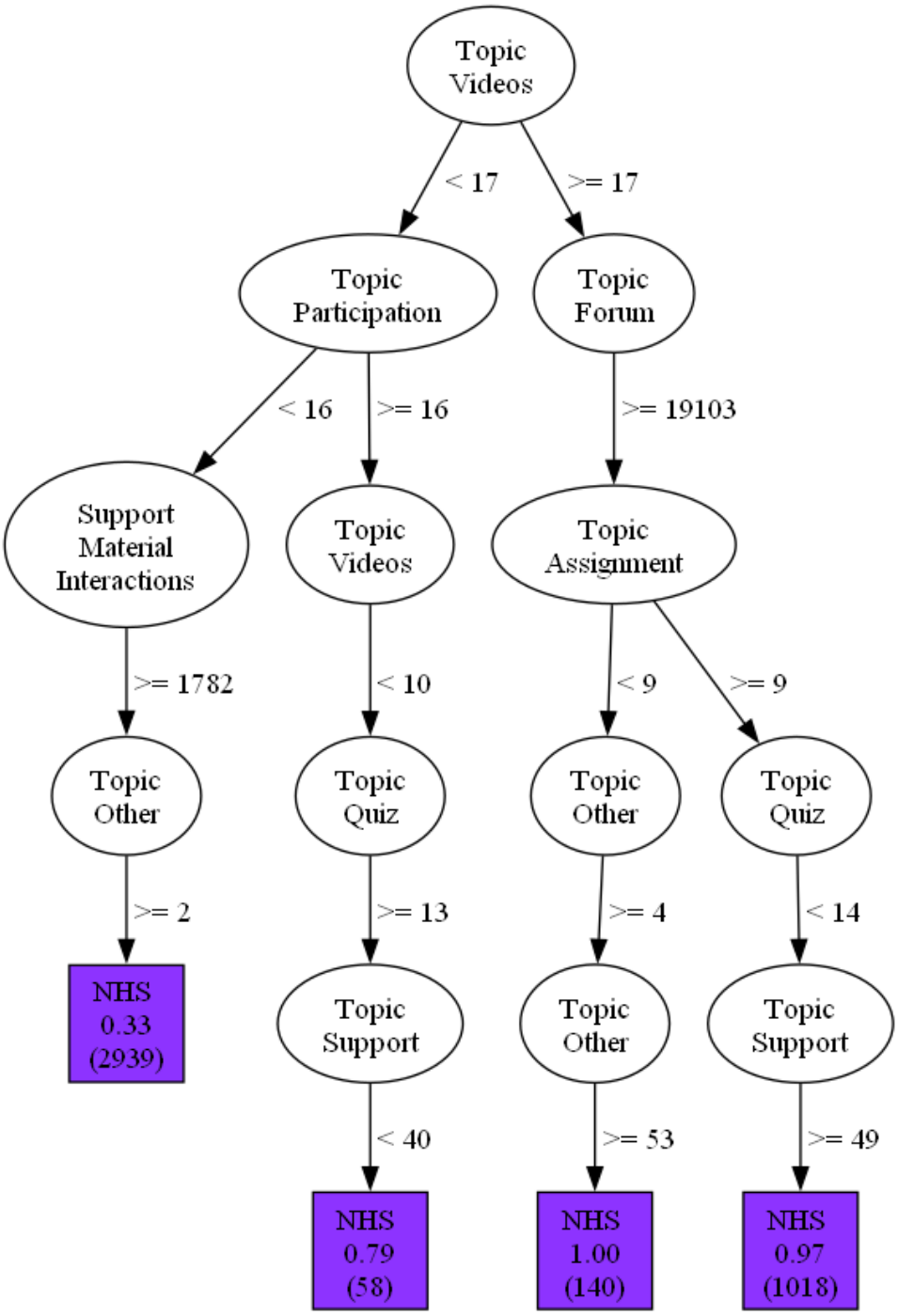

Figure 64 - Path for NHS college classification (reduced to depth of 5)

Most paths to predict NHS affiliation are very highly accurate, with generally above 79% accuracy. With the remaining path is still double the chance of randomly predicting a college. This suggests that topic composition for NHS is very important for predictive purposes.

Predicting at-risk students for NHS, as shown in Figure 65, is relatively straightforward. Factors such as fewer Days in Topic, reduced interactions in the weeks leading up to the exams, and diminished engagement with assignments from the middle of the semester up to just before the exams (Assignment Sem Period 4 and Assignment Sem Period 5) all suggest poor outcomes.

Again, for predicting Fails, the paths shown are very high (91%) to moderately high (52%), suggesting that Fails are relatively straight forward, and reliable to predict from the shown LMS attributes.

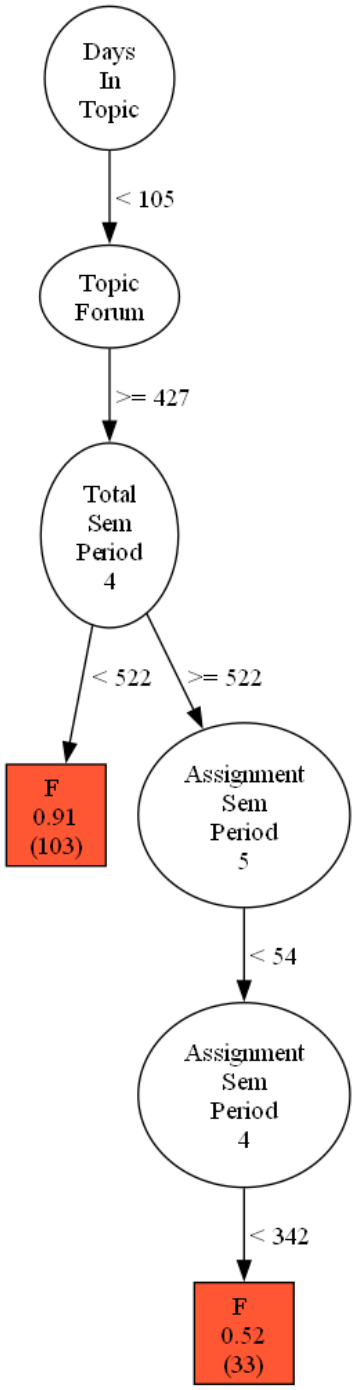

Figure 65 - NHS path for fail grades (reduced to 5 levels)

For NHS, predicting students who achieve high distinctions, as shown in Figure 66, involves indicators such as a large number of Days in Topic, a high number of social interactions during the middle of the semester (Social Period 4), and a large number of unique quizzes viewed. In terms of topic composition, this includes a larger number of 'Other' materials (Topic Other), a lower number of forum posts, fewer participation objects (Topic Participation), and fewer videos (Topic Videos).

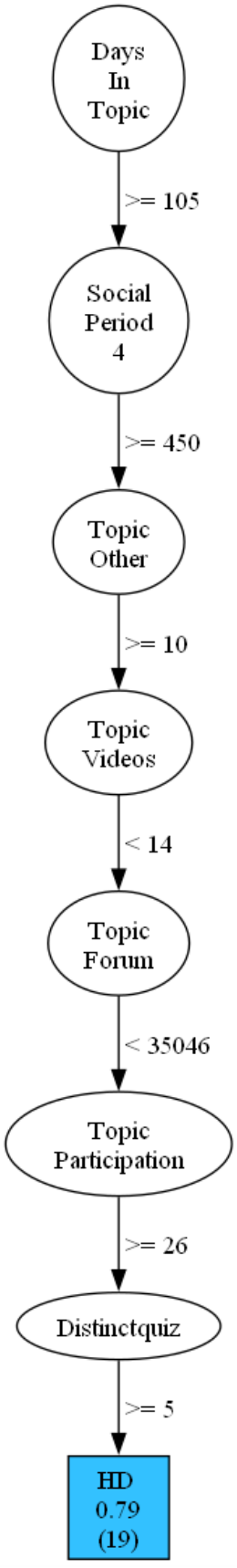

Figure 66 - NHS path for high distinctions (reduced to 7 levels)

The cognitive load theory discussed by Sweller, van Merriënboer & Paas (2019) supports the observation that lower-performing students may struggle to prioritise essential materials, which could lead to extensive but less effective interactions with the LMS. In addition, Rasheed & Wahid (2021) found that learners facing challenges may over rely on available resources, reflecting this behaviours observed by lower-performing students.

In the same way as predicting a Fail is for NHS is very reliable, predicting a HD is also very straight forward and reliable, at 79% accuracy, and only one path, suggesting that each of the variables shown have significant importance. However, as mentioned previously for other colleges, this tree may suggest there is overfitting occurring, for HDs for the NHS College.

Success in NHS would appear to be related to both the level of activity, and with the utilisation of quizzes and assignments. As described in Section 3.4 Exploratory data analysis, NHS showed higher Fail grades than expected, and fewer HD grades than expected. They also, much like MPH, showed a larger number of 'High' activity students, and fewer 'Low', and 'Medium' activity students. NHS also had more quizzes than most colleges (like MPH), but as shown in Figure 66, the tree suggests that students viewing more distinct quizzes, is related to a HD outcome.

5.4.6. Science and Engineering

Among all the decision tree paths, the one for S&E, as shown in Figure 67, exhibits the most complexity. It suggests a significant emphasis on topic composition attributes, including Topic Videos, Topic Participation, and Topic Forum. Additionally, non-topic composition attributes like fewer Days in Topic and a greater number of interactions with 'Other' type objects are predictive of S&E outcomes.

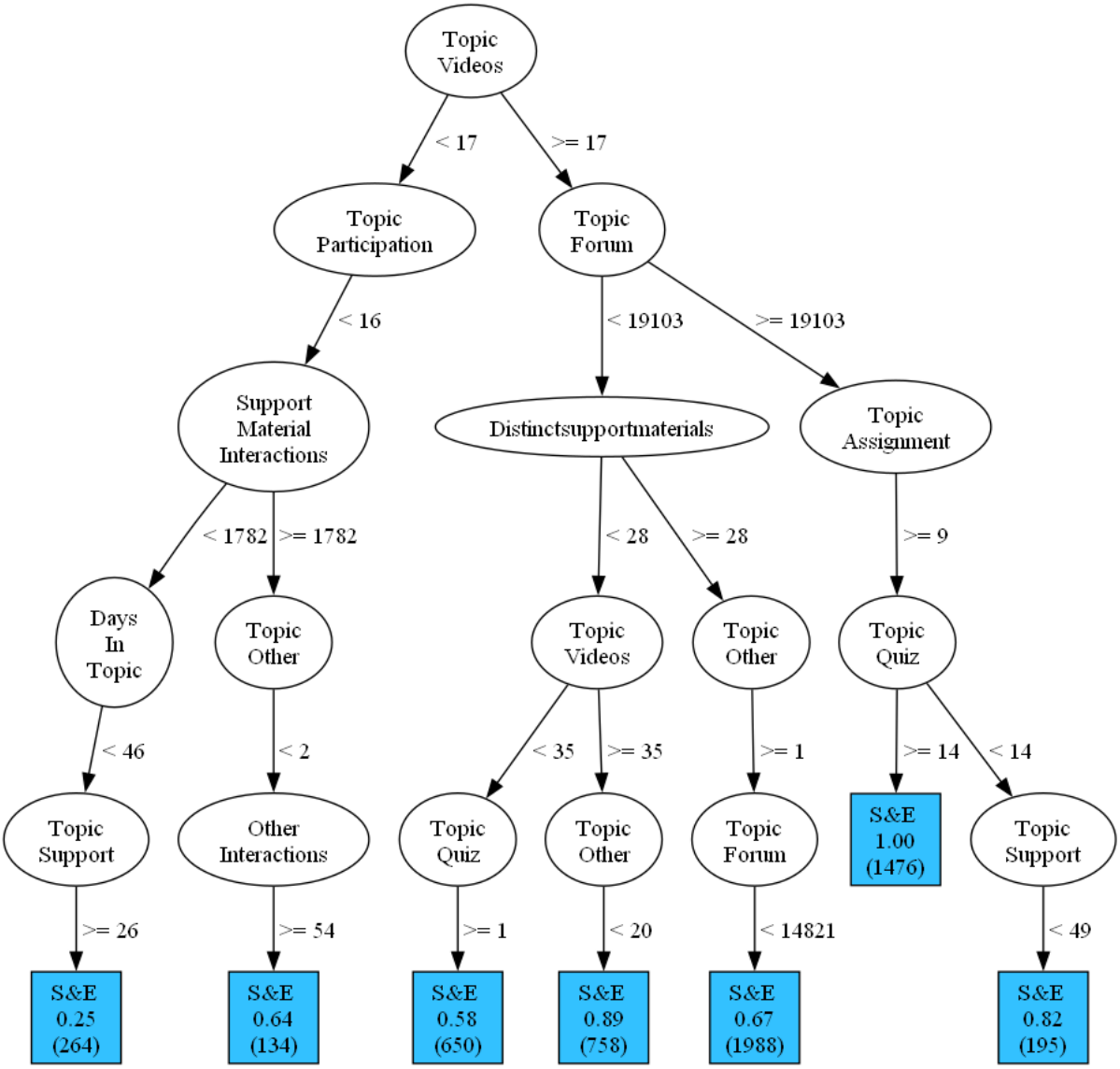

Figure 67 - Path for S&E college classification (reduced to depth of 5)

However, regardless of the complexity, the accuracy of the paths are all relatively high for predicting S&E affiliation, with only one path dropping to 25%, and the rest being above 58% and reaching 100% in one instance.

While the prediction of at-risk students for S&E (Figure 68), is more complicated than other colleges, it does appear to have more student-centric attributes, at higher levels of the decision tree. These include attributes such as, Days Active, Days in Topic, Distinct Quiz, and Participation Sem Period 5.

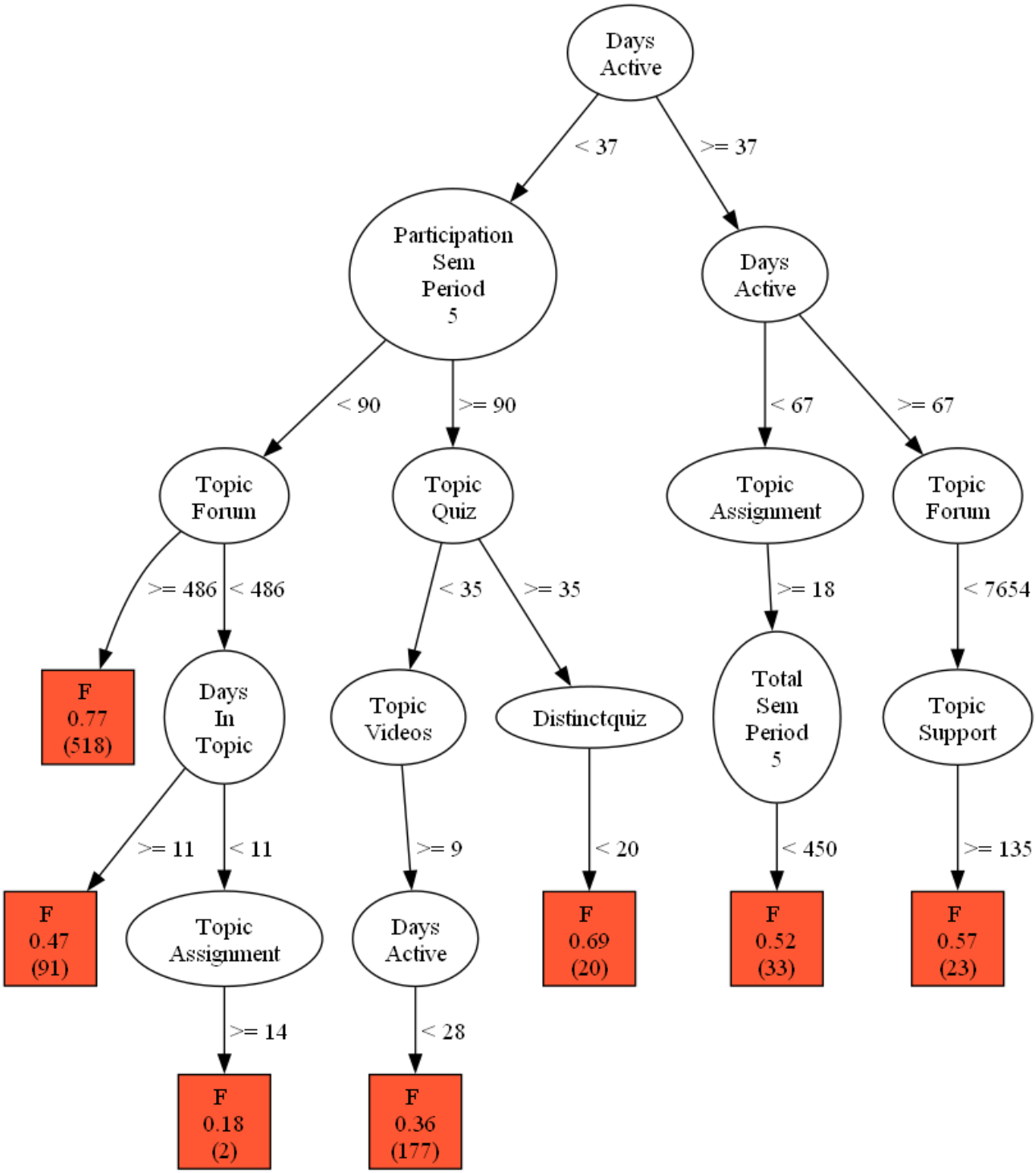

Figure 68 - S&E path for fail grades (reduced to 5 levels)

This path also begins with Days Active as a critical attribute. Students who demonstrate higher activity (more than 37 Days Active) are further evaluated based on their engagement with participation materials during the exam period (Participation Sem Period 5). Additionally, lower overall engagement during the exam period is a predictor for students with medium to low days of activity. Topic composition-related attributes such as Topic Quizzes, Topic Videos, Topic Assignments, and Topic Forum posts are also indicative of students at risk of failing.

Accuracy for predicting at-risk students is mostly medium to high, with one being worse than random chance (18% compared to 20% randomly guessing grades), again suggesting the complexity for the college with regards to prediction.

Finally, the path for predicting the high distinction path for S&E Figure 69, is one of the most complicated of all the colleges. Again, Days Active playing a major role in classification. In addition, Days in Topic, interactions with assignments, support materials, and social materials before the exam period (Assignment Sem Period 4, Support Sem Period 4, and Social Period 4), all play important roles.

The accuracy of the paths involved are mostly low to medium, again, underscoring the complexity and difficulties in prediction for S&E.

S&E often provided more extensive content in areas such as lecture videos and quizzes, with the number of video components in S&E topics notably higher than in other colleges. As described in Section 3.4 Exploratory data analysis, S&E was shown to have less Fail grades than expected, and more HD grades than expected. Activity levels were also increased, with fewer students in ‘Low’ activity, and more in ‘High’ and ‘Medium’ levels of activity.

S&E was also shown to have larger numbers of videos, quizzes, assignments, and support materials. While higher forum activity than most, but less than BGL and EPS. Due to the complicated nature of the trees shown above, it is not as easy to suggest what component is more important, however, general interaction with the LMS would be suggested, along with interactions during the later period of the semester.

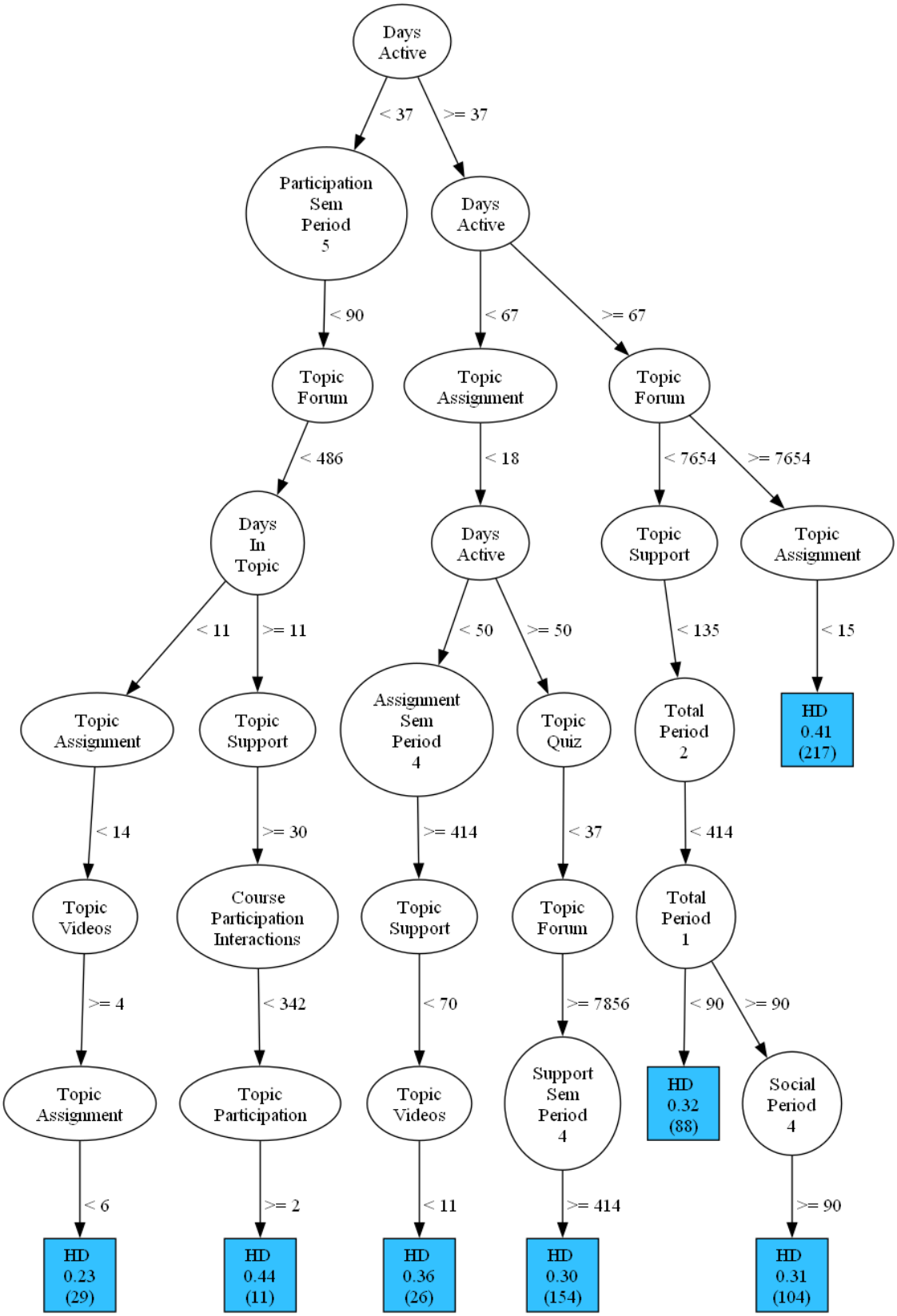

Figure 69 - S&E path for high distinctions (reduced to 7 levels)

5.4.7. Summary of college differences

This section synthesises the distinctions among the colleges based on the decision tree analyses presented, focusing on topic composition, student engagement, and predictors of academic outcomes.

College of Business, Government, and Law (BGL)

Shows a preference for Topic Videos and Topic Participation as positive indicators, while low Days in Topic and Support Material Interactions suggest limited LMS engagement. Early semester assignment interactions are key for identifying at-risk students. High distinctions correlate with extensive engagement with videos in the mid semester and exam periods.

Education, Psychology, and Social Work (EPS)

Attributes such as Topic Videos and Topic Participation are generally favourable indicators for EPS. Lower values for Days in Topic and Support Material Interactions suggest low engagement with the LMS. A significant number of early semester interactions with assignments is indicative of at-risk students, while high distinctions correlate with substantial Topic Assignments and Topic Quizzes. EPS highlights the importance of early semester engagement and a balanced mix of assignments and quizzes for academic success.

Humanities, Arts, and Social Sciences (HAS)

Points to the importance of Topic Videos and non-topic attributes like Days in Topic and Distinct Support Materials in classification. A medium to large number of forum posts is predictive of at-risk students, while smaller forums and increased engagement with support materials are linked to high distinctions. HAS suggests that a nuanced approach to forum size and support material engagement can influence outcomes.

Medicine, and Public Health (MPH)

Classification leans on Topic Videos, Topic Participation, and Topic Forum. Higher interaction with quizzes and distinct support materials indicates engaged learning. Fewer Days Active and reduced exam period engagement signal potential failure, whereas lower Days Active combined with higher early semester interactions predict high distinctions. MPH emphasises the value of early and diverse interactions with course materials for academic excellence.

Nursing, and Health Sciences (NHS)

Emphasises topic composition, with Support Material Interactions being a standout attribute for student engagement. A larger number of forum posts and quizzes, along with a lower number of videos, predict NHS outcomes. Key indicators for high distinctions include extensive Days in Topic, mid-semester social interactions, and engagement with quizzes. NHS reveals a pattern where diverse material engagement and timely participation correlate with success.

Science and Engineering (S&E)

Demonstrates a complex interplay of Topic Videos, Topic Participation, and Topic Forum, alongside non-topic attributes like Days in Topic. S&E stands out for its extensive content across lectures, quizzes, assignments, and support materials, with higher forum activity compared to most colleges except BGL and EPS. Engaging with the LMS, especially during later semester periods, is crucial for achieving high distinctions in S&E.

Summary

While there are common threads such as the importance of Topic Videos and participation across colleges, each college has unique predictors of student success and risk factors. BGL and EPS highlight the critical timing of engagement, HAS and NHS underscore the role of support materials and forum interactions, and MPH and S&E emphasise the breadth of engagement with course materials. This nuanced understanding aids in tailoring interventions and support mechanisms to enhance student outcomes across different academic disciplines.

5.5. Overall university perspective

The research suggests significant variability in grade distribution, student engagement with the LMS, and interaction patterns across colleges as shown in the results in 3.4 Exploratory data analysis. In addition, results from Section 4.5 Experiment 2 results suggest each are significantly different in topic composition to be able to reliably predict college membership from LMS data on topic composition, and usage patterns. The differences in how colleges utilise LMS components point to varied pedagogical strategies and resource allocations, with the College of Science and Engineering standing out for its extensive provision of content.

For a deeper analysis on attribute usage than can be shown in a reduced depth decision tree, the following sections will outline the per-college full decision usages of attributes, grouped by attribute type. This will identify attributes used to classify both high performers (HD), and failing students (F), and which colleges utilise (or do not utilise at all) certain attributes, represented by a numeral (for utilises attribute) or a blank cell (for does not utilise). Each row represents a different attribute (grouped by type for each table), while columns represent college-outcome combinations for each attribute (if the attribute is used in the context of an F or HD outcome), and if the attribute was used in a positive or negative context (represented by green for a positive interaction, and red for negative interaction). Note, that attributes can have both positive and negative effects on outcomes. This analysis does not consider decision point values, only the usage of said attributes, to show the usefulness in classification scenarios.

Enrolment

The overall usage of enrolment attributes as shown in Table 46, suggests each attribute is used in some capacity for each college. With some having more usage across both trees, such as S&E (but in a more mixed fashion). Days Active is universally utilised across all colleges to classify both failing and high-performing students. The usage context (positive or negative) varies, indicating its significance in predicting student outcomes. For instance, BGL, EPS, HAS, and S&E demonstrate a broader application of this attribute in both positive and negative contexts for both HD and F classifications, suggesting a nuanced understanding of student engagement over time.

Table 46 - Enrolment attributes comparison (by college and F/HD)

College	BGL				EPS				HAS				MPH				NHS				S&E			
Grade	F		HD		F		HD		F		HD		F		HD		F		HD		F		HD	
Outcome																								
Effect on Outcome	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Days Active	1	3		2		3	2	1		2	4	3		1	1				1		3	4	3	3
Days In Topic	2	2			2	2	1			1		1	1	1			1	1	2		2	2	1	1

Days in Topic, like Days Active, is widely used, however, with a more balanced representation between positive and negative contexts.

Bulk interactions

As shown in Table 47, Total Interactions across various components is shown to be more varied. Activity Interactions sees minimal use, only positively for failing students in MPH, indicating limited application, whereas Assignment Interactions is predominantly used to differentiate outcomes in EPS, HAS, and MPH, with a mixed context of positive and negative implications. Course Participation Interactions is only utilised in S&E, in both positive context and negative, suggesting a unique emphasis on participation in this college.

Table 47 - Total interaction attributes comparison (by college and F/HD)

College	BGL				EPS				HAS				MPH				NHS				S&E			
Grade Outcome	F		HD		F		HD		F		HD		F		HD		F		HD		F		HD	
Effect on Outcome	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Activity Interactions														1										
Assignment Interactions					1	1	2			3		1												
Course Participation Interactions																					1	1		1
Forum Activity Interactions														1								1		
Support Material Interactions						1				1														
Avg. Sec Between actions						1												1			1	1	1	1

For Forum Activity, this shows limited use across colleges, with negative use for failing students in HAS and S&E, indicating specific contexts where forum engagement is preventative of failure. Interactions is used by EPS and HAS in a negative context (F), suggesting for both colleges, a consideration of interaction quality or quantity in predicting failures. Support Material Interactions, again with EPS and HAS, is shown in a negative context (F) for EPS but with a positive context for HD students in HAS, suggesting variations in how support material engagement correlates with student outcomes. Avg. Sec Between Actions is almost exclusively used in S&E (with only NHS utilising it for a positive context for HD), with both in positive and negative contexts for HD and F students.

Distinct interactions

The usage of distinct components Table 48, sees limited use across colleges, primarily being focused on certain components for each college. For example, both BGL and MPH utilise Distinct Assign, but BGL uses it as a positive predictor for HD students, and MPH uses it as a negative predictor for Fail students. Suggesting the attribute is positive, but just in different contexts, preventing failure, or ensuring high distinctions. Distinct Lecture videos is solely utilised by S&E, and as a negative for Fail students, while Distinct Quizzes is also utilised by S&E, but as a negative predictor for both Fail and HD students. Finally, Distinct Support Materials is utilised by HAS, MPH, and NHS, however for completely different reasons. HAS utilise it as a negative for HD students, MPH utilises it as a negative for F students, and NHS utilise it as a positive for HD students.

Table 48 - Distinct interaction attributes comparison (by college and F/HD)

College	**BGL**				**EPS**				**HAS**				**MPH**				**NHS**				**S&E**			
Grade Outcome	**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**	
Effect on Outcome	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**
Distinct Assign			1											1										
Distinct Lecture																							1	
Distinct Quiz																					2		2	
Distinct Support Materials											1			1					1					

Time of day engagement

Interactions across the time-of-day attributes is again more varied between colleges, as shown in Table 49. Assignment across time periods use is varied usage across colleges, with specific time periods being significant in predicting outcomes. For example, Assignment Period 6 is negatively associated with F students in BGL, while Assignment Period 7 has a positive association for HD students in S&E. Participation use varies across all time periods. Notable periods include Participation Period 5, which is positively linked to HD students in several colleges (EPS, NHS, S&E), and Participation Periods 1 and 8, each negatively associated with F students in BGL.

Table 49 - Time of day attributes comparison (by college and F/HD)

College	**BGL**				**EPS**				**HAS**				**MPH**				**NHS**				**S&E**			
Grade Outcome	**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**		**F**		**HD**	
Effect on Outcome	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**
Assignment Period 2																			1					
Assignment Period 4																								1
Assignment Period 6		1																						
Assignment Period 7																							1	1
Participation Period 8	1																							
Participation Period 1	1																							
Participation Period 2													1						1					
Participation Period 4	1																							
Participation Period 5				1																	1	1		1
Social Period 4													1				1	1	1				1	1
Social Period 5										1														
Social Period 6				1			1																	
Support Period 2																								1
Total Period 1			1																				1	1
Total Period 2															1						1		1	2
Total Period 3		1																						
Total Period 4													1	1										
Total Period 5																					1			
Total Period 8										1													1	

Social activity across time periods varies across colleges. Social Period 4 is noteworthy, positively associated with HD outcomes in NHS and S&E, indicating that engagement in social activities during this period may contribute to higher performance. Support Period 2 sees limited use, being positively influencing HD outcomes in S&E, suggesting that using support at this period correlates with higher achievement. Total Periods Represent overall engagement across each period, Total Period 2 shows a positive association with HD students in MPH and S&E, and Total Period 1 is linked positively with HD students in BGL and negatively in S&E, highlighting the importance of earlier engagement in the day.

Time of semester engagement

For the time of the semester attributes, shown in Table 50, each attribute highlights a different aspect of activity for each college. Activity Sem Periods show varied significance across the colleges, with notable emphasis on later periods (e.g., Activity Sem Period 4 and 5) for both F and HD outcomes in some colleges, suggesting critical times for engaging students to influence their academic success.

Table 50 - Time of semester attributes comparison (by college and F/HD)

College	BGL				EPS				HAS				MPH				NHS				S&E			
Grade Outcome	F		HD		F		HD		F		HD		F		HD		F		HD		F		HD	
Effect on Outcome	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Activity Sem Period 1																			1				1	
Activity Sem Period 2													1											
Activity Sem Period 3																							1	
Activity Sem Period 4					1	1									1	1					1			
Activity Sem Period 5																	1	1			1	1	1	
Assignment Sem Period 1							1	1																
Assignment Sem Period 2															1									
Assignment Sem Period 3							1																	
Assignment Sem Period 4		1				1				1		1						1				1	1	
Assignment Sem Period 5						1	1	1		1				1				1					1	
Participation Sem Period 5	1	1				1																		1
Social Sem Period 2											1						1							
Social Sem Period 4																								
Support Sem Period 1						1																		
Support Sem Period 2	1																						1	
Support Sem Period 3														1								1		
Support Sem Period 4											2												2	
Total Sem Period 1																								1
Total Sem Period 2					1	1																		
Total Sem Period 4	2	1			1	2			1	1				1			1	2	2		2	3	2	1
Total Sem Period 5						1			1	1	1	2	1	1								3		
Video Sem Period 3			1																			1		
Video Sem Period 4			1																1					

Assignment Sem Periods suggest the importance of assignment engagement throughout the semester, with Assignment Sem Period 4 and 5 being particularly pivotal for several colleges. These periods are associated with both positive and negative outcomes, indicating the critical timing of assignments in determining student success. Participation Sem Period 5 shows a mixed contexts with F outcomes in BGL, and a negative context for F students in EPS.

Social Sem Periods have limited but specific impact, with Social Sem Period 2 positively affecting HD outcomes in HAS, but also positively affecting F outcomes for NHS, indicating that early-semester social engagement is positive for HAS, but may be negative for NHS students. Support Sem Periods indicate that seeking support at various times can have a significant impact, with Support Sem Period 4 being notably associated with HD outcomes in HAS, and S&E highlighting the value of late-semester support.

Total Sem Periods reflect overall engagement across the semester, with Total Sem Period 4 showing extensive associations with both F and HD outcomes across several colleges. This suggests a broad understanding of how engagement intensity and timing affect academic performance. Finally, Video Sem Periods have limited usage but indicate that engagement with video content during the mid-semester (Periods 3 and 4) can influence HD outcomes, particularly in BGL NHS, and S&E, underscoring the importance of multimedia resources in student learning.

Topic composition

Topic composition was universally used in all the college affiliation predictions to a certain degree and varied in type across colleges. As shown in Table 51, total number of Topic Assignments has varied importance across colleges, with a notable emphasis in S&E for a negative context for HD outcomes, but a positive context for F outcomes.

Table 51 - Topic composition attributes comparison (by college and F/HD)

College	BGL				EPS				HAS				MPH				NHS				S&E			
Grade	F		HD		F		HD		F		HD		F		HD		F		HD		F		HD	
Outcome																								
Effect on Outcome	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Topic Assignment	1				2	1	2	2	1			1				1				1	4	3		4
Topic Forum	4	4		1	3	2	2	3	4	2	5	4	2				3		2		1	3	3	3
Topic Other	1				2	2	1	2				1							1	2	1			
Topic Participation	1		1		1	2	1	1														1	1	
Topic Quiz	1				1		1						1							1	1	3		1
Topic Support	2	2			1	1	3	1	1	1			1		1						4	2	1	2
Topic Videos	1				2	2	2	1	1			1	1	1	1	1					3	1	2	2

The Topic Forum posts attribute is widely used across all colleges with a strong correlation to both F and HD outcomes, indicating the forum's central role in student engagement and academic performance. The number of forum posts is a significant predictor in colleges like BGL and S&E and HAS.

Topic Other is less frequently mentioned but indicates that engagement with other types of topic-related content can influence outcomes, especially in. Topic Participation generally has a positive impact on HD outcomes in colleges like BGL, EPS, and S&E. Topic Quiz is similar to Topic Assignment, with quizzes notably important in S&E for F students, but negative for BGL, MPH, and NHS. Topic Support indicates a broad impact across colleges, particularly in EPS and S&E, where total support materials are positive for EPS HD students, and negative for S&E F students. Finally, Topic Videos is utilised across most colleges, especially in S&E for HD and F outcomes, indicating that video content is a significant component of engaging and effective learning experiences.

5.6. Chapter summary

Chapter 5 investigates the relationship between Learning Management System (LMS) usage and academic outcomes across different academic disciplines, drawing on comprehensive data analyses and experiments. This chapter is structured to first identify common LMS features and engagement patterns across colleges, then moves into detailed examinations of each of the six colleges' unique interactions with the LMS, concluding with overarching insights for educational strategies and LMS design.

The chapter opens by addressing the universal aspects of LMS engagement, highlighting the shared attributes that influence student performance across all colleges. This sets the stage for a deeper, college-specific analysis, where distinct or significant LMS features influencing student outcomes are brought to the forefront. Each college's section reveals how specific attributes, from topic composition to engagement timing, uniquely contribute to defining academic success or risk within their disciplines.

A significant portion of the discussion is dedicated to the ways in which colleges differ in their LMS usage. For example, the College of Business, Government, and Law shows a pronounced reliance on video content and participation metrics, while the College of Science and Engineering emphasise a complex interplay of various content types and engagement patterns. These insights underline the importance of discipline specific LMS strategies to foster academic achievement.

Overall, this chapter argues for a customised approach to both LMS design and instructional design. This is due to the presented needs and engagement patterns between colleges. It suggests that while there are common threads in LMS activity and utilisation, such as the consistent importance of videos and active participation with the LMS, the context of these interactions, such as timing (time of day, and time of semester) and the type of content interacted with (videos, activities, quizzes, forum posts, etc.), varies significantly across colleges.

The chapter concludes by suggesting the importance of, and the implications of these findings for both universities and educators. Highlighting the potential of utilising LMS data to enhance pedagogical practices and student outcomes, by personalising learning experiences to accommodate the distinct requirements of each academic discipline/college. This analysis not only highlights the complexity of LMS engagement across colleges but identifies areas for future research and development in educational technology and instructional design.

6. Conclusions

6.1. Chapter overview

This chapter discusses the implications of the findings from this research regarding LMS use and its impact on student performance across various disciplines. Providing conclusions following the analysis of results from the exploratory data analysis, principal component analysis, and machine learning experiments. These conclusions will detail how LMS engagement correlates with academic outcomes and explain the importance of tailored pedagogical approaches across disciplines, as well as the need for LMS designers to target those unique requirements to better support students and educators and improve the likelihood of successful academic outcomes. While the conclusions presented relate to Flinders University primarily, insights gained from this research can be used for future research into other institutions, as well as being a useful guide for educators regarding customisable educational content, from a discipline-based perspective.

The chapter will address how the research has answered the outlined research questions (as described in Section 1.2.1), utilising results from experiments detailed in Section 4, to describe the relationship between LMS usage and student academic performance. It will outline specific LMS features and student engagement patterns that are useful as predictors for overall student success as well as showing differences in LMS engagement across colleges. This detailed analysis will show the unique disciplinary differences that influence student LMS engagement and, consequently, student learning outcomes.

This chapter details the unique contribution of the research, emphasising discipline-specific insights into LMS usage, and advocates for personalised engagement and LMS material design. Specific recommendations for each college are outlined, informed by LMS usage patterns and pedagogical needs. Additionally, it proposes strategies to optimise LMS components based on discipline specific criteria, serving as a guide for educators to refine teaching practices and create more effective LMS implementations.

Finally, this chapter outlines how the research findings can inform and improve instructional design practices, specifically improving the analysis phase of the ADDIE model to tailor educational content and implementations.

6.2. Answers to research questions

The findings presented in Section 4 Results and analysis presented in Section 5 Discussion, provide critical insights into the research questions posed in this thesis. Each experiment was meticulously crafted to address these questions, with some directly providing answers, while others complement additional experiments to offer comprehensive insights, ensuring all research questions have been thoroughly addressed.

Results presented in Section 3.4 Exploratory data analysis provide answers, covering research questions RQ1.1, RQ1.2, and RQ1.3. This is shown in the analysis of grade distributions, which identifies that there are discrepancies between the colleges, before even going into any LMS engagement. Additionally, attendance levels, activity over time (and semester), and usage of LMS components have been shown to be different across colleges, providing support for research questions RQ2.1, RQ2.2, and RQ2.3.

Topic composition is another factor that was identified early in the exploratory data analysis phase, with different colleges showing significantly different compositions of topic materials, this feature was key for beginning to answer research question RQ2.3.

Results presented in Section 4.3 Principal Component Analysis (PCA), allow for both a direct answer to research question RQ1.4, as well as additional information on attribute patterns, which are directly linked with research question RQ2.1.

While the initial predictive nature of grades was not as high as was hoped, Section 4.4 Experiment 1 results did identify general factors that are common across colleges, with regards to predicting performance. This was crucial for research questions RQ1.1, and RQ1.2. This does suggest that predictive analytics could be utilised to identify not only student performance but also the engagement patterns, preferences in LMS components, and the variability in topic construction across colleges. Results from Section 4.5 Experiment 2 results, can be utilised to directly answer RQ1.3, as well as provide more support for answering research questions RQ2.1, and RQ2.2.

Finally, the results discussed in Section 4.6 Experiment 3 results: Predictive analytical models for E-Learning by discipline, enable direct answering of research questions RQ1.3, and RQ1.5, as well as research questions RQ2.1, RQ2.2, and RQ2.3. Providing both college specific information about usage, engagement, as well as topic composition, like Experiment 2, but adding in additional results on predicting student outcomes, and the benefit of approaching this process from a college-centric view.

6.2.1. How does LMS use across discipline impact student performance? (RQ1)

How does LMS usage differ across disciplines, and how are these differences associated with student performance metrics? (RQ1.1)

This research has shown that not only does student usage patterns of the LMS differ across college, but it also differs significantly. With results from Section 3.4 Exploratory data analysis showing statistically significant differences across colleges. Further analysis in Section 4.4 Experiment 1 results, and Section 4.5 Experiment 2 results outline predictable differences between colleges, so much so, that the models created in Experiment 2 performed vastly superior with predicting college affiliation, as opposed to Experiment 1 predicting grade outcomes. Finally, considering results from Section 4.6 Experiment 3 results, which were in general higher performing regarding the predictive capability of grades for each algorithm, would suggest that these differences can be utilised to better predict student performance, via college specific interventions, as opposed to a one-size-fits-all approach.

Which specific features of LMS usage are significant predictors of student academic performance? (RQ1.2)

This research has shown that some of the most important factors that can be utilised as both predictors for student performance overall, and in a college-by-college context, are enrolment factors (such as days active in the LMS), and specific utilisation of topic components. Additionally, the composition of the topics themselves can be a predictor of student performance, as discussed in Section 5.4 Discipline-specific analysis, most of the topics had a strong preference for topic composition, for both predicting membership, as well as HD paths, and Fail paths.

This suggests that not only is the engagement with the LMS important, engagement with the right type of materials important (and different for each college), but also how the topics themselves can adapt to these differences, and provide students optimal numbers of materials, and types of materials. Additionally, this is supported by current research into personalisation of LMS resources improving student outcomes (Mikić et al. 2022; Xie et al. 2019), as well as reducing cognitive load of students (Lange 2023).

How can predictive analytics models, incorporating LMS usage data, enhance the identification of at-risk students across different colleges? (RQ1.3)

This research has shown that at-risk students (students that achieve a Fail grade), are some of the more predictable groupings of students. With most of the Fail paths discussed in Section 5.4 Discipline-specific analysis, showing greater accuracy than paths depicting HD's (high distinctions), or other grade outcomes. This suggests that with proper monitoring of aspects of LMS usage (days active, especially in certain times of the semester), as well as better provision of materials tailored to the specific college, at-risk students can be better served via thorough investigation of LMS usage data, as well as a more customised approach.

Is dimensionality reduction necessary to accurately capture the essential aspects of LMS use, and what impact does this reduction have on the performance of predictive models? (RQ1.4)

This question was answered in Section 4.4.5 Experiment 1 results summary. Dimensionality reduction was ultimately not performed due to it being unnecessary, and potentially harmful in that it may lose some of the intricate nature of the dataset. Therefore, the dataset was kept at the size it was after the preprocessing stage discussed in Section 3.3 Data pre-preprocessing.

6.2.2. Do colleges differ significantly in approach and consistency? (RQ2)

RQ2.1. In what ways do colleges differ in terms of student engagement patterns, and how are these differences reflected in academic outcomes? (RQ2.1)

From the results of Section 3.4 Exploratory data analysis, and Section 4.3 Principal Component Analysis (PCA), several factors can be identified to explain the differences between colleges. One of the most important factors is that of engagement with the LMS, which has been shown to be a predictor of student performance as mentioned in Section 4.4 Experiment 1 results, and Section 4.5 Experiment 2 results.

Key Findings on College Differences and Academic Outcomes:

Engagement with the LMS was identified as a critical predictor of academic success, with varying levels of engagement observed across colleges.

Differences in Student Activity was shown in generally lower engagement, notably in colleges such as BGL, EPS, and HAS, where there are higher than expected numbers of students with low engagement levels. In contrast, higher Engagement levels were observed in MPH, NHS, and S&E, indicating more engaged students.

How do student behaviours, as captured through LMS data, vary across colleges, and what implications do these variations have for instructional design and student support services? (RQ2.2)

First, it must be acknowledged that the data available for this research was a limiting factor with answering this research question as thoroughly as possible. While the general patterns identified do provide a significant window into what types of engagement is performed, a more detailed dataset, with access to student demographics, and enrolment information (year level, enrolment type, and so on), future research into this area would provide a significant increase in information and potential insights.

From what was available, it was shown that colleges had significant differences in recorded student behaviours, with some colleges displaying higher levels of activity and engagement in the LMS compared to others. The numbers of specific topic components were also shown to be different across colleges, with usage of various component types also varying across colleges and being used for predicting at-risk students as well as high performing students.

The focus on content type (primarily videos for most colleges), as well as the identification of low engagement amongst students would seem to be some of the most important differences between colleges, and reflect in in student outcomes, which have also been shown to be different cross colleges, suggesting better knowledge of which aspects of topics need to be focused, on is necessary, as well as better monitoring of student engagement levels.

The data reveals a spectrum of engagement levels, with certain colleges exhibiting higher activity within the LMS. This variance in student behaviour points to differing needs and preferences, which instructional design must address to enhance learning outcomes.

Implications for instructional design

The observed variations in student behaviours necessitate a customised approach to instructional design, several suggestions of the types of customisations are shown in the following:

Customisation of content delivery

An LMS feature that may benefit students would be the adaptation of content types to match the engagement preferences observed in each college. For colleges like EPS and BGL, where engagement might be lower, integrating more interactive and varied content types could be beneficial.

While not explored in this thesis, the literature suggest that it may be beneficial to develop adaptive learning paths that adjust not only to engagement levels but also to individual learning styles, cognitive loads, and chronotypes. This could involve offering content in different formats and at various complexity levels, allowing students to engage with material in a way that best suits their cognitive and temporal preferences. This could potentially be done through the incorporation of the Felder-Silverman (1988) learning style model, Gardner's (1983) theory of multiple intelligences, or Kolb's (1981) Learning Style Inventory, depending on the preference of education staff, and availability of tests. This addition into the LMS may offer a more personalised learning experience for students and could include diagnostic tools for students to identify their learning styles upon course enrolment, enabling the LMS to tailor content delivery formats (e.g., visual, tactile, auditory) accordingly.

Implement strategies to manage cognitive load more effectively by personalising content presentation. This could include segmenting information into smaller, manageable units, using multimedia principles to balance verbal and visual information, and providing scaffolding where necessary to support learners' cognitive processing capabilities.

Dynamic adjustment of course components

The ability to adjust the visibility or appearance of component types within the LMS based on the engagement levels and preferences, for example, more engagement with quizzes for S&E, and HAS, or with the forums for EPS and NHS. Instructional design should be flexible, allowing for the incorporation of more or less of certain components (e.g., videos, quizzes) as needed, and the utilisation of recommendations made by this research to mediate which components. Plugins for LMSs such as the Kaiss, Mansouri & Poirier (2023) 'LearningPartnerBot' chatbot, or similar, would be ideal as that was shown to help recommend learning components via identifying a students the Felder & Silverman (1988) Learning Style.

Implications for student support

Given the variations shown in student behaviours with the LMS, student support services should be aware of the specific engagement patterns of each college, recommendations of what support services may be utilised are as follows:

Current interventions

In a systematic review of the online student support strategies and interventions reported between 2010 and 2020 by Rotar (2022), suggest that the effectiveness of the implementation of these strategies and interventions are dependent on when the intervention occurs, and how embedded within the LMS they are. Successful interventions of student support are suggested to be made early in the stage of education, such as at the intake of the student, or identified through behaviours in LMS logs.

Targeted support for at-risk students

It would be generally beneficial to utilising LMS usage data to better identify students who may be at risk of failing due to low levels of engagement (either with the LMS in general, or with components such as assignments, quizzes, or videos, depending on the college preference). This echoes the suggestions by Rotar (2022), and early identification of these students would produce the best outcomes. Additionally, integrating programs aimed at enhancing student self-efficacy, such as information on using the LMS, and providing positive feedback upon achieving set tasks. This would involve creating more interactive and engaging content that provides immediate feedback, thus helping students feel more competent and in control of their learning process. Student usage data is always generated by the LMS, through automatic logging and would simply require use of the data to better offer targeted interventions for support services, additional modules for feedback and monitoring would require more implementation.

Recognition programs for high performers

There is potential in developing recognition programs for students that demonstrate high engagement and performance, as this may encourage further engagement and high achievement. Colleges that have low levels of engagement such as BLG and EPS may be high priority for this approach, especially where the increased levels of engagement are key for higher performance. As mentioned above, this may help to assist with overall self-efficacy if students and provide a level of positive reinforcement for desirable tasks.

Customised scheduling

Offer flexibility in scheduling learning activities and assessments to accommodate different student chronotypes. This might include providing options for morning or evening activities and assessments to ensure students can work at their optimal times.

Implications for instructional design

As discussed by Spatioti, Kazanidis, and Pange (2022) the ADDIE instructional design model chosen for this research, is suited to diverse applications and domains. Therefore, with the application of plugins such as the Kaiss, Mansouri & Poirier (2023) 'LearningPartnerBot', or via AI analysis of LMS logs similar to the research presented by Kanchon et al. (2024), an accurate profile of students can be identified.

Enhance interactive learning

Further develop the LMS to dynamically adjust content and instructional components based on a comprehensive profile of each student, including their learning style, chronotype, self-efficacy levels, and cognitive load capacity. This could lead to a more engaging and less overwhelming learning experience. Especially in colleges with lower engagement, enhancing the number of materials made available (or shown from an available library of materials), would help to better stimulate student interest and participation.

Monitor and respond to engagement patterns

Implement more sophisticated tools for monitoring and responding to changes in student engagement patterns, taking into consideration the comprehensive set of individual differences outlined. This would allow for timely and effective adjustments in instructional design and student support. Additionally, the continuous monitoring of LMS data to identify shifts in student behaviour would allow for a better approach to identifying at-risk students, as well as potential high performing students that would need extra materials/motivation. This added feature would allow for timely adjustments in instructional design and support services.

Summary of differences

As shown in Table 52, differences vary significantly across colleges regarding activity levels and engagement, indicating the need for tailored instructional and support strategies.

There is an emphasis on videos across most colleges, with potential to expand to more interactive components based on student engagement data. There is a potential benefit of utilising the LMS data to identify at-risk and high-performing students, guiding targeted support and recognition efforts.

Table 52 - Differences between behaviours across colleges

Aspect	Observation	College examples	Recommended actions
Engagement & activity levels	Varied levels of LMS activity and engagement	Higher: S&E, NHS Lower: BGL, EPS	Customise instructional design per college needs. Targeted support for lower engagement colleges.
Content type usage	Differences in the usage of videos, quizzes, forums	Videos: Predominant across all colleges Quizzes: More engaged in S&E, HAS Forums: Higher usage in EPS, NHS	Diversify content types based on college preferences. Increase interactive elements in courses.
Identification of student groups	LMS patterns help identify at-risk and high performers	At-risk: Notable in BGL, EPS High performers: Frequent in S&E, NHS	Early intervention for at-risk students. Recognition programs for high performers.

The variations in student behaviours across colleges, as captured through LMS data, offer valuable insights for refining instructional design and enhancing student support services. By recognising and responding to these differences, educational institutions can create more responsive and effective learning environments. Tailoring instructional content and support services to the specific engagement patterns and needs of students across different colleges will not only improve student outcomes but also enrich the overall educational experience. Integrating continuous monitoring of engagement data will ensure that instructional design and support services remain dynamic and responsive to student needs.

What are the distinctive pedagogical approaches adopted by different colleges as evident from the LMS data, and how do these approaches correlate with student engagement and performance? (RQ2.3)

While it cannot be truly proven, due to the nature of the student dataset (only observations of activity, topic construction, and student outcomes are available to analyse), it can be inferred from observation. Teaching pedagogical approaches appear to vary significantly across colleges, as evidenced by the different structures of topic content provided by the LMS shown in Section 3.4 Exploratory data analysis. For example, Section 4.2.6 and Section 7.1 Appendix A: Additional tables and figures, show the large differences in topic composition across a variety of components.

This investigation into the distinctive pedagogical approaches adopted by different colleges, as evident from the LMS data, uncovers a complex tableau of teaching strategies and their impacts on student engagement and performance. The analysis provides a nuanced understanding of how varied educational methodologies correlate with levels of student interaction within the LMS, ultimately affecting academic outcomes.

Pedagogies across colleges

The analysis reveals distinct pedagogical profiles for each college, showcasing strategic pedagogical alignments with their unique educational goals and subject matters. For instance, the S&E utilises a multimedia-intensive and assessment-driven strategy, characterised by a larger average number of lecture videos and quizzes. This suggests an emphasis on delivering comprehensive content and regular knowledge assessments.

In contrast, the EPS and NHS prioritise interactive and discussion-based learning, evidenced by a greater average number of forum posts and participation materials. This approach likely fosters a more collaborative and reflective learning environment.

Pedagogies and student outcomes

Statistical exploration indicates significant correlations between pedagogical elements and student outcomes. Multimedia content and assessments in S&E are linked to higher engagement levels, potentially leading to improved performance, suggesting that engaging multimedia content coupled with frequent assessments can enhance learning effectiveness.

The HAS college, with its larger average number of videos, supports diverse learning preferences, possibly contributing to enhanced conceptual understanding and engagement.

Temporal effects on student engagement and performance

The semester-based analysis of engagement and performance highlights the impact of pedagogical timing. Increased interaction with quizzes and videos around mid-semester in colleges like HAS and S&E correlates with an uptick in performance, indicating that strategic distribution of resources can optimise student success.

Course structure on engagement

The course structure significantly affects engagement. For example, the BGL and HAS, with a larger average number of assignments, encourage regular student engagement, leading to a more consistent learning journey throughout the semester.

Analysis of Engagement Levels

Colleges like S&E and NHS, which demonstrate high engagement levels, share a common approach of incorporating a balanced mix of multimedia content, interactive activities, and regular assessments. This multifaceted strategy contrasts with colleges showing lower engagement levels, highlighting the effectiveness of diverse learning activities in maintaining student interest and participation.

Recommendations for improving pedagogical strategies

Based on the insights from the analysis, tailored recommendations are proposed for each college to bolster student engagement and performance. For example, with colleges like BGL and EPS, integrating more multimedia teaching aids could address lower engagement levels, enriching the learning experience.

Encouraging more interactive and discussion-based activities in S&E and HAS may promote a deeper understanding of complex concepts. Adjusting course structures in colleges such as NHS to include regular, formative assessments could keep students actively engaged with the course material, enhancing learning outcomes.

Summary of important pedagogical elements by college

Larger average number of **Videos** (HAS, MPH, and S&E).
Larger average number of **Quizzes** (HAS and S&E).
Larger average number of **Assignments** (BGL, HAS, and S&E).
Larger average number of **Forum** posts (EPS, MPH, and NHS).
Larger average number of **Participation** materials (BGL, EPS, and NHS).
Larger average number of **Support** materials (EPS, and NHS).

The differences suggest a focus on specific components, as well as LMS usage such as more forum usage, or more interactions with participation materials. Focus on elements, such as Assignments and Quizzes, would suggest a more practical approach. Whereas, a more interactive approach involves forum usage, Participation materials, would imply a more discussion-based approach. The distinctive pedagogical approaches of different colleges, as reflected in LMS data, exhibit clear correlations with student engagement and performance levels. This detailed analysis not only elucidates the diverse strategies across colleges but also highlights the potential for pedagogical refinements to improve educational outcomes. Future research should include qualitative feedback from students to further refine teaching methodologies, aligning them with student learning preferences and needs. By evolving pedagogical strategies based on comprehensive data analysis and feedback, colleges can create more engaging and effective learning environments, catering to the diverse educational needs of their student body.

6.3. Research contribution

This study made a significant theoretical contribution to the field of E-Learning of a predictive data analytics model for E-Learning across disciplines to impact student performance. Unlike past studies that have focused on a specific discipline (as discussed in Section 1.4.1), this study is innovative by integrating the machine learning and E-Learning literature to develop a model across disciplines. The predictive model evaluates students' performance against E-Learning pedagogical approaches across various disciplines, thereby assisting in the identification of best practices for each field.

Additionally, the research was innovative as it employed tree-based machine learning algorithms to not only predict the college of a student through usage data accurately but also extract and leverage crucial topic structure features. This approach elucidated effective pedagogical strategies in each discipline, enhancing our understanding of E-Learning dynamics. Thirdly, the study introduced a targeted and nuanced approach, utilising instructional design models like ADDIE to provide educators with structured yet adaptable guidance. This represents a departure from the one-size-fits-all solutions predominant in existing E-Learning research, offering a tailored methodology suited to distinct educational domains.

The analysis contributes discipline-specific insights into LMS usage that is not commonly discussed, as referenced in Section 1.4 Significance and contribution of the research, and from Section 2.5.5 Interdisciplinary differences. This research has uncovered a wide variety of college/discipline-based differences in both LMS usage, as well as topic composition. Highlighting the importance of not only personalised engagement, but personalised design of LMS materials across colleges.

It demonstrates that while some disciplines may require more interactive elements such as forums and quizzes, others benefit from extensive support materials and diverse learning activities. This understanding is critical at multiple stages of the instructional design process, as referenced in Section 2.4.4 Instructional design models. Understanding both what students needs are, as well as what teaching practices, and materials are available is critical.

The research provides a data-driven approach to investigating the differences between different disciplines/colleges, regarding LMS usage, and highlight a wide variety of educational strategies and differences in student utilisation of materials. This highlights a distinct need for a customised approach rather than a one-size-fits-all approach. Additionally, these results emphasise the benefit of machine learning, and datamining in identifying successful student LMS engagement patterns, and how individual differences in learning styles and teaching methodologies can be utilised to influence LMS design and delivery.

6.4. Implications for teaching practices

The overall findings suggest that teaching practices may benefit from a more discipline specific understanding of how various LMS components are used by students. This section will outline the primary differences between colleges at a teaching level and recommend approaches for best attending to these differences.

6.4.1. College of Business, Government, and Law (BGL)

Features of college

Topics from this college appear to have a larger focus on videos, assignments, and participation materials, while interacting with fewer support materials, and overall having lower engagement with the LMS.

BLG overall has a medium number of videos on average per topic, as well as a medium number of assignments on average. This would suggest a more practical approach, with less focus on interactive LMS materials, and student interactions potentially focused outside the LMS.

Inferred pedagogy

Given BGL's reliance on structured content like videos and assignments, a behaviorist approach emphasising clear learning outcomes and immediate feedback on quizzes may enhance learning efficiency. Incorporating instructivist elements, such as direct instruction through video lectures, can provide a solid foundation in complex subjects. LMS design should facilitate these pedagogies by allowing for the easy creation of assessment materials that offer instant feedback and supporting diverse video content.

Recommendations

While there is an overall lower average in activity levels for BLG, these levels are still important with regards to identifying at-risk students. This would indicate that what interactions do occur on the LMS are important. Therefore, it would be encouraged to monitor overall student activity levels, especially with video content on the LMS, as these interactions have been shown to be most beneficial to students during mid semester, and at exam times, where the LMS materials are likely used for revision and supplementary to in-person interactions.

6.4.2. College of Education, Psychology, and Social Work (EPS)

Features of college

Topics from this college appear to focus assignments, and support materials, while interacting with the LMS in a similarly low level as BGL. However, EPS was also noted for being one of the colleges with the largest on average forum usage (aside for NHS), and the largest user of participation materials, suggesting student-teacher, and student-student communication via LMS is crucial. Additionally, failing EPS students appearing to have even lower levels of activity supporting the importance of these interactions, much like with BLG.

Overall, EPS appears to have on average, a very high number of participation materials, a medium number of support materials, and a relatively high number of forum posts, suggesting a focus on interactions with the topic, and with other students and teachers.

Inferred pedagogy

EPS's focus on forums and participation suggests a constructivist approach, where learning is built through interaction and reflection. The LMS should support collaborative projects and discussion boards that encourage active participation and facilitate peer learning. Features enabling learners to construct knowledge through dialogue, group work, and the integration of theory into practice can also be useful.

Recommendations

It would be recommended that EPS educators focus on student communication via forums, or through participation, additionally, identifying at-risk students would involves recognising those with limited activity, especially those who have not engaged with the topic at all for certain periods, most likely missing out on valuable interactions with students and teachers.

6.4.3. College of Humanities, Arts, and Social Sciences (HAS)

Features of college

Topics from this college appear to focus on videos, participation materials, and forum posts, while like BLG, and EPS, have a larger number of low engagement students. Compared to other colleges, the number of interactions with participation materials was relatively high, but more engagement was shown with 'Other' type materials, which are in general associated with LMS background tasks, general engagement with the LMS, but not to specific educational content.

HAS was also shown to have on average a fairly low number of quizzes and assignments. Additionally, showing a general focus on interaction with the LMS and in topic participation, but not specifically to many forum posts, or to high levels of student-student interactions.

Inferred pedagogy

HAS's emphasis on videos, participation, and forums aligns with constructivist approach of learning through exploration and interaction. The integration of connectivist principles, such as networking through social media-like forums within the LMS, can foster community among students. Designing the LMS to support user-generated content and peer feedback can further enhance the learning experience.

Recommendations

While HAS didn't appear to have a larger number of assignment materials than other colleges, assignments were shown to be important for the prediction of at-risk students, especially from students with low interactions with assignments in the later periods of the semester. This is also suggested with an overall low number of days engaged with the topic, being predictive of a Fail. While engagement with assignments and support materials, and social interaction materials is important for high achievers.

Interestingly, the size of the topic's forum was also predictive of higher performing students; topics with smaller forums showing higher likelihood of having HD students.

It would be recommended for HAS topics, to focus on getting students engaged with the LMS, as well as interacting with assignments and support materials and early interactions with other students and teachers.

6.4.4. College of Medicine, and Public Health (MPH)

Features of college

Topics from this college were shown to have a larger number of quizzes on average (second only to S&E), and a lower on average number of participation materials. In addition, topics with many videos, and students with low interactions during the exam period were shown to be potential predictors of failure. MPH was also shown to have fewer students with a low level of engagement with the LMS, and more medium to highly engaged students.

This would suggest a preference for frequent, interactive assessment, rather than static videos requiring no interactivity, but not for general interactive purposes or topic related participation.

Inferred pedagogy

The focus on quizzes and interactivity in MPH suggests a blend of cognitivist strategies for deep understanding and behaviorist approaches for reinforcement learning. The LMS should offer adaptive learning paths that adjust to individual student performance, providing tailored resources as needed. Incorporating simulation-based learning modules can also support practical application of theoretical knowledge.

Recommendations

The negative association with topics consisting of many videos is interesting, considering the college was shown to have on average a relatively high number of videos. In addition, the preference for more days active with the LMS along with the negative association to videos, suggest that the focus for MPH should be in developing, and delivering more interactive content (less videos, more quizzes), as well as maintaining overall student engagement especially early on, and during exam periods, to better promote positive student outcomes.

6.4.5. College of Nursing, and Health Sciences (NHS)

Features of college

This college, like MPH, showed a greater number medium to highly active students, with a medium number of forum posts on average, assignments, quizzes, and participation materials, while a relatively larger on average number of support materials.

For MPH, the primarily emphasis would appear to be assignments and quizzes, as well as level of engagement.

Inferred pedagogy

NHS's balanced use of LMS components calls for a cognitivist approach to facilitate the understanding of complex concepts through diverse materials. Constructivist elements, such as case-based learning within the LMS, can encourage application of knowledge in real-world scenarios. Features that support interactive case studies and virtual simulations can enhance the learning experience.

Recommendations

For NHS, engagement with the topic appears to be a very large predictor of at-risk students, with assignment interactions during mid semester, and in the exam periods, showing as being predictive of a Fail. Additionally, the number of quizzes viewed also showed a positive relation to high performers, indicating that quizzes may play a large part of assessment, in addition to regular assignments.

Overall, this would suggest a strong focus on providing engaging quizzes and assignments, and monitoring student activity levels, especially during the middle and at the end of the semester.

6.4.6. College of Science and Engineering (S&E)

Features of college

For this college, the engagement with the LMS is the most predictive aspect, where other colleges are defined more by their content, S&E is more related to levels of interaction. S&E was shown to have on average, the largest number of videos, quizzes, and assignments, which may be indicative of the technical nature of the subjects, potentially for more knowledge in implementing these materials into the LMS. S&E also was shown to have less low engages students, and more medium and highly engaged students. It is also worth noting, that S&E has nearly three times the number of videos on average, compared to the next highest college (MPH), showing a large investment into videos.

Inferred pedagogy

S&E's extensive use of multimedia content and high engagement rates suggest a constructivist approach, where hands-on problem-solving and project-based learning is important. Incorporating connectivist elements, such as integration with external resources and platforms for coding practice or design projects, can keep students engaged and up to date with industry standards. The LMS should facilitate easy access to external tools and resources, promoting a culture of continuous learning and connection.

Recommendations

Overall, S&E was harder to directly recommend solutions due to the complicated nature of their decision tees, however, common aspects such as days active with the LMS, and more focus on students interacting with assignments, support materials, and social materials before the exam period would be something to recommend for positive student outcomes.

6.4.7. Overview

In general, educators should consider integrating more interactive and supportive LMS components that align with the needs of their students, such as forums and quizzes for Science and Engineering based topics and a diverse selection of support materials for Nursing and Health Sciences related topics. Additionally, by tailoring the LMS design and instructional strategies to align with the pedagogical approaches mentioned to be best suited for each college, educators can create more engaging and effective learning experiences that are customised to suit the diverse needs of students across various disciplines.

6.5. Implications for LMS design

This research identifies the necessity for a LMS to be flexible and adaptable, accommodating a wide range of teaching and learning styles across different disciplines. The results from this research suggest that the LMS design process should not adopt a one-size-fits-all policy regarding academic disciplines, but rather follow a more customisable methodology (as will be discussed in Section 6.7), to meet the specific discipline-based educational needs of students. The following detail several practical examples of how LMS platforms can integrate the findings from this research.

LMS dashboards

To better address the engagement patterns of students shown across disciplines, LMS platforms could include a customisable dashboard to allow instructors to highlight (or to hide) certain types of content based on the specific needs of their discipline, as was mentioned in Section 2.6.1. For example, disciplines that rely heavily on visual learning, such as Science and Engineering, could benefit from enhanced capabilities for embedding and interacting with video content, including interactive features like embedded quizzes or discussion prompts within videos. The flexibility to adapt LMS dashboards aligns with findings on the importance of personalisation in E-Learning environments (Brusilovsky & Millán 2007; Xie et al. 2019).

College specific learning pathways

An LMS could offer tools that allow instructors to create learning paths tailored to their discipline's unique requirements, as was discussed in Section 2.6.2, regarding customisation of content. For example, instructors from the College of Humanities, Arts, and Social Sciences could involve integrating external resources and forums to foster broader discussions and encourage interaction around course materials. This approach is supported by connectivist theories (Siemens 2004) and the need for constructivist learning paths that engage learners in active knowledge-building processes (Piaget & Inhelder 1967).

Quiz modules and support materials

Similar to the previous recommendation, and as discussed in Section 2.6.2 about creating dynamic difficulty levels of materials, based on student performance, in that case, with the KT-IDEM model by Pardos & Heffernan (2011) In disciplines like Medicine and Public Health, where interactive assessments were shown in the results to be particularly important, LMS platforms could provide advanced quiz modules supporting adaptive learning. For Nursing and Health Sciences, easy creation, and distribution of supplementary materials, such as case studies, could support extensive support material needs. This aligns with the findings of personalised learning strategies outlined in Section 2.3.4 by Fariani, Junus & Santoso (2022), where adaptive learning methods were linked to higher student satisfaction and engagement.

Engagement monitoring

Incorporating analytics tools within the LMS to monitor student engagement levels can help identify at-risk students early, as discussed in Section 2.6.2. This feature would be particularly useful in disciplines like Business, Government, and Law, where the results showed early access to crucial content like video lectures is essential for student success. This is further validated by research on learning analytics that highlights the potential of micro-level data to optimise real-time interactions and improve student outcomes (Avella et al. 2016; Fischer et al. 2020).

Forums

Given the importance of forum interactions in certain colleges, LMS platforms could include features that enhance engagement, such as gamification elements for active participation or automated prompts encouraging contributions. Examples of the beneficial impact of gamification include research by; Romsi, Widodo & Slamet (2024) finding a positive impact of gamification on slow-learners, Subiyantoro et al. (2024) finding a significant increase student engagement and motivation in learning, and Yu, Yu & Li (2024) finding a positive effect on educational outcomes. This builds on the gamification discussion in Section 2.2.3, where tailored gamification was shown to improve student engagement through interactive, data-driven adaptations (Denden et al. 2024).

By implementing these recommendations, LMS designers can ensure that the platform is not only consistent with the diverse needs of students and instructors across different disciplines but also provide a more engaging and effective learning experience for students. These enhancements, grounded in the findings of this research, highlight the critical role of LMS design in adapting to the evolving landscape of higher education and the specific requirements of each academic discipline.

6.6. Summary of findings

This research embarked on an extensive examination of the FLO LMS engagement and performance data across different colleges. Utilising rigorous methodological framework consisting of statistical analyses, Principal Component Analysis (PCA), and machine learning experiments. The primary aim was to uncover the interrelationships between student engagement patterns within the LMS and their academic outcomes, as well as to predict college affiliation and student grades with high accuracy. Here are the key findings from the study:

Grade distribution and LMS engagement

Significant disparities in grade distributions were identified across colleges, suggesting variations in grading standards or teaching methodologies. The analysis of LMS engagement, including student attendance and activity levels, revealed patterns of student interaction with the LMS that correlated with academic performance, highlighting the influence of engagement on academic outcomes.

PCA results

PCA provided a deeper understanding of the dataset's dimensionality, emphasising the complexity and interconnectedness. While not necessary for attribute reduction, the analysis highlighted the richness of the dataset and the importance of retaining comprehensive attribute sets for in-depth analysis.

Predicting student grades and college affiliation

Experiment 1: Focused on predicting grades across all colleges, identifying critical engagement metrics and LMS components that correlate with academic success. The results emphasised the complicated nature of academic achievement, highlighting the importance of active engagement and resource utilisation within the LMS.

Experiment 2: Aimed at predicting college affiliation, demonstrating that specific patterns of LMS interaction are indicative of college affiliation. This finding suggests that colleges exhibit unique LMS usage patterns, topic construction, and pedagogical approaches.

Experiment 3: Focused on college-specific grade predictions, further refining the understanding of academic performance within distinct educational contexts. This experiment reinforced the critical role of engagement metrics in predicting grades, offering information into tailored pedagogical strategies for enhancing student learning experiences.

Variations between colleges

The research identified significant college-based differences in LMS engagement and academic performance, highlighting the wide variety of educational practices and student behaviours across different academic disciplines. However, these variations also suggest opportunities for targeted interventions and customised educational strategies to support student success.

Implications for educational practice

The research outcomes offer valuable insights for educators, administrators, and LMS designers, suggesting that a deep understanding of student engagement patterns through data-driven analysis, along with the alignment of LMS features with overall pedagogical objectives are very important for optimising student learning experiences and academic outcomes.

6.7. Contribution to education practitioners and instructional designers

To better support educational design practices, results from this research can be applied to modifying (as shown in Figure 70) the ADDIE model of instructional design (as mentioned in Section 2.4.4 Instructional design models). While other instructional design models may also benefit, ADDIE's effectiveness across various E-Learning environments (Spatioti, Kazanidis & Pange 2022), improvements in student performance (Almelhi 2021) and flexibility across different applications and domains (Spatioti, Kazanidis & Pange 2022) highlight why this instructional model was specifically chosen for these recommendations. The following recommendations to further improve the Analysis phase are supported by the research, particularly those discussed in Section 6.4 Implications for teaching practices.

Analysis phase

The analysis phase is the most critical for instructional design, as it is where the educational objectives, and needs analysis for students takes place. This phase is key for implementing any predictive analytics that could better identify the needs of learners, and from any specific discipline-based needs that may arise. Integrate predictive analytics into analysis: To enhance the analysis phase, incorporate predictive analytics to better identify at-risk students and tailor content accordingly. This approach can leverage LMS usage data to anticipate student needs and potential failure points, allowing for a more proactive design of educational materials and interventions.

BGL: Behaviourist or Instructivist pedagogies recommended - Quizzes, and video lectures.

EPS: Constructivist pedagogy is recommended - Collaborative projects / discussion boards.

HAS: Constructivist pedagogy is recommended - Networking and social media-like forums, user-generated content and peer feedback.

MPH: Cognitivist for deep understanding and behaviorist for reinforcement learning. Simulation-based learning, and practical application of theoretical knowledge.

NHS: Cognitivist pedagogy is recommended - Case-based learning, such as interactive case studies and virtual simulations.

S&E: Constructivist pedagogy is recommended - Hands-on problem-solving and project-based learning.

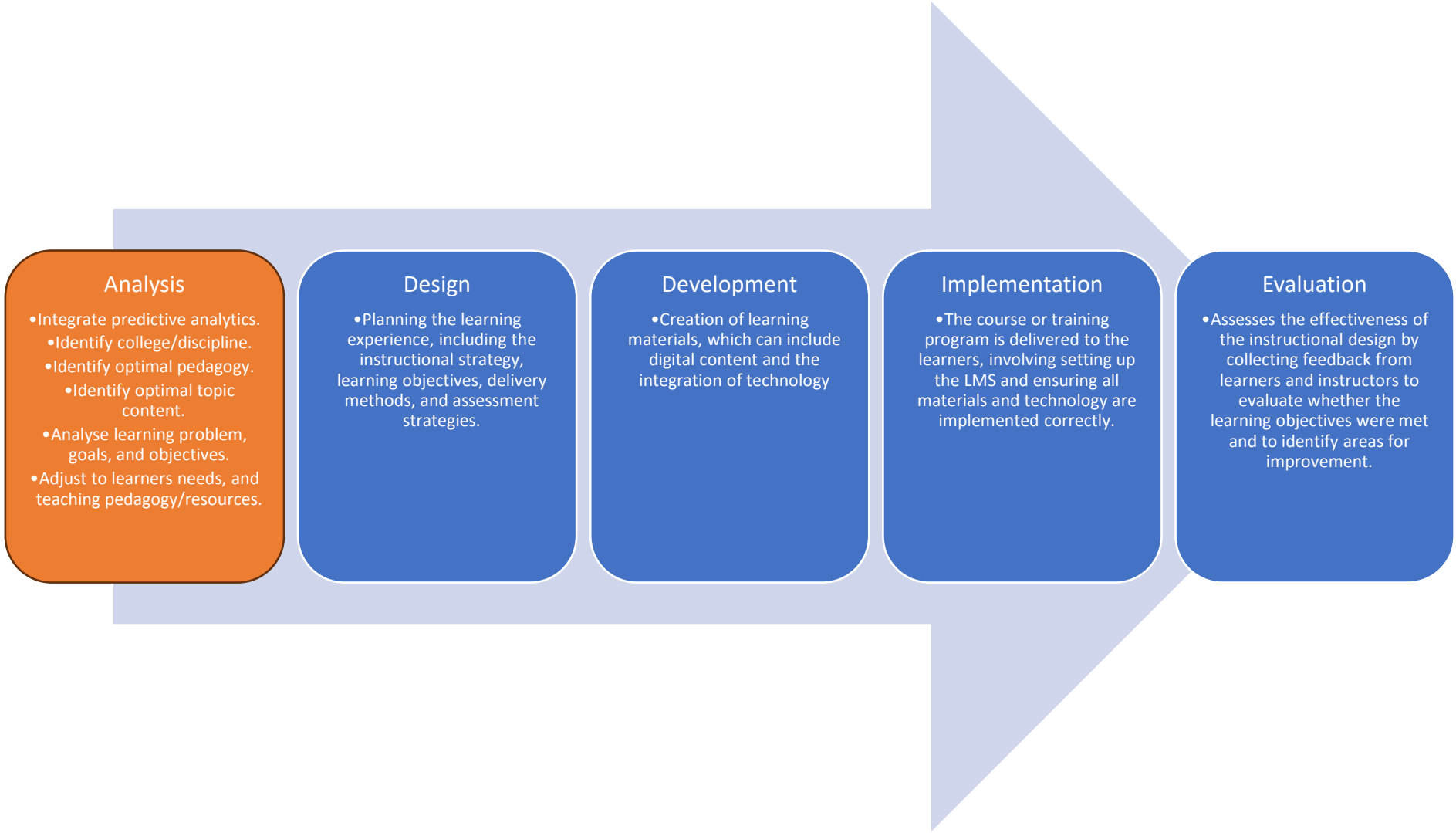

Figure 70 - ADDIE model (Grafinger 1988) with adjusted analysis phase developed from this thesis

6.8. Directions for future research

One of the most important potential future uses for this research methodology is to re-apply the processes from this study on current LMS data, given that the dataset for this research was acquired pre-COVID19. This would add additional layers of analysis, especially with regards to activity patterns and preferences. This will provide future insights about the direct implications of the COVID-19 Pandemic, and its effect on universities, and LMS usage.

Additional research into both the patterns of activity at the end of semester, activity throughout the semester, would be worthwhile. Such as how students presented with this information compared to those with no such intervention change their study patterns, or overall learning strategies. While this is a more direct method (rather than data analysis), it would provide additional information for analytical research as well. Additionally, further research into patterns for credit/pass/distinction paths would help to identify what works well or needs work for most students, not just the at-risk and high performers.

Identifying pre-COVID-19 and post-COVID-19 usage patterns (temporal aspects would be beneficial to compare, especially during lock-down periods), as well as grade distributions, and changes in topic composition, would provide a wealth of information, especially in a cross-discipline perspective.

Specific research into S&E college patterns to better unpack the substantial differences that were shown (such as the large volume of videos identified), would potentially be very rewarding.

Further research into better parameterisation and adjustment of the current decision tree algorithms, and how they may be further improved is a natural extension of this research. With more time to perform additional tests on existing and future data, further insights may be identified. Furthermore, deeper analysis, and comparison of more algorithms, aside for just the better performing ones would also potentially see interesting results.

Additionally, the exploration of black-box style Neural Networks is recommended for future research, as this area was avoided in this research context due to the lack of interpretability of the resultant models. With the identification of college-based patterns being shown to not only be possible, but to be relatively accurate with the chosen techniques from this study, incorporating higher performing models would be beneficial for research into performance and classification.

Another direction for research could potentially be in adapting all stages of the ADDIE methodology not just the Analysis stage. This would delve more into areas outside of predictive analytics however, but the other stages would potentially be suitable for a similar analysis and improvement.

Finally, this research can be presented to both Flinders Learning and Teaching executives as well as presented in further published work for further comment, feedback and confirmation of potential effectiveness.

6.9. Limitations

The primary limitation of this research was the age of the dataset. With many landscape altering events occurring between the date of the dataset's retrieval and today. Additionally, the dataset is from only one institution (Flinders University) and represents only Flinders University students and colleges from the standard Semester 1 and Semester 2 enrolment. The data also did not include any demographic information that would have provided a large amount of additional information for research. Nevertheless, the insights derived from this research is valuable in comparing E-Learning approaches and impacts across disciplines which directly addresses the key research questions.

6.10. Chapter summary

Reflecting on the journey undertaken in this research, there has been significant strides in both theoretical and practical areas of E-Learning. Theoretically, this research has developed predictive data analytics models that leverage machine learning to explore E-Learning across various disciplines. The use of these models in a targeted college-by-college approach has been shown to be innovative for evaluating student performance against a wide variety of E-Learning pedagogical strategies aiding in the identification of potential best practices for each academic field.

Practically, the insights derived from the research helps to better allow educators, with data-driven strategies to tailor E-Learning implementations more effectively. By implementing relatively adaptable and customisable educational approaches, educators can significantly enhance their pedagogical effectiveness and improve student engagement across different disciplines. This research highlights the use of tree-based machine learning algorithms to dissect LMS usage patterns across disciplines, offering educators additional insights into how they can better optimise their digital learning environments.

Looking at the potential future of research into E-Learning, this research lays the groundwork for further exploration into the evolving landscape of E-Learning, particularly in the context of college specific practices, and more interestingly with post-COVID-19 educational practices across colleges. Future research should focus on advanced analytical models and cross-institutional studies to broaden the scope of understanding and application of E-Learning strategies.

In summary, this research has carved out an important niche in the intersection of E-Learning, machine learning, and college/discipline differentiation. Providing a useful framework for educators, that examines the impact of digital engagement on student educational outcomes. By offering a blend of theoretical insights and practical recommendations, this research contributes to the enhancement of E-Learning practices, paving the way for a more informed and effective use of digital platforms in education.

7. Appendices

7.1. Appendix A: Additional tables and figures

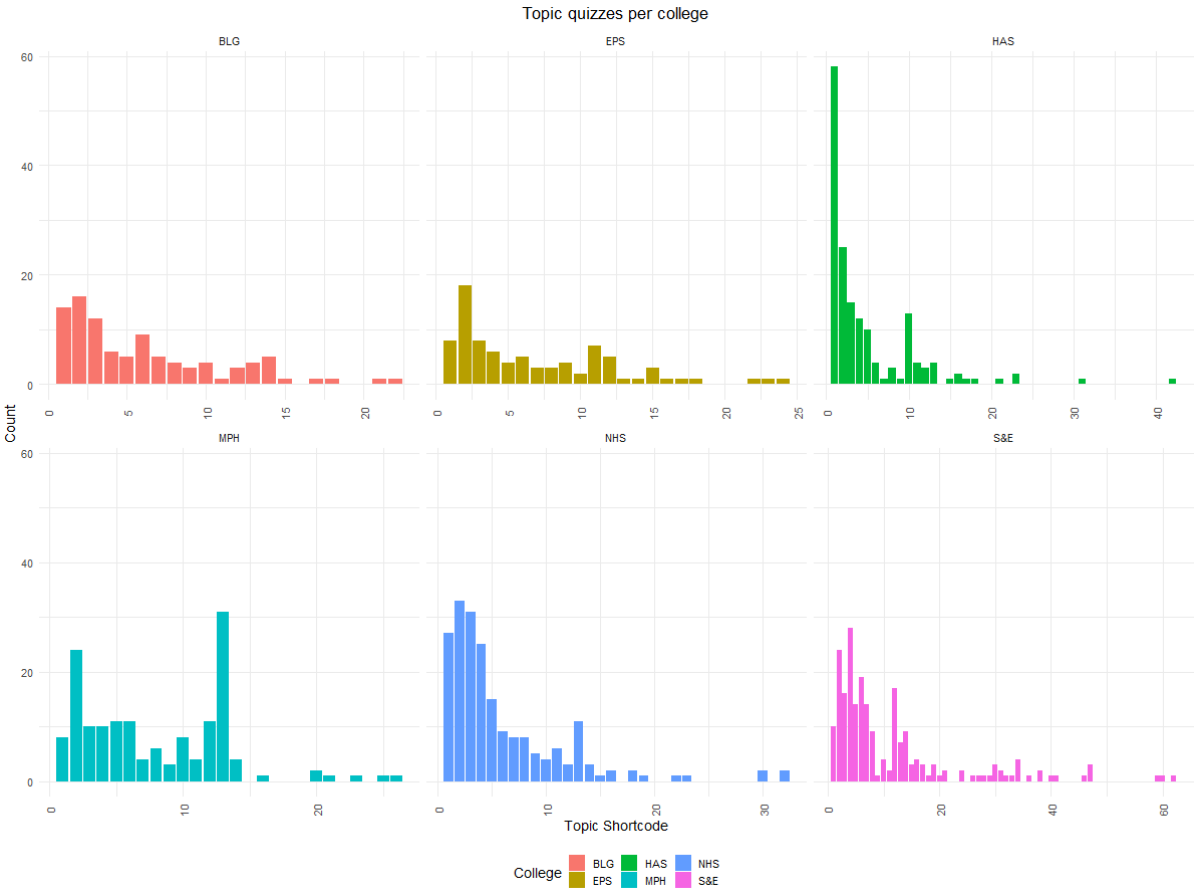

Figure 71 - Topic quiz modules grouped by college

Table 53 - Topic quiz modules grouped by college

college	quizzes sum	quizzes avg	quizzes max
BGL	581	0.87	22
EPS	572	0.82	24
HAS	799	0.67	42
MPH	1,216	2.95	26
NHS	1,178	2.04	32
S&E	2,433	2.94	62

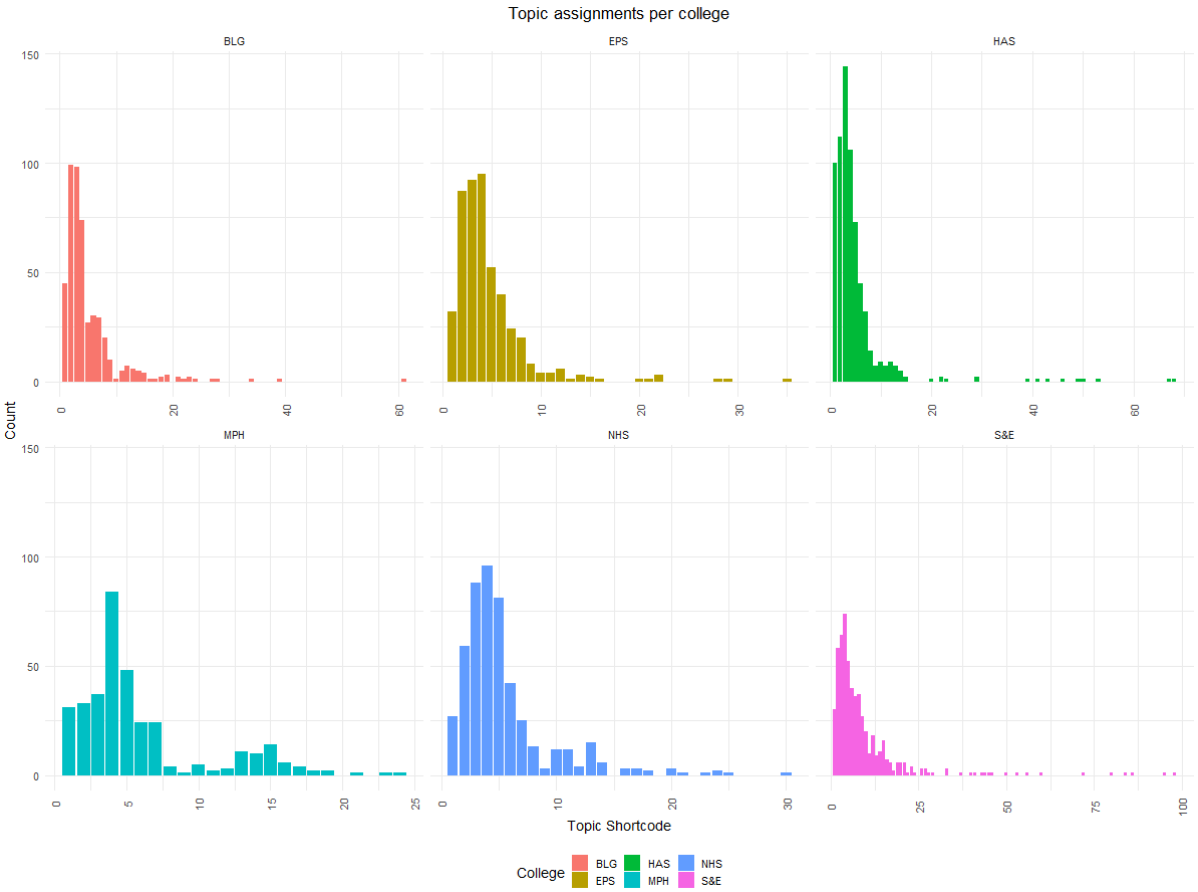

Figure 72 - Topic assignment modules grouped by college

Table 54 - Topic assignment modules grouped by college

college	assignments sum	assignments avg	assignments max
BGL	2,407	3.60	61
EPS	2,218	3.18	35
HAS	3,281	2.75	68
MPH	2,044	4.96	24
NHS	2,677	4.64	30
S&E	5,055	6.11	98

Figure 73 - Topic forum posts grouped by college

Table 55 - Topic forum posts grouped by college

college	forum sum	forum avg	forum max
BGL	1,906,267	2,849.43	56,615
EPS	3,205,033	4,591.74	84,909
HAS	1,444,448	1,210.77	55,303
MPH	1,328,809	3,225.26	32,130
NHS	3,189,397	5,527.55	83,496
S&E	2,054,149	2,483.86	149,631

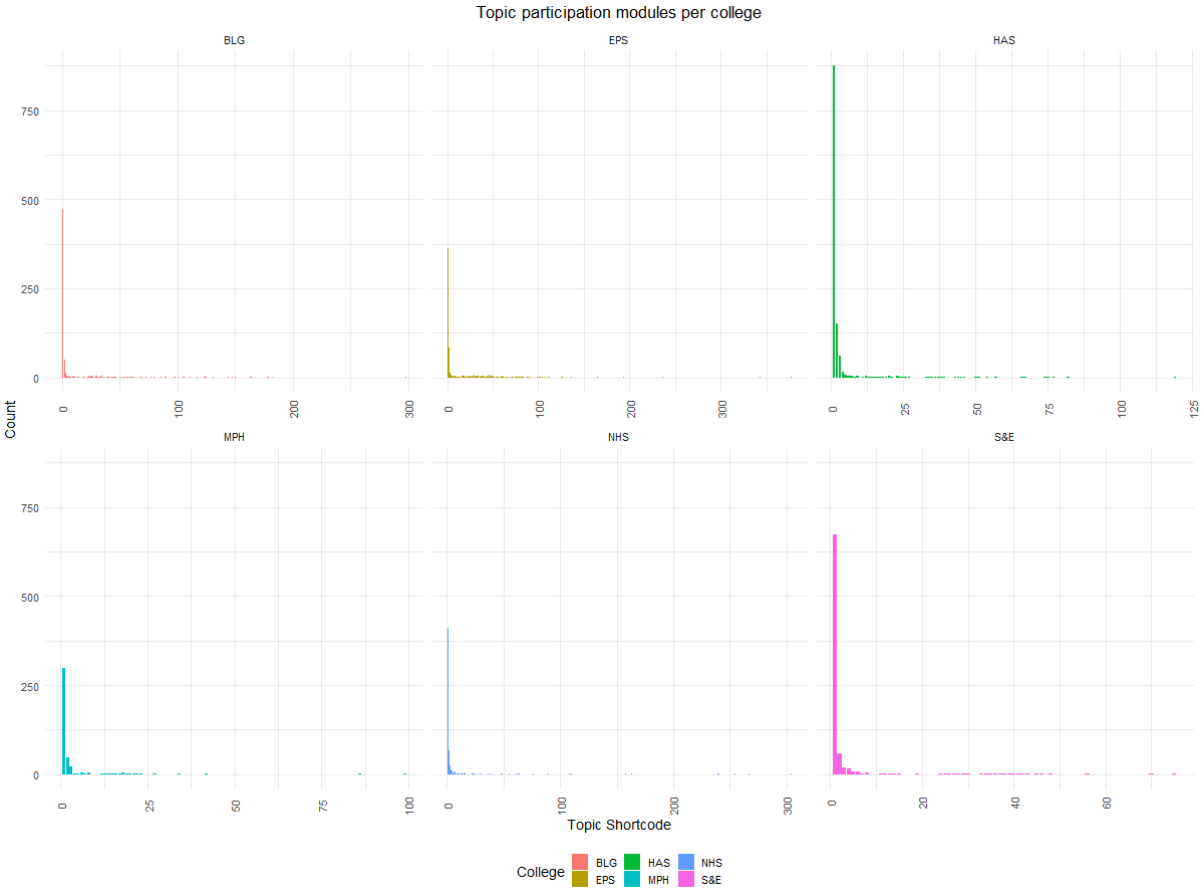

Figure 74 - Topic participation modules grouped by college

Table 56 - Topic participation modules grouped by college

college	participation sum	participation avg	participation max
BGL	7,346	10.98	298
EPS	11,606	16.63	377
HAS	3,790	3.18	119
MPH	1,228	2.98	99
NHS	3,423	5.93	304
S&E	2,302	2.78	75

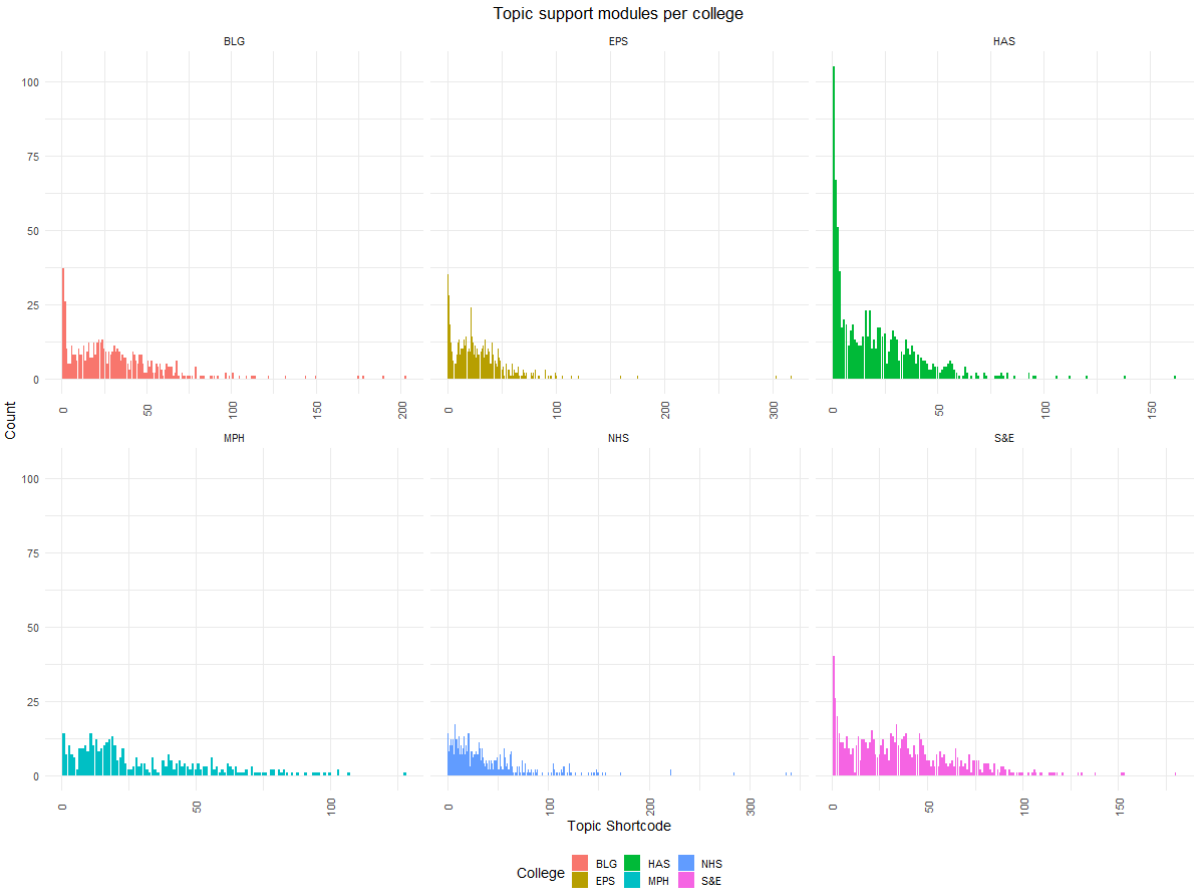

Figure 75 - Topic support modules grouped by college

Table 57 - Topic support modules grouped by college

college	support sum	support avg	support max
BGL	17,494	26.15	203
EPS	17,045	24.42	317
HAS	16,984	14.24	162
MPH	11,041	26.80	128
NHS	20,934	36.28	341
S&E	26,463	32.00	180

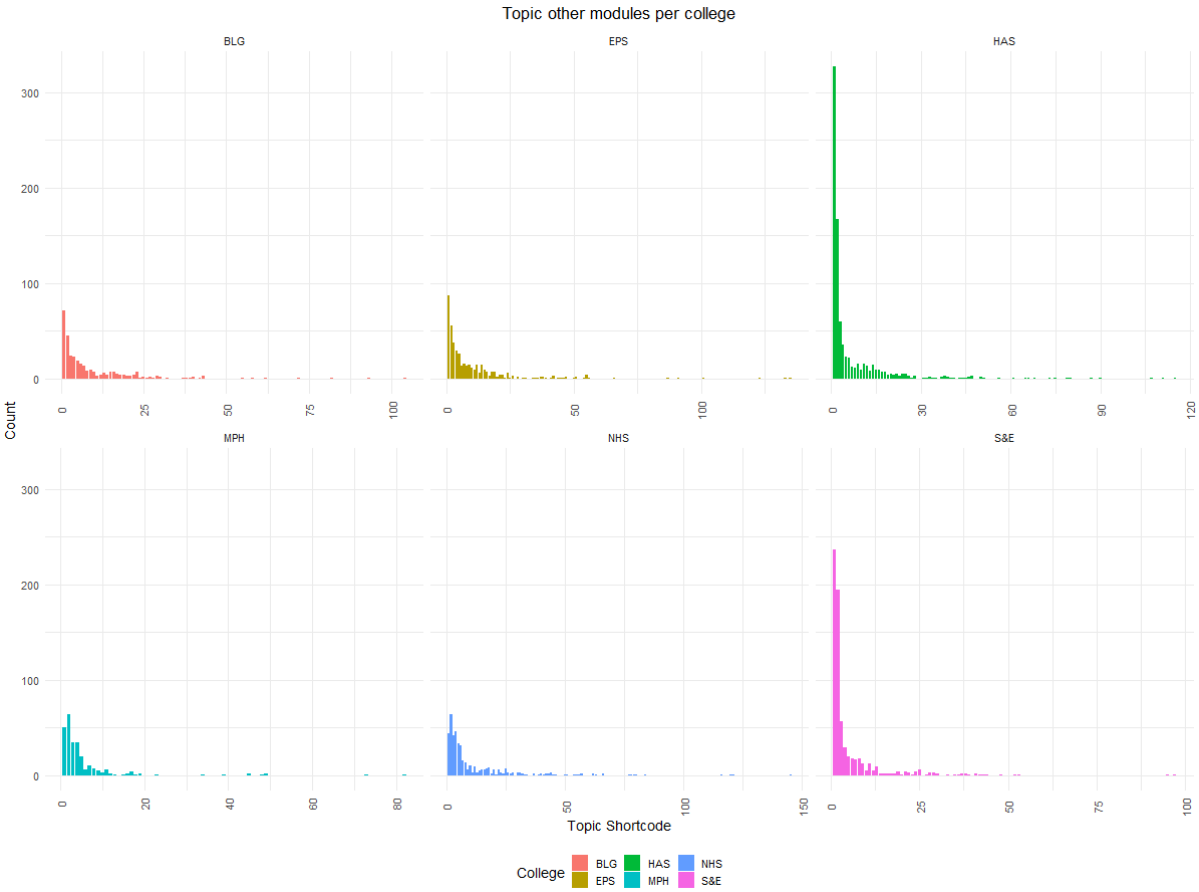

Figure 76 - Topic 'other' modules grouped by college

Table 58 - Topic 'other' modules grouped by college

college	other sum	other avg	other max
BGL	3,111	4.65	104
EPS	4,687	6.71	135
HAS	5,850	4.90	115
MPH	1,517	3.68	82
NHS	5,261	9.12	145
S&E	3,634	4.39	97

Table 59 - Performance of ensemble grade prediction models (with college)

	Weighted auROC	Difference
REPTree	0.68	-
RandomForest	0.71	+4.41%
RotationForest	0.70	+2.94%
AdaBoost (RT)	0.68	+0%

Table 60 - Performance of ensemble grade prediction models (without college)

	Weighted auROC	Difference
REPTree	0.70	-
RandomForest	0.72	+2.86%
RotationForest	0.71	+1.43%
AdaBoost (RT)	0.70	0%

Table 61 - Performance of ensemble college prediction models (with grade)

	Model Size	Difference
REPTree	193,526.88	-
RandomForest	490,144,973.28	+99.96%
RotationForest	4,541,505.88	+95.74%
AdaBoost (RT)	197,522.32	+2.02%

Table 62 - Performance of ensemble college prediction models (without grade)

	Model Size	Difference
REPTree	195,009.72	-
RandomForest	492,437,264.08	+99.96%
RotationForest	4,749,426.72	+95.89%
AdaBoost (RT)	193,269.96	-0.9%

7.2. Appendix B: Initial analysis into predictive analytics and E-Learning (pilot study into MOOC student performance)

7.2.1. Introduction of pilot study

Before the underlying hypothesis and overall methodology necessary for this research was developed, a small pilot study was developed. This was to help identify what would be necessary for literature requirements, methodological approaches, software, and tools necessary, and more importantly, what data would be required for a valid analysis.

This process was undertaken over several months at the beginning of candidature and was later published in the e-Proceeding of the 5th Global Summit on Education 2017, held in Kuala Lumpur, Malaysia (Wilden, Shillabeer & deVries 2017). Passages from the following sections have been included from that publication and are clearly identified to satisfy Flinders University rules for inclusion of published material in thesis.

7.2.2. Predictors for success (Wilden, Shillabeer & deVries 2017)

Given that the intent underpinning e learning participation is varied, it is no surprise that there is conflicting research regarding predictors of e learning success. Some suggest that more interactions and engagement with online activities leads to greater success and a lower chance of withdrawing (Castaño-Muñoz, Duart & Sancho-Vinuesa 2014; Morris, Finnegan & Wu 2005; Ramos & Yudko 2008). However, others (Beaudoin 2002; Garrison, D Randy & Cleveland-Innes 2005) argue that frequent interaction does not necessarily correlate to better performance, and that the quality rather than quantity of interactions is the real predictor of success.

Research also suggests that beyond quantity and quality of interaction, type of interaction may be a predictive factor in success. Forum posts in particular have been identified by some as an important determinant (Kizilcec, Piech & Schneider 2013; Morris, Finnegan & Wu 2005), however others suggest that there is no real statistically significant link (Beaudoin 2002; Picciano 2002; Ramos & Yudko 2008). Further research is clearly required.

Although there is little agreement, most published research on E-Learning success factors suggests that there are three key factors that influence the potential for success: Quantity of interactions, Quality of interactions and Content Type for interaction.

The quantity of interactions suggests the level to which the content is providing relevant and unknown information (to the participant). Low interaction counts suggest that the material is irrelevant, already known or beyond the ability of the participant to understand.

Quality of interactions can be measured by analysing the breadth of engagement by a participant. This could be used to identify the intent of the participant as low-quality engagement focussed on information gathering would suggest that the participant is interested only in acquiring new information (e.g. by teachers for classroom use) rather than engaging in deeper learning to gain certification. For a participant engaging only at the level of information gathering, measuring success through grades, discussions or posts would not be appropriate.

Content type can be evaluated through interactions with the various classes of content provided in the e learning offering. E learning has a very broad spectrum of content delivery modes including, text, video, blogs, tests, forums, posts, collaborative sharing, and interactive communications channels such as skype. This is a relatively easy factor to measure by identifying the various content types and counting the number of interactions for each.

Results of measuring these three factors can then be analysed against outcomes to investigate whether there is any statistical correlation between the factors (individually or together) and a successful outcome. An identified correlation would provide evidence to support the use of the factors in creating a predictive model to measure success in e learning and could influence better future design of e learning offerings for a particular participant group. Given that some content types have a far greater cost of production, for providers it is important to understand the impact of each content type so that the greatest return on investment can be realised.

The focus of the work presented in this paper is on testing the published success factors to determine if there is any evidence to support that they can facilitate or influence a successful outcome in e learning.

7.2.3. Methodology of the pilot study (Wilden, Shillabeer & deVries 2017)

Several datasets are available to support the discussed work. For initial factor testing the Harvard/MIT online course dataset was selected as it contains all required data attributes. The data is fully documented, not aggregated and from the MITx and HarvardX (2014) Dataverse which is a reputable source. The selected dataset contains raw data for two years of open online courses from fall 2012 to summer 2014. The dataset contains 20 variables for 641,138 rows of data and includes enrolments in science and non-science focussed courses by both local and international participants with all levels of educational background. The full data description can be downloaded from the associated website (Ho et al. 2014). The dataset includes a broad cross section of populations and study interests.

The following four steps were undertaken to complete the analysis.

1. Relevant data attributes were selected to facilitate testing of each factor (Table 63). The dataset for analysis was reduced from 20 attributes to the following 3 sub population attributes and 4 interaction attributes, no rows were removed during this step.

    - Sub population Identification.
        - viewed Identifying participants who viewed the course.
        - explored Identifying participants who accessed at least half of the course content.
        - certified Identifying participants who earned a certificate.
    - Interactions (Quantity/Quality/Type).
        - nevents Total number of interactions between the participant and the course.
        - ndays_act Total number of unique days that the student interacted with the course.
        - nplay_video Total number of video play events recorded for the participant.
        - pchapters Total percentage of chapters viewed by the participant for a course.

Table 63 - Attributes and values selected for analysis of each success factor (Wilden, Shillabeer & deVries 2017)

Factor	Attributes	Values
Quantity	Certified	Categorical "1" or "0"
	Number of days active	Continuous 1 205
	Chapters viewed	Continuous 1 100
Quality	Certified	Categorical "1" or "0"
	Number of days active	Continuous 1 205
	Chapters viewed	Continuous 1 100
	Video Plays	Continuous 1 98,517
Type	Certified	Categorical "1" or "0"
	Chapters viewed	Continuous 1 100
	Video Plays	Continuous 1 98,517

2. The data was cleaned to remove erroneous values and ensure consistency. The data owners had implemented a de identification process prior to releasing the dataset. Rows that had unique combinations of quasi-identifiers were deleted. Deleted rows included unique or potentially identifiable combinations of number of forum posts, course ID, gender, year of birth country, start date, last days active, and number of days active (Ho et al. 2014).
3. After selecting appropriate data for the analysis, WEKA 3.8 data mining software (Frank, Hall & Witten 2016) was used for association rule mining to identify patterns between the focus attributes for each factor and a successful outcome as measured by certification status. Certification was selected as the success measure as this was the intended outcome by the provider of the dataset. For all data mining runs, support was set at 0.05 and confidence at 0.75.
4. Results of the analysis were graphed for presentation using both Microsoft Excel (Microsoft 2016) and Weka.

7.2.4. Results of pilot study (Wilden, Shillabeer & deVries 2017)

Quantity

The results shown in Table 64 suggest a positive association with certification based on quantity of interactions, with number of days active and percentage of chapters viewed showing the strongest relationship to success.

Table 64 - Quantity vs success (Wilden, Shillabeer & deVries 2017)

	Certified			
	r	**r^2**	**t**	**p**
nevents	0.630	0.397	440.687	< 0.05
ndays_act	0.697	0.485	527.295	< 0.05
nplay_video	0.320	0.103	183.615	< 0.05
pchapters	0.661	0.437	478.238	< 0.05
nforum_posts	0.131	0.017	71.975	< 0.05

Figure 77 shows the certification rate decreases rapidly after the first few interaction count increments. Beyond 22,000 interactions the analysis showed only 3 additional certified participants per 100 extra interactions and a decreasing total marginal benefit with increasing interactions. An 80% certification rate occurs between 8,800 and 20,000 interactions. Success rates do not increase beyond this point regardless of the number of extra interactions.

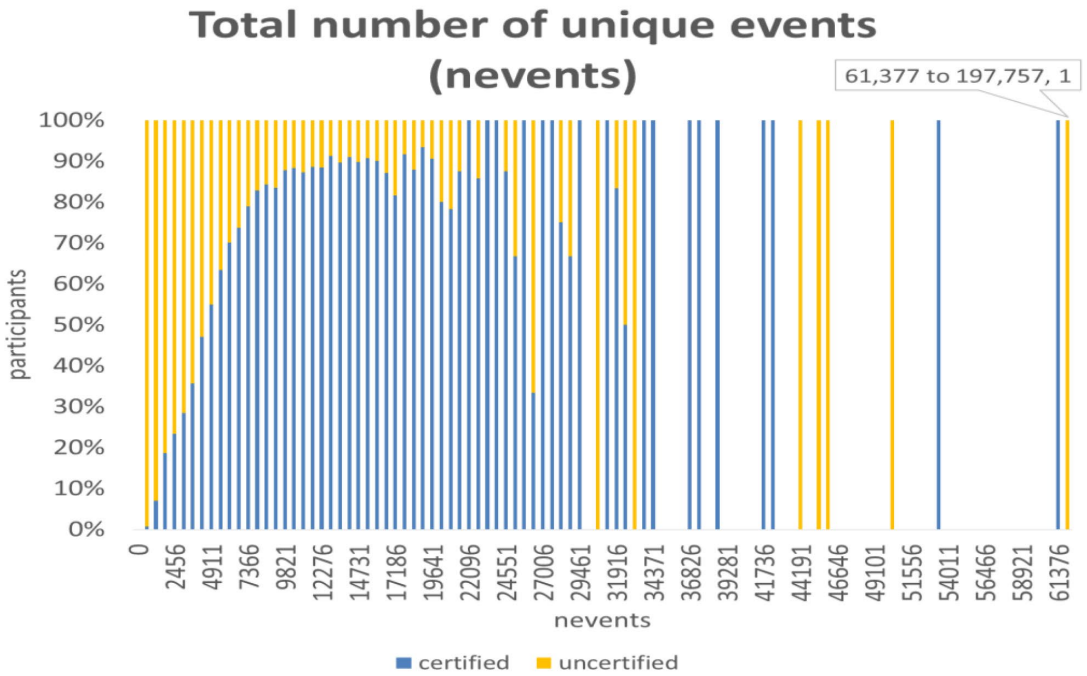

Figure 77 - Effect on success of number of interactions (Wilden, Shillabeer & deVries 2017)

These findings are contrary to published research which suggests that there is a continuous positive relationship between number of interactions and improved success. Whilst the results are only applicable to the dataset used, it does suggest that there can be defined thresholds of interaction (effort required) for optimal success in e learning courses. Requiring students to engage with materials beyond the identified threshold is unnecessary and may lead to participant disenfranchisement and lethargy.

Number of days active

Number of days active has a similar impact on success as compared to the number of interactions. Figure 78 shows that overall, a larger number of days active generally results in a greater success rate although there is again an upper threshold. The greatest positive impact on outcomes is seen between 50 to 150 active days. A greater number of interactions across a higher number of days is generally associated with greater success but it is not absolute. It is not sufficient to simply log on each day, there must also be active engagement with materials and learning opportunities over a defined time-period. Overlaying Figure 77 and Figure 78 reveal that the upper threshold for success is shown to be approximately 150 days or 5 months and 30,000 interactions. This substantiates traditional higher education learning cycles.

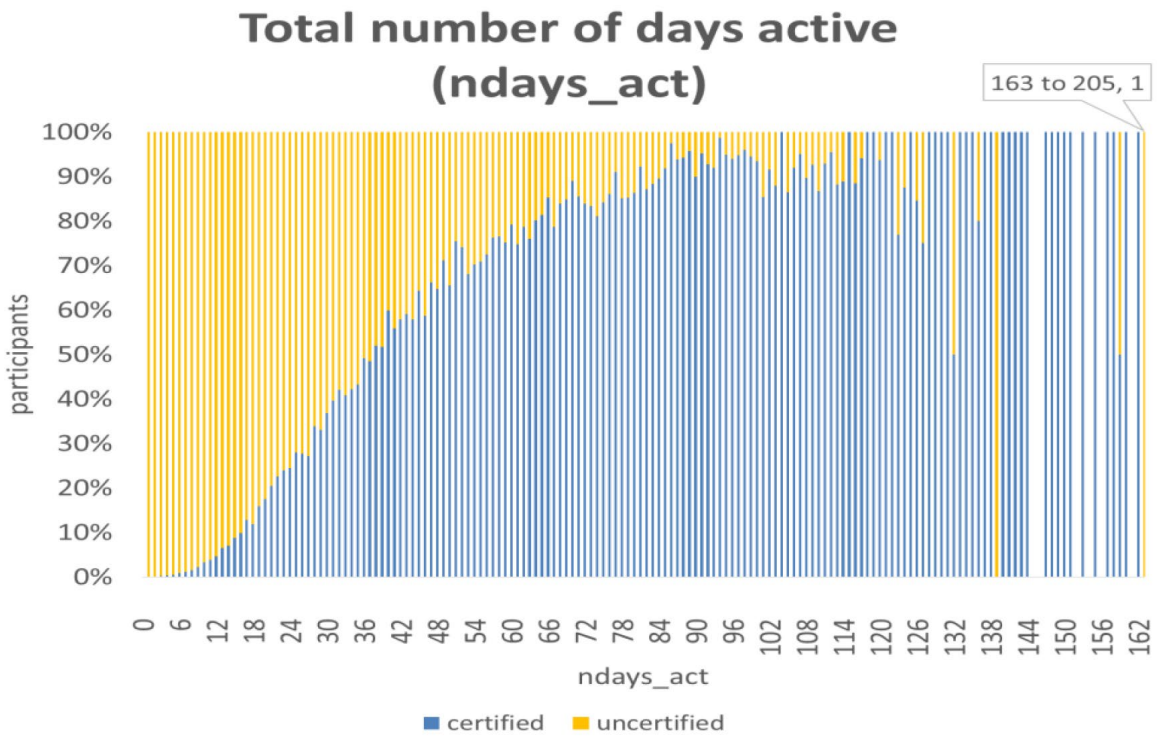

Figure 78 - Effect on success of number of days engaged (Wilden, Shillabeer & deVries 2017)

Number of video play events

Video play events do not show an association with certification. Figure 79 shows there is very little relationship between the success rate and number of video play events. The number of video plays was very similar between participants who achieved a successful outcome and those who did not. The total number of events suggests that a greater number of total interactions can influence a positive outcome, but this is not dependent upon the number of video plays alone. 85.99% of participants achieving a successful outcome, have 900 or fewer video plays during active course duration. Success rate plateaus beyond 900 views and suggests that increased video plays do not contribute to increased chance of success. Whilst video is a key ingredient in most e learning delivery the analysis presented here does not substantiate the investment required to develop such materials for purpose.

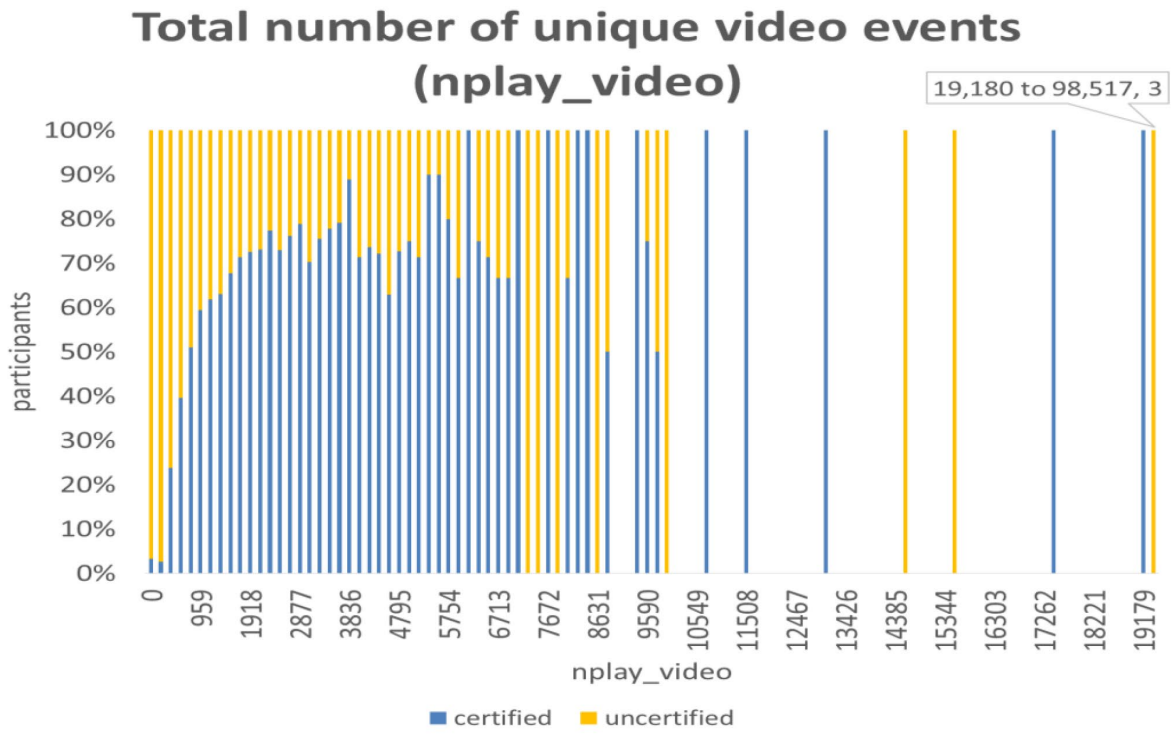

Figure 79 - Effect on success of number of videos played (Wilden, Shillabeer & deVries 2017)

Chapters viewed

Chapters viewed had the greatest statistical effect on success when compared to all other attributes, the 'explored' attribute (participants who viewed more than a set threshold of the topic), was shown to have a correlation with certification with a confidence factor of 0.89. Figure 80 shows that certification was only likely by participants viewing over 50% of chapters, suggesting a strong positive correlation between chapter interaction and success. There was however an equally likelihood of failure even when 100% of chapters were viewed (as seen in Figure 80) but this could reflect the broad intention of participants with some using the platform for information retrieval only. Further investigation is required to determine if those who intended to certify were unsuccessful or if only those who did not aim for this outcome were unsuccessful.

Results show a slow increase in certification up until approximately 75% of chapters viewed, and then a steep increase up towards 100% chapters viewed. Unlike the previous attribute results, most successful outcomes occur at the higher end of the interaction scale, with the success rate continuing to increase up to the maximum threshold for the attribute.

Only 18.02% of certified participants viewed 75% or less of the course content. It is not until participants view 85% of the chapters that the cumulative certification total increases to 34.26% and until 98% of chapters viewed that the cumulative certification total reaches 77.19%. Clearly the greater the percentage of chapters viewed, the greater chance of a successful outcome.

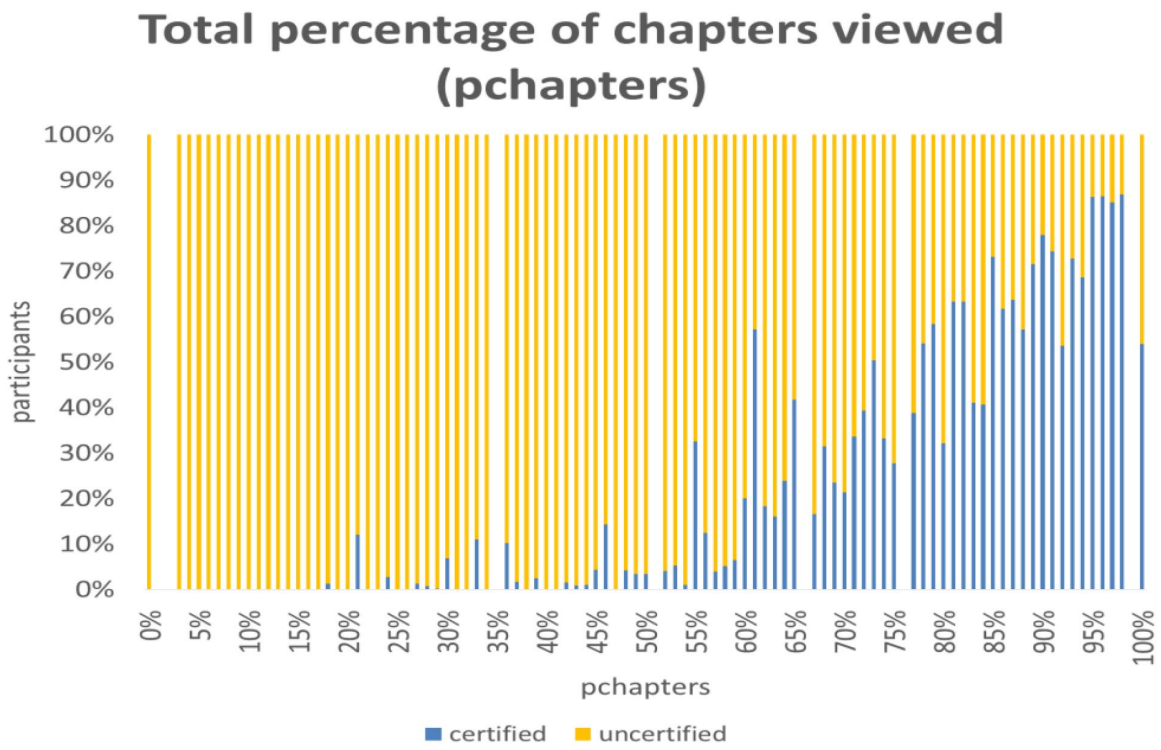

Figure 80 - Effect on success of number of chapters viewed (Wilden, Shillabeer & deVries 2017)

7.2.5. Conclusions of pilot study (Wilden, Shillabeer & deVries 2017)

Three key success factors were identified in the literature:

1. Quantity of participant interactions with course content
2. Quality of participant interaction with course content
3. Type of interaction with course content.

Analysis of these factors using broad spectrum data collected from 2 years of online course offerings by Harvard and MIT Universities revealed some unexpected results. Two of the reported key success factors showed little association with a successful outcome.

The quantity of interactions was correlated with a successful outcome but quickly reached a plateau which evidenced that there is an effort threshold for success and even when increasing interaction significantly there was minimal or no associated improvement in outcomes.

Courses offering certification as a reward for high quality learning interaction were not more likely to measure a successful outcome across participants. This again disputes the published success criteria. Findings presented here suggest that setting or encouraging high participation requirements would not be an incentive to engage or a predictor of success.

Of the published success factors, only Type showed evidence of being correlated. It was interesting to note that the new multimedia delivery methods most integrated into E-Learning courses had the least influence on successful outcomes. Only the traditional information delivery method of book chapters was shown to have a positive correlation with potential for a successful outcome. This suggests that traditional teaching materials should not be overlooked when designing e learning courses. Whilst no single multimedia type was correlated with positive outcomes the overall interaction rate was correlated. This suggests that providing a range of modes for interacting with course content presents a positive impact on the chance of success, but none present a definitive success factor in isolation.

7.2.6. Additional information since pilot study

Section 7.2.2 discusses the predictors of E-Learning success, emphasising interactions, engagement, and content type as key factors. Recent research by Qi, Zhang, et al. (2022), further supports these points by demonstrating a significant positive correlation between learners' E-Learning behaviour and learning outcomes.

This newer research not only confirms the importance of quality and quantity of interactions but also incorporates the use of machine learning to analyse and predict E-Learning success based on specific behaviours and engagement patterns, offering a more nuanced understanding of the factors contributing to E-Learning success, which plays a large part in the current research.

The intention of this pilot research was to either prove or disprove current (at the time), thoughts on the efficacy of multimedia, and other types of interactions in an LMS. While limited to the scope of a MOOC from the MITx and HarvardX (2014) dataset, results from and insights gained from the pilot study would influence the course of this research.

7.2.7. Reflection on methodological approach

The methodology described for the pilot study in Section 7.2.3 has several strengths, such as leveraging a large, well-documented dataset from a reputable source, and a clear, structured approach to data analysis. However, there are areas where critiques and recommendations would have enhanced the methodology. These critiques were critical in the evaluation of the current research methodology.

The process of data cleaning and attribute selection is crucial for the integrity of the analysis, and therefore a level of transparency is necessary when reporting findings. A more detailed description of these processes would increase transparency and allow for replicability by other researchers. This was especially important with the current research, given the nature of the dataset, and care was taken to meticulously detail every step taken, for both transparency purposes, but for completeness as well.

The pilot study classifies success primarily through certification status, which was primarily a limitation on the availability of additional grade information from the dataset. While this is a clear and measurable outcome, E-Learning success is generally more complicated. This was taken into consideration with regards to this study, where additional information on student outcomes were requested, and ultimately led to a much richer analysis of the dataset, when compared to the pilot study.

The lack of more informative machine learning algorithms, or more advanced explorative data analysis made the pilot study severely lacking in detail and complexity. Only basic analysis was performed, and more advanced software and algorithms were not utilised. While this was to be expected for an initial pilot study to justify the potential of the initial research questions (of can success be predicted in E-Learning); the methodology should have been significantly bolstered. These critiques were addressed with this research, with a large array of techniques, and software solutions utilised.

7.2.8. Analysis of results

The results from the pilot study, while limited, provide several insights into the predictors of E-Learning success, such as the importance of interaction quantity and quality. Notably, the findings suggest diminishing returns on learner engagement beyond certain thresholds, challenging the assumption that more interactions always lead to better outcomes. The methodology uses relatively simple statistical analysis to uncover these relationships, therefore, the methodology of this research intended to make sure these limitations were addressed.

The pilot study focuses on exploring predictors of E-Learning success through the analysis of MITx and HarvardX (2014) student performance. As mentioned, the methodologies and statistical analyses used in the pilot study are basic and primarily exploratory, investigating the relationship between various forms of student interaction with course content (quantity, quality, and type) and learning success, measured by certification status.

Comparing the pilot study and the analysis performed in this research, reveals several areas where the pilot study could have been enhanced or expanded to better account for error and provide more nuanced insights into E-Learning success factors.

As discussed in the upcoming Section 3.4 Exploratory data analysis, there is great importance in the use of exploratory data analysis in uncovering patterns, trends, and anomalies in data before formal hypothesis testing. The pilot study's methodology could have benefited from a more thorough exploratory analysis phase to identify underlying structures or potential biases in the data. For instance, visualisations and descriptive statistics could have provided deeper insights into the distribution of interactions, video plays, and chapters viewed, which might have influenced the design of subsequent analyses.

The pilot study also primarily relies on correlation analyses to explore the relationship between interaction metrics and success. While this technique is useful, this approach could be supplemented with the more advanced statistical tests mentioned in Section 3.4 Exploratory data analysis, such as Pearson's (1900) Chi-squared test for categorical data or the Kruskal-Wallis (1952) test for comparing medians across multiple groups. These tests could offer more detailed insights into the independence of variables and differences in interaction patterns among successful and unsuccessful students.

The pilot study does not explicitly address the distribution of the data, which can significantly impact the choice of statistical tests. As mentioned in Section 3.4 Exploratory data analysis, the Anderson-Darling (1952) test for assessing normality, would have been appropriate for this purpose. Applying this test could have helped determine the most appropriate statistical techniques for the pilot study's data, potentially suggesting the use of non-parametric tests where normality assumptions are violated.

With multiple comparisons being made, the pilot study could have controlled for the risk of Type I errors (false positives) through techniques such as the Bonferroni (1936) correction, as discussed in Section 3.4 Exploratory data analysis. This would ensure the integrity and reliability of the findings, particularly when making multiple pairwise comparisons between interaction metrics and success outcomes.

Section 3.4 Exploratory data analysis, outlines the comprehensive approach this research undertook regarding predictive analytics, with statistical testing to inform predictive modelling. The pilot study's methodology could have been expanded to include predictive models that account for the complex, nature of E-Learning success. Techniques such as regression analysis (for continuous outcomes) or logistic regression (for binary outcomes like certification status) could provide more detailed predictions of success based on interaction metrics.

The pilot study's analysis could also have ben deepened by incorporating multivariate analysis techniques mentioned in Section 3.4 Exploratory data analysis. This would allow for the exploration of how multiple factors interact to influence E-Learning success, moving beyond simple bivariate correlations to more complex models that can account for interactions among variables.

Some of the potential consequences of not applying the stated testing methodologies, include the misinterpretation of data, inaccurate predictions, and invalid conclusions. Ultimately, this greatly affects the reliability of research findings, which rely on the appropriate selection and application of statistical tests. Failing to use the correct tests can invalidate the research conclusions, undermining the study's contribution to the field.

In summary, while the pilot study provides valuable initial insights into predictors of E-Learning success, its methodology and analytical approaches could be significantly enhanced by incorporating the statistical techniques and exploratory analysis strategies outlined in Section 3.4 Exploratory data analysis. This would not only strengthen the reliability of the findings but also provide a more detailed understanding of the factors contributing to E-Learning success.

7.2.9. Critique of pilot study

The conclusions of the pilot study challenge common beliefs about E-Learning success factors, as mentioned in Section 7.2.2, and discussed in work by Mayer (2009) on multimedia learning. The pilot study suggests that quantity and quality of interactions are not as strongly correlated with success as previously thought. With traditional content delivery, such as reading book chapters, being found to have a positive correlation with success, suggesting the importance of integrating traditional teaching materials into E-Learning courses. This contrasts with current trends emphasising multimedia and interactive content. However, the pilot study was relatively lacking in detail, having access only to public MOOC data, that itself was limited in outcome information, and richer interaction information, and utilising only basic methods of analysis. This positively affected the current research, with the foreknowledge of what would be necessary, and what additional information should be retrieved.

7.2.10. Importance of pilot study on this research

This pilot study was included within this research as both a record of progress, and for the reasoning behind crucial decisions that made this study as robust and detailed as it is. Documenting these crucial initial steps shows that while mistakes were made (in lack of analysis, and of not utilising potential resources), these mistakes can help to form the basis for better research, and for a level of transparency with what came before, and influenced decisions regarding the objectives, and decisions made.

7.3. Appendix C: MATLAB script for PCA, clustering and figures

```
%% Import Data
opts = delimitedTextImportOptions("NumVariables", 138);
opts.DataLines = [2, Inf];
opts.Delimiter = ",";
opts.VariableNames = ["grade", "days_active", "days_in_topic", "topic_videos", "topic_quiz",
"topic_assignment", "topic_forum", "topic_participation", "topic_support", "topic_other",
"video_percent", "quiz_percent", "assignment_percent", "topic_forum_usage",
"participation_percent", "support_percent", "other_percent", "AvgSecBetweenActions",
"Interactions", "DistinctLecture", "DistinctQuiz", "DistinctAssign", "ForumActivity",
"DistinctParticipation", "DistinctSupportMaterials", "DistinctOtherInteractionMaterials",
"Assignment_interactions", "Video_interactions", "Support_material_interactions",
"Activity_interactions", "Social_interactions", "Course_participation_interactions",
"Other_interactions", "total_period_1", "total_period_2", "total_period_3",
"total_period_4", "total_period_5", "total_period_6", "total_period_7", "total_period_8",
"assignment_period_1", "video_period_1", "support_period_1", "activity_period_1",
"social_period_1", "participation_period_1", "other_period_1", "assignment_period_2",
"video_period_2", "support_period_2", "activity_period_2", "social_period_2",
"participation_period_2", "other_period_2", "assignment_period_3", "video_period_3",
"support_period_3", "activity_period_3", "social_period_3", "participation_period_3",
"other_period_3", "assignment_period_4", "video_period_4", "support_period_4",
"activity_period_4", "social_period_4", "participation_period_4", "other_period_4",
"assignment_period_5", "video_period_5", "support_period_5", "activity_period_5",
"social_period_5", "participation_period_5", "other_period_5", "assignment_period_6",
"video_period_6", "support_period_6", "activity_period_6", "social_period_6",
"participation_period_6", "other_period_6", "assignment_period_7", "video_period_7",
"support_period_7", "activity_period_7", "social_period_7", "participation_period_7",
"other_period_7", "assignment_period_8", "video_period_8", "support_period_8",
"activity_period_8", "social_period_8", "participation_period_8", "other_period_8",
"total_sem_period_1", "total_sem_period_2", "total_sem_period_3",
"total_sem_period_4", "total_sem_period_5", "assignment_sem_period_1",
"video_sem_period_1", "support_sem_period_1", "activity_sem_period_1",
"social_sem_period_1", "participation_sem_period_1", "other_sem_period_1",
"assignment_sem_period_2", "video_sem_period_2", "support_sem_period_2",
"activity_sem_period_2", "social_sem_period_2", "participation_sem_period_2",
"other_sem_period_2", "assignment_sem_period_3", "video_sem_period_3",
"support_sem_period_3", "activity_sem_period_3", "social_sem_period_3",
"participation_sem_period_3", "other_sem_period_3", "assignment_sem_period_4",
"video_sem_period_4", "support_sem_period_4", "activity_sem_period_4",
"social_sem_period_4", "participation_sem_period_4", "other_sem_period_4",
"assignment_sem_period_5", "video_sem_period_5", "support_sem_period_5",
"activity_sem_period_5", "social_sem_period_5", "participation_sem_period_5",
"other_sem_period_5", "college"];
opts.VariableTypes = ["categorical", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
```

```
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "double", "double", "double", "double",
"double", "double", "double", "double", "double", "categorical"];
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";
opts = setvaropts(opts, ["grade", "college"], "EmptyFieldRule", "auto");
dat = readtable("R:\\DatasetAccess\\ processed.csv", opts);
clear opts

%% Data Preparation
[x, grades, col, labels] = prepareData(dat);

%% PCA Analysis
[U, S, V, x_standardised] = standardiseAndSVD(x);
scores = calculateScores(U, S);
[val, index1] = sort(-V(:,1).*V(:,1));
vi_1 = [val, index1];
[val, index2] = sort(-V(:,2).*V(:,2));
vi_2 = [val, index2];
[val, index3] = sort(-V(:,3).*V(:,3));
vi_3 = [val, index3];
% limit of top 20 (+ve and -ve) loading components
lim = 20;

%% Visualise components
[l1] = plotTopLoading(V, 1, labels, lim);
[l2] = plotTopLoading(V, 2, labels, lim);
[l3] = plotTopLoading(V, 3, labels, lim);
[l4] = plotTopLoading(V, 4, labels, lim);
[l5] = plotTopLoading(V, 5, labels, lim);
[l6] = plotTopLoading(V, 6, labels, lim);
[l7] = plotTopLoading(V, 7, labels, lim);
[l8] = plotTopLoading(V, 8, labels, lim);
[l9] = plotTopLoading(V, 9, labels, lim);
[l10] = plotTopLoading(V, 10, labels, lim);
```

```
%% Visualisation
% PCA/LSA Results
plotScreePlot(S);
plotCumulativeVariance(S);
%%
% components 1 and 2
plotBiPlot(U, S, V, labels, 1, 2, [index1(1:lim) index2(1:lim)]);
% components 1 and 3
plotBiPlot(U, S, V, labels, 1, 3, [index1(1:lim) index3(1:lim)]);
% components 1 and 4
plotBiPlot(U, S, V, labels, 1, 4, [index1(1:lim) index4(1:lim)]);
% components 2 and 3
plotBiPlot(U, S, V, labels, 2, 3, [index2(1:lim) index3(1:lim)]);
% components 2 and 4
plotBiPlot(U, S, V, labels, 2, 4, [index2(1:lim) index4(1:lim)]);
% components 3 and 4
plotBiPlot(U, S, V, labels, 3, 4, [index3(1:lim) index4(1:lim)]);

%% Function Definitions
function [x, grades, col, labels] = prepareData(dat)
    x = table2array(dat(:, 2:137));
    grades = table2array(dat(:, 1));
    col = table2array(dat(:, 138));
    labels = dat.Properties.VariableNames(2:137);
end
function [U, S, V, x_standardised] = standardiseAndSVD(x)
    means = mean(x);
    stds = std(x);
    x_standardised = (x - means) ./ stds;
    [U, S, V] = svd(x_standardised, 'econ');
end
function scores = calculateScores(U, S)
    scores = U * S;
end
function plotScreePlot(S)
    figure;
    plot(diag(S).^2 / sum(diag(S).^2), '-o');
    title('Scree plot');
    xlabel('Principal component');
    ylabel('Variance explained (%)');
end
function plotCumulativeVariance(S)
    figure;
    plot(cumsum(diag(S).^2) / sum(diag(S).^2), '-o');
    title('Cumulative variance explained by principal components');
    xlabel('Number of principal components');
    ylabel('Cumulative variance explained (%)');
```

```
end
function [varListWithLoadings] = plotTopLoading(V, componentNumber, variableNames, lim)
    figure;
    % Replace underscores with spaces and capitalise
    processedVariableNames = cellfun(@(name) ...
        [upper(name(1)), lower(strrep(name(2:end), '_', ' '))], ...
        variableNames, 'UniformOutput', false);
    loadings = V(:, componentNumber);
    [~, sortOrder] = sort(abs(loadings), 'descend');
    % Select the indices of the top 'lim' loadings
    topVariables = sortOrder(1:lim);
    % Get the corresponding processed variable names and actual loading values
    topVariableNames = processedVariableNames(topVariables);
    actualTopLoadings = loadings(topVariables);
    bar(actualTopLoadings);
    xlabel('Variables');
    ylabel(['Loadings on PC', num2str(componentNumber)]);
    title(['Top ', num2str(lim), ' Loadings for PC', num2str(componentNumber)]);
    set(gca, 'xtick', 1:lim, 'xticklabel', topVariableNames);
    xtickangle(45);
    grid on;
    % Ensure both topVariableNames and actualTopLoadings are column vectors
    topVariableNames = topVariableNames(:);
    actualTopLoadings = actualTopLoadings(:);
    % Combine names and loadings into a cell array
    varListWithLoadings = [topVariableNames, num2cell(actualTopLoadings)];
end
function plotBiPlot(U, S, V, labels, x, y, ind)
    figure;
    % LSA application for better visualisation of components
    L = sqrt(S);
    scores = U(ind, [x, y]) * L([x, y], [x, y]);
    loadings = V(ind, [x, y]) * L([x, y], [x, y]);
    modLabels = arrayfun(@num2str, 1:length(labels), 'UniformOutput', false);
    hBiplot = biplot(loadings, 'Scores', scores, 'VarLabels', modLabels(ind));
    title(sprintf('Biplot of principal components %d and %d', x, y));
    % Find all text objects in the biplot
    hText = findall(hBiplot, 'Type', 'text');
    % Set the font size of the text objects
    set(hText, 'FontSize', 12);
end
```

7.4. Appendix D: R script for exploratory statistics and figures

```
## Libraries
if (!require(farff)) install.packages("farff")
library("farff")
if (!require(ggplot2)) install.packages("ggplot2")
library("ggplot2")
if (!require(forcats)) install.packages("forcats")
library("forcats")
if (!require(tidyr)) install.packages("tidyr")
library("tidyr")
if (!require(dplyr)) install.packages("dplyr")
library("dplyr")
if (!require(data.table)) install.packages("data.table")
library("data.table")
if (!require(reshape2)) install.packages("reshape2")
library("reshape2")
if (!require(nortest)) install.packages("nortest")
library("nortest")
if (!require(dunn.test)) install.packages("dunn.test")
library("dunn.test")
if (!require(lsr)) install.packages("lsr")
library("lsr")
if (!require(knitr)) install.packages("knitr")
library("knitr")
if (!require(MASS)) install.packages("MASS")
library("MASS")
if (!require(stats)) install.packages("stats")
library("stats")
if (!require(multcompView)) install.packages("multcompView")
library("multcompView")
if (!require(broom)) install.packages("broom")
library("broom")
if (!require(effsize)) install.packages("effsize")
library("effsize")
if (!require(scales)) install.packages("scales")
library("scales")
if (!require(rms)) install.packages("rms")
library("rms")
if (!require(VennDiagram)) install.packages("VennDiagram")
library("VennDiagram")
if (!require(grid)) install.packages("grid")
library("grid")
## Libraries end
## Files
path = "R:\\DatasetAccess\\processed.csv"
```

```
path2 = "R:\\ SQL\\topic_dat.csv"
out <- read.csv(path)
topicdata <- read.csv(path2)
writePath = "R: \\R_files\\"
#newout <- out[ -c(1,3:4,6:7) ]
#write.csv(newout, "R:\\ R_files \\newout.csv", row.names=TRUE)
## Files end
## Data preparation
# grade to factor
out$grade <- as.factor(out$grade)
out$grade <- factor(out$grade, levels = c("F","P","CR","DN","HD"))
# remove '0' colleges identified (not part of any college)
out <- out[(out$college>=1) ,]
topicdata <- topicdata[(topicdata$college>=1) ,]
# numerical to character (college)
out$college <- as.character(out$college)
out$college[out$college == '1'] <- 'BGL'
out$college[out$college == '2'] <- 'EPS'
out$college[out$college == '3'] <- 'HAS'
out$college[out$college == '4'] <- 'MPH'
out$college[out$college == '5'] <- 'NHS'
out$college[out$college == '6'] <- 'S&E'
topicdata$college <- as.character(topicdata$college)
topicdata$college[topicdata$college == '1'] <- 'BGL'
topicdata$college[topicdata$college == '2'] <- 'EPS'
topicdata$college[topicdata$college == '3'] <- 'HAS'
topicdata$college[topicdata$college == '4'] <- 'MPH'
topicdata$college[topicdata$college == '5'] <- 'NHS'
topicdata$college[topicdata$college == '6'] <- 'S&E'
# numerical to character (pass_fail)
out$pass_fail <- as.character(out$pass_fail)
out$pass_fail[out$pass_fail == 0] <- 'F'
out$pass_fail[out$pass_fail == 1] <- 'P'
# convert college character to factors
out$college <- as.factor(out$college)
topicdata$college <- as.factor(topicdata$college)
# convert pass_fail character to factors
out$pass_fail <- as.factor(out$pass_fail)
out$pass_fail <- factor(out$pass_fail, levels = c("F","P"))
# trim topic codes to just topic area (letters)
out$topic_shortcode <- substr(out$topic_shortcode, 1, 4)
topicdata$topic_shortcode <- substr(topicdata$topic_shortcode, 1, 4)
# convert topic code character to factors
out$topic_shortcode <- as.factor(out$topic_shortcode)
topicdata$topic_shortcode <- as.factor(topicdata$topic_shortcode)
# Mapping grades to numerical values for correlation analysis
grade_mapping <- c('HD' = 5, 'DN' = 4, 'CR' = 3, 'P' = 2, 'F' = 1)
```

```
out$grade_numeric <- as.numeric(recode(out$grade, !!!grade_mapping))
#
out_summary1 <- out %>%
  dplyr::group_by(college, grade) %>%
  dplyr::summarize(across(starts_with("total_period"), list(
    max = ~ max(.x, na.rm = TRUE),
    mean = ~ mean(.x, na.rm = TRUE),
    median = ~ median(.x, na.rm = TRUE)
  ), .names = "{.col}_{.fn}"))
out_summary_long1 <- out_summary1 %>%
  pivot_longer(
    cols = -c(college, grade),
    names_to = "time_period_stat",
    values_to = "interactions"
  )
out_summary_long1 <- out_summary_long1 %>%
  tidyr::extract(
    col = time_period_stat,
    into = c("time_period", "stat"),
    regex = "(total_period_[1-8]+)_(max|mean|median)"
  )
out_summary_long1 <- out_summary_long1 %>%
  pivot_wider(
    id_cols = c(college, grade, time_period),
    names_from = stat,
    values_from = interactions
  )
#
out_summary_long1$time_period <- factor(out_summary_long1$time_period, levels =
unique(out_summary_long1$time_period))
time_agg <- aggregate(mean ~ time_period + grade, data = out_summary_long1, mean)
time_wide <- reshape2::dcast(time_agg, time_period ~ grade, value.var = "mean")
time_long <- melt(time_wide, id.vars = "time_period", variable.name = "grade", value.name
= "mean")
#
max_y_1 <- max(out_summary_long1$mean, na.rm = TRUE)
out_summary2 <- out %>%
  dplyr::group_by(college, grade) %>%
  dplyr::summarize(across(starts_with("total_sem_period"), list(
    max = ~ max(.x, na.rm = TRUE),
    mean = ~ mean(.x, na.rm = TRUE),
    median = ~ median(.x, na.rm = TRUE)
  ), .names = "{.col}_{.fn}"))
out_summary_long2 <- out_summary2 %>%
  pivot_longer(
    cols = -c(college, grade),
    names_to = "sem_period_stat",
```

```
   values_to = "interactions"
  )
out_summary_long2 <- out_summary_long2 %>%
  tidyr::extract(
   col = sem_period_stat,
   into = c("sem_period", "stat"),
   regex = "(total_sem_period_[1-8]+)_(max|mean|median)"
  )
out_summary_long2 <- out_summary_long2 %>%
  pivot_wider(
   id_cols = c(college, grade, sem_period),
   names_from = stat,
   values_from = interactions
  )
#
out_summary_long2$sem_period <- factor(out_summary_long2$sem_period, levels =
unique(out_summary_long2$sem_period))
sem_agg <- aggregate(mean ~ sem_period + grade, data = out_summary_long2, mean)
sem_wide <- reshape2::dcast(sem_agg, sem_period ~ grade, value.var = "mean")
sem_long <- melt(sem_wide, id.vars = "sem_period", variable.name = "grade", value.name =
"mean")
#
max_y_2 <- max(out_summary_long2$mean, na.rm = TRUE)
#
percent_columns <-
c('days_active','video_percent','quiz_percent','assignment_percent','topic_forum_usage',
'participation_percent','support_percent','other_percent', 'grade_numeric')
correlation_matrix <- cor(out[, percent_columns], use = "complete.obs")
melted_correlation_matrix <- melt(correlation_matrix, varnames = c('Variable1', 'Variable2'))
## Data preparation end
### Exploratory data analysis
# Chi-Square Test
grade_college_table <- table(out$college, out$grade)
chisq_test_college <- chisq.test(grade_college_table)
#print(chisq_test_college)
chisq_test_college <- chisq.test(grade_college_table)
expected <- chisq_test_college$expected
observed <- as.matrix(grade_college_table)
comparison <- data.frame(observed = as.vector(observed),
                         expected = as.vector(expected),
                         college = rep(colnames(grade_college_table), nrow(grade_college_table)),
                         grade = rep(rownames(grade_college_table), each =
ncol(grade_college_table)))
#print(comparison)
# Generate all pairwise combinations of colleges
colleges <- unique(out$college)
college_pairs <- combn(colleges, 2, simplify = FALSE)
```

```
# Chi-squared tests of college pairs
results <- lapply(college_pairs, function(pair) {
  college1 <- out[out$college == pair[1], ]
  college2 <- out[out$college == pair[2], ]
  combined <- rbind(college1, college2)
  test_result <- chisq.test(table(combined$college, combined$grade))
  # Return a list with the test results and the pair of colleges
  list(pair = pair, p_value = test_result$p.value, test_result = test_result)
})
# Bonferroni correction
p_values <- sapply(results, function(x) x$p_value)
adjusted_p_values <- p.adjust(p_values, method = "bonferroni")
# Combine results with adjusted p-values
final_results <- lapply(seq_along(results), function(i) {
  list(
    pair = results[[i]]$pair,
    original_p_value = results[[i]]$p_value,
    adjusted_p_value = adjusted_p_values[i],
    test_result = results[[i]]$test_result
  )
})
#print(final_results)
## Days Active
max_days <- max(out$days_active, na.rm = TRUE)
out$activity_level <- cut(out$days_active,
                          breaks = c(0, max_days/3, 2*max_days/3, max_days),
                          labels = c("Low", "Medium", "High"),
                          include.lowest = TRUE)
table_data <- table(out$college, out$activity_level)
# Chi-squared test
chi_squared_test <- chisq.test(table_data)
#print(chi_squared_test)
residuals <- chi_squared_test$residuals
n <- sum(table_data)
adjusted_residuals <- residuals / sqrt((1 - rowSums(table_data) / n) * (1 -
colSums(table_data) / n))
adjusted_residuals_df <- as.data.frame(as.table(adjusted_residuals))
names(adjusted_residuals_df) <- c("College", "Activity_Level", "Adjusted_Residual")
# Filter out the residuals that are greater than 2
significant_residuals <- subset(adjusted_residuals_df, abs(Adjusted_Residual) > 2)
#print(significant_residuals)
model <- polr(grade ~ days_active + college, data = out, Hess=TRUE)
#summary(model)
## Days Active
max_days <- max(out$days_active, na.rm = TRUE)
out$activity_level <- cut(out$days_active,
                          breaks = c(0, max_days/3, 2*max_days/3, max_days),
```

```
                labels = c("Low", "Medium", "High"),
                include.lowest = TRUE)
table_data <- table(out$college, out$activity_level)
# Chi-squared test
chi_squared_test <- chisq.test(table_data)
# Check expected counts for Chi-squared test
expected_counts <- chi_squared_test$expected
low_count_cells <- sum(expected_counts < 5)
cat("Number of cells with expected counts < 5:", low_count_cells, "\n")
#print(chi_squared_test)
residuals <- chi_squared_test$residuals
n <- sum(table_data)
adjusted_residuals <- residuals / sqrt((1 - rowSums(table_data) / n) * (1 -
colSums(table_data) / n))
adjusted_residuals_df <- as.data.frame(as.table(adjusted_residuals))
names(adjusted_residuals_df) <- c("College", "Activity_Level", "Adjusted_Residual")
# Filter out the residuals that are greater than 2
significant_residuals <- subset(adjusted_residuals_df, abs(Adjusted_Residual) > 2)
#print(significant_residuals)
model <- polr(grade ~ days_active + college, data = out, Hess=TRUE)
# nominal_test(model)
# summary(model)
## Interactions
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(Interactions)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_Interactions <- kruskal.test(Interactions ~ college, data = out)
# Dunn's test
dunn_test_Interactions <- dunn.test(out$Interactions, out$college, method="bonferroni")
## days_active
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(days_active)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_days_active <- kruskal.test(days_active ~ college, data = out)
# Dunn's test
dunn_test_days_active <- dunn.test(out$days_active, out$college, method="bonferroni")
## video_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(video_percent)$p.value)
```

```
#print(results_normality)
# Kruskal-Wallis test
kruskal_video_percent <- kruskal.test(video_percent ~ college, data = out)
# Dunn's test
dunn_test_video_percent <- dunn.test(out$video_percent, out$college,
method="bonferroni")

## quiz_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(quiz_percent)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_quiz_percent <- kruskal.test(quiz_percent ~ college, data = out)
# Dunn's test
dunn_test_quiz_percent <- dunn.test(out$quiz_percent, out$college, method="bonferroni")
## assignment_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(assignment_percent)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_assignment_percent <- kruskal.test(assignment_percent ~ college, data = out)
# Dunn's test
dunn_test_assignment_percent <- dunn.test(out$assignment_percent, out$college,
method="bonferroni")
## participation_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(participation_percent)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_participation_percent <- kruskal.test(participation_percent ~ college, data = out)
# Dunn's test for post-hoc analysis
dunn_test_participation_percent <- dunn.test(out$participation_percent, out$college,
method="bonferroni")
## support_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(support_percent)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_support_percent <- kruskal.test(support_percent ~ college, data = out)
```

```
# Dunn's test for post-hoc analysis
dunn_test_support_percent <- dunn.test(out$support_percent, out$college,
method="bonferroni")
## other_percent
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(other_percent)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_other_percent <- kruskal.test(other_percent ~ college, data = out)
# Dunn's test
dunn_test_other_percent <- dunn.test(out$other_percent, out$college,
method="bonferroni")
## topic_forum_usage
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(topic_forum_usage)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_topic_forum_usage <- kruskal.test(topic_forum_usage ~ college, data = out)
# Dunn's test
dunn_test_topic_forum_usage <- dunn.test(out$topic_forum_usage, out$college,
method="bonferroni")
## ForumActivity
# Anderson-Darling test
results_normality <- out %>%
  group_by(college) %>%
  summarise(ad_test = ad.test(ForumActivity)$p.value)
#print(results_normality)
# Kruskal-Wallis test
kruskal_ForumActivity <- kruskal.test(ForumActivity ~ college, data = out)
# Dunn's test for post-hoc analysis
dunn_test_ForumActivity <- dunn.test(out$ForumActivity, out$college,
method="bonferroni")
# Function to perform Dunn's test
perform_dunns_test <- function(data, period_column) {
  results <- do.call(rbind, lapply(split(data, data[[period_column]]), function(df) {
    test_result <- dunn.test(df$mean, df$college, method = "bonferroni")
    data.frame(
      Period = unique(df[[period_column]]),
      college_Pair = test_result$comparisons,
      Z = test_result$Z,
      P_Value = test_result$P,
      Adjusted_P_Value = test_result$P.adjusted,
      Significance = test_result$P.adjusted < 0.05
```

```
    )
  }))
  return(results)
}
# Anderson-Darling test
ad_test_time <- ad.test(out_summary_long1$mean)
# Kruskal-Wallis Test
kruskal_time <- kruskal.test(mean ~ college, data = out_summary_long1)
# Dunn's test
dunns_results_time <- perform_dunns_test(out_summary_long1, "time_period")
rownames(dunns_results_time) <- NULL
#print(dunns_results_time[dunns_results_time$Significance=='TRUE',])
## Time
results_time <- data.frame(Period = character(), Pair = character(), Cohen_d = numeric(),
stringsAsFactors = FALSE)
time_periods <- unique(out_summary_long1$time_period)
colleges <- c('BGL', 'EPS', 'HAS', 'MPH', 'NHS', 'S&E')
college_pairs <- combn(colleges, 2, simplify = FALSE)
for (time_period in time_periods) {
  for (pair in college_pairs) {
    data_college1 <- subset(out_summary_long1, college == pair[1] & time_period ==
time_period)
    data_college2 <- subset(out_summary_long1, college == pair[2] & time_period ==
time_period)

    if (nrow(data_college1) > 0 & nrow(data_college2) > 0) {
      d <- cohen.d(data_college1$mean, data_college2$mean)
      results_time <- rbind(results_time, data.frame(Period = as.character(time_period), Pair =
paste(pair, collapse = " - "), Cohen_d = d$estimate))
    }
  }
}
#print(results_time[abs(results_time$Cohen_d)>0.5,])
## Semester
# Anderson-Darling test
ad_test_sem <- ad.test(out_summary_long2$mean)
# Kruskal-Wallis Test
kruskal_test_sem <- kruskal.test(mean ~ college, data = out_summary_long2)
# Dunn's test
dunns_results_semester <- perform_dunns_test(out_summary_long2, "sem_period")
rownames(dunns_results_semester) <- NULL
#print(dunns_results_semester[dunns_results_semester$Significance=='TRUE',])
# Cohen d
results_semester <- data.frame(Period = character(), Pair = character(), Cohen_d =
numeric(), stringsAsFactors = FALSE)
semester_periods <- unique(out_summary_long2$sem_period)
for (sem_period in semester_periods) {
```

```
  for (pair in college_pairs) {
    data_college1 <- subset(out_summary_long2, college == pair[1] & sem_period ==
sem_period)
    data_college2 <- subset(out_summary_long2, college == pair[2] & sem_period ==
sem_period)

    if (nrow(data_college1) > 0 & nrow(data_college2) > 0) {
      d <- cohen.d(data_college1$mean, data_college2$mean)
      results_semester <- rbind(results_semester, data.frame(Period =
as.character(sem_period), Pair = paste(pair, collapse = " - "), Cohen_d = d$estimate))
    }
  }
}
#print(results_semester[abs(results_semester$Cohen_d)>0.5,])
#significant_dunns$Pair <- significant_dunns$College_Pair
#significant_dunns <- subset(significant_dunns, select = -College_Pair)
#combined_results <- merge(significant_dunns, significant_cohens, by = c("Period", "Pair"))
#final_summary <- combined_results[, c("Period", "Pair", "Z", "P_Value",
"Adjusted_P_Value", "Cohen_d")]
#print(final_summary)
## Topic components
# Anderson-Darling test
ad_test_topics_video <- ad.test(topicdata$lecture_videos)
ad_test_topics_quiz <- ad.test(topicdata$quiz_count)
ad_test_topics_assignment <- ad.test(topicdata$assignment_count)
ad_test_topics_forum <- ad.test(topicdata$forum_activity)
ad_test_topics_participation <- ad.test(topicdata$participation_activity_count)
ad_test_topics_support <- ad.test(topicdata$support_matrials_count)
ad_test_topics_other <- ad.test(topicdata$other_interactions_count)
# Kruskal-Wallis Test
kw_test_topics_video <- kruskal.test(lecture_videos ~ college, data = topicdata)
kw_test_topics_quiz <- kruskal.test(quiz_count ~ college, data = topicdata)
kw_test_topics_assignment <- kruskal.test(assignment_count ~ college, data = topicdata)
kw_test_topics_forum <- kruskal.test(forum_activity ~ college, data = topicdata)
kw_test_topics_participation <- kruskal.test(participation_activity_count ~ college, data =
topicdata)
kw_test_topics_support <- kruskal.test(support_matrials_count ~ college, data = topicdata)
kw_test_topics_other <- kruskal.test(other_interactions_count ~ college, data = topicdata)
# Function to perform Dunn's test
perform_dunns_test_topics <- function(data, component_column) {
  data <- na.omit(data[, c(component_column, 'college')])
  test_result <- dunn.test(x=data[[component_column]],
                           g=data[['college']],
                           method="bonferroni")
  result <- data.frame(
    college_Pair = test_result$comparisons,
    Z = test_result$Z,
```

```
    P_Value = test_result$P,
    Adjusted_P_Value = test_result$P.adjusted,
    Significance = test_result$P.adjusted < 0.05
  )
  return(result)
}

# Perform Dunn's test
# Videos
dunns_results_topics_videos <- perform_dunns_test_topics(topicdata, "lecture_videos")
rownames(dunns_results_topics_videos) <- NULL
#print(dunns_results_topics_videos[dunns_results_topics_videos$Significance=='TRUE',])
# Quiz
dunns_results_topics_quiz <- perform_dunns_test_topics(topicdata, "quiz_count")
rownames(dunns_results_topics_quiz) <- NULL
#print(dunns_results_topics_quiz[dunns_results_topics_quiz$Significance=='TRUE',])
# Assignment
dunns_results_topics_assignment <- perform_dunns_test_topics(topicdata,
"assignment_count")
rownames(dunns_results_topics_assignment) <- NULL
#print(dunns_results_topics_assignment[dunns_results_topics_assignment$Significance=='T
RUE',])
# Forum
dunns_results_topics_forum <- perform_dunns_test_topics(topicdata, "forum_activity")
rownames(dunns_results_topics_forum) <- NULL
#print(dunns_results_topics_forum[dunns_results_topics_forum$Significance=='TRUE',])
# Participation
dunns_results_topics_participation <- perform_dunns_test_topics(topicdata,
"participation_activity_count")
rownames(dunns_results_topics_participation) <- NULL
#print(dunns_results_topics_participation[dunns_results_topics_participation$Significance=
='TRUE',])
# Support
dunns_results_topics_support <- perform_dunns_test_topics(topicdata,
"support_matrials_count")
rownames(dunns_results_topics_support) <- NULL
#print(dunns_results_topics_support[dunns_results_topics_support$Significance=='TRUE',])
# Other
dunns_results_topics_other <- perform_dunns_test_topics(topicdata,
"other_interactions_count")
rownames(dunns_results_topics_other) <- NULL
#print(dunns_results_topics_other[dunns_results_topics_other$Significance=='TRUE',])

### Visualisations
## Grades
png(file=paste(writePath, 'heat_corr.png'))
ggplot(melted_correlation_matrix, aes(Variable1, Variable2, fill = value)) +
```

```
  geom_tile() +
  geom_text(aes(label = sprintf("%.2f", value)), color = "black", size = 3) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(fill = "Correlation")
dev.off()
## Grade distribution
png(file=paste(writePath, 'grade.png'))
ggplot(data.frame(out), aes(x=grade, fill = college)) +
  geom_bar() +
  labs(title="Grades per college") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="College")) +
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
# percentage for each grade within each college
grade_percentage <- out %>%
  group_by(college, grade) %>%
  summarise(student_count = n()) %>%
  group_by(college) %>%
  mutate(percentage = student_count / sum(student_count))
png(file=paste(writePath, 'grade_per.png'))
ggplot(grade_percentage, aes(x=grade, y=percentage * 100, fill = college)) +
  geom_bar(stat="identity") +
  labs(title="Grades per college (percentage)", y="Percentage (%)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="College")) +
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'days_active.png'))
ggplot(data.frame(out), aes(x=days_active, fill = college)) +
  geom_bar() + labs(title="Days active per college") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(fill=guide_legend(title="College")) +
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
## time
png(file=paste(writePath, 'time_mean_grade.png'))
ggplot(out_summary_long1, aes(x = time_period, y = mean, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Mean interactions over time periods by college",
       x = "Time period",
       y = "Mean interactions") +
```

```
  scale_x_discrete(labels = c("total_period_1" = "12am-2:59am",
                              "total_period_2" = "3am-5:59am",
                              "total_period_3" = "6am-8:59am",
                              "total_period_4" = "9am-11:59am",
                              "total_period_5" = "12pm-2:59pm",
                              "total_period_6" = "3pm-5:59pm",
                              "total_period_7" = "6pm-8:59pm",
                              "total_period_8" = "9pm-11:59pm")) +
  coord_cartesian(ylim = c(0, max_y_1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()



png(file=paste(writePath, 'time_max_grade.png'))
ggplot(out_summary_long1, aes(x = time_period, y = max, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Maximum interactions over time periods by college",
       x = "Time period",
       y = "Maximum interactions") +
  scale_x_discrete(labels = c("total_period_1" = "12am-2:59am",
                              "total_period_2" = "3am-5:59am",
                              "total_period_3" = "6am-8:59am",
                              "total_period_4" = "9am-11:59am",
                              "total_period_5" = "12pm-2:59pm",
                              "total_period_6" = "3pm-5:59pm",
                              "total_period_7" = "6pm-8:59pm",
                              "total_period_8" = "9pm-11:59pm")) +
  coord_cartesian(ylim = c(0, max_y_1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
png(file=paste(writePath, 'time_median_grade.png'))
ggplot(out_summary_long1, aes(x = time_period, y = median, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Median interactions over time periods by college",
       x = "Time period",
       y = "Median interactions") +
  scale_x_discrete(labels = c("total_period_1" = "12am-2:59am",
                              "total_period_2" = "3am-5:59am",
```

```
                    "total_period_3" = "6am-8:59am",
                    "total_period_4" = "9am-11:59am",
                    "total_period_5" = "12pm-2:59pm",
                    "total_period_6" = "3pm-5:59pm",
                    "total_period_7" = "6pm-8:59pm",
                    "total_period_8" = "9pm-11:59pm")) +
  coord_cartesian(ylim = c(0, max_y_1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
##
### Heatmap
png(file=paste(writePath, 'time_grade_heat.png'))
  ggplot(time_long, aes(x = grade, y = time_period, fill = mean)) +
    geom_tile() +
    geom_label(aes(label = round(mean, 1)), color = "black", fill = "white", size = 3,
            fontface = "bold", label.padding = unit(0.20, "lines"), label.size = 0.15) +
    scale_fill_viridis_c() +
    labs(x = "Grade", y = "Time period", title = "Mean LMS interactions by time period and
grade") +
    scale_y_discrete(labels = c("total_period_1" = "12am-2:59am",
                            "total_period_2" = "3am-5:59am",
                            "total_period_3" = "6am-8:59am",
                            "total_period_4" = "9am-11:59am",
                            "total_period_5" = "12pm-2:59pm",
                            "total_period_6" = "3pm-5:59pm",
                            "total_period_7" = "6pm-8:59pm",
                            "total_period_8" = "9pm-11:59pm"),
                limits = rev(levels(time_long$time_period))) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()
# semester2
png(file=paste(writePath, 'sem_mean_grade.png'))
ggplot(out_summary_long2, aes(x = sem_period, y = mean, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Mean interactions over semester periods by college",
       x = "Semester period",
       y = "Mean interactions") +
  scale_x_discrete(labels = c("total_sem_period_1" = "Week 0-2",
                    "total_sem_period_2" = "Week 3-6",
                    "total_sem_period_3" = "Week 7-Break",
                    "total_sem_period_4" = "Week 9-12",
                    "total_sem_period_5" = "Week 13+")) +
```

```
  coord_cartesian(ylim = c(0, max_y_2)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
##
## Heatmap
png(file=paste(writePath, 'sem_grade_heat.png'))
ggplot(sem_long, aes(x = grade, y = sem_period, fill = mean)) +
  geom_tile() +
  geom_label(aes(label = round(mean, 1)), color = "black", fill = "white", size = 3,
          fontface = "bold", label.padding = unit(0.20, "lines"), label.size = 0.15) +
  scale_fill_viridis_c() +
  labs(x = "Grade", y = "Semester period", title = "Mean LMS interactions by semester period
and grade") +
  scale_y_discrete(labels = c("total_sem_period_1" = "Week 0-2",
                              "total_sem_period_2" = "Week 3-6",
                              "total_sem_period_3" = "Week 7-Break",
                              "total_sem_period_4" = "Week 9-12",
                              "total_sem_period_5" = "Week 13+"),
                   limits = rev(levels(sem_long$sem_period))) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()
##
png(file=paste(writePath, 'sem_avg_grade.png'))
ggplot(out_summary_long2, aes(x = sem_period, y = mean, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Mean interactions over semester periods by college",
       x = "Semester period",
       y = "Mean interactions") +
  scale_x_discrete(labels = c("total_sem_period_1" = "Week 0-2",
                              "total_sem_period_2" = "Week 3-6",
                              "total_sem_period_3" = "Week 7-Break",
                              "total_sem_period_4" = "Week 9-12",
                              "total_sem_period_5" = "Week 13+")) +
  coord_cartesian(ylim = c(0, max_y_2)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
png(file=paste(writePath, 'sem_max_grade.png'))
ggplot(out_summary_long2, aes(x = sem_period, y = max, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
```

```
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Maximum interactions over semester periods by college",
       x = "Semester period",
       y = "Maximum interactions") +
  scale_x_discrete(labels = c("total_sem_period_1" = "Week 0-2",
                              "total_sem_period_2" = "Week 3-6",
                              "total_sem_period_3" = "Week 7-Break",
                              "total_sem_period_4" = "Week 9-12",
                              "total_sem_period_5" = "Week 13+")) +
  coord_cartesian(ylim = c(0, max_y_2)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
png(file=paste(writePath, 'sem_med_grade.png'))
ggplot(out_summary_long2, aes(x = sem_period, y = median, group = grade, color = grade,
shape = grade)) +
  geom_line() +
  geom_point() +
  facet_wrap(~college, scales = "fixed") +
  labs(title = "Median Interactions over Semester Periods by college",
       x = "Semester Period",
       y = "Mean Interactions") +
  scale_x_discrete(labels = c("total_sem_period_1" = "Week 0-2",
                              "total_sem_period_2" = "Week 3-6",
                              "total_sem_period_3" = "Week 7-Break",
                              "total_sem_period_4" = "Week 9-12",
                              "total_sem_period_5" = "Week 13+")) +
  coord_cartesian(ylim = c(0, max_y_2)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
dev.off()
# days active and grades by college
png(file=paste(writePath, 'Days_active_col_grade.png'))
ggplot(out, aes(x = grade, y = days_active, fill = grade)) +
  geom_boxplot() +
  labs(title = "Total days active and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
# percentages
png(file=paste(writePath, 'Video_percent.png'))
ggplot(out, aes(x = grade, y = video_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage of videos viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
```

```
png(file=paste(writePath, 'quiz_percent.png'))
ggplot(out, aes(x = grade, y = quiz_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage of quiz modules viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5)) +
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'assignment_percent.png'))
ggplot(out, aes(x = grade, y = assignment_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage of assignment modules viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'participation_percent.png'))
ggplot(out, aes(x = grade, y = participation_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage of participation modules viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'support_percent.png'))
ggplot(out, aes(x = grade, y = support_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage of support modules viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'other_percent.png'))
ggplot(out, aes(x = grade, y = other_percent, fill = grade)) +
  geom_boxplot() +
  labs(title = "Percentage other modules viewed and grade by college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
# totals
png(file = paste(writePath, 'Video_total.png'))
ggplot(out[out$Video_interactions > 0, ], aes(x = grade, y = Video_interactions, fill = grade))
+
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total video modules viewed by grade and college",
    y = expression("Log10(Video Interactions)")
  ) +
  theme(axis.text.x = element_text(hjust = 0.5)) +
  facet_wrap(~ college, scales = "fixed", ncol = 2)
```

```
dev.off()
png(file=paste(writePath, 'Quiz_total.png'))
ggplot(out[out$Activity_interactions > 0, ], aes(x = grade, y = Activity_interactions, fill =
grade)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total quiz modules viewed by grade and college",
    y = expression("Log10(Quiz interactions)")
  ) +
  labs(title = "Total quiz modules viewed by grade and college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'Assignment_total.png'))
ggplot(out[out$Assignment_interactions > 0, ], aes(x = grade, y = Assignment_interactions,
fill = grade)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total assignment modules viewed by grade and college",
    y = expression("Log10(Assignment interactions)")
  ) +
  labs(title = "Total assignment modules viewed by grade and college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'Participation_total.png'))
ggplot(out[out$Course_participation_interactions > 0, ], aes(x = grade, y =
Course_participation_interactions, fill = grade)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total participation modules viewed by grade and college",
    y = expression("Log10(Participation interactions)")
  ) +
  labs(title = "Total participation modules viewed by grade and college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
png(file=paste(writePath, 'Support_total.png'))
ggplot(out[out$Support_material_interactions > 0, ], aes(x = grade, y =
Support_material_interactions, fill = grade)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total support modules viewed by grade and college",
```

```
    y = expression("Log10(Support interactions)")
  ) +
  labs(title = "Total support modules viewed by grade and college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()

png(file=paste(writePath, 'Other_total.png'))
ggplot(out[out$Other_interactions > 0, ], aes(x = grade, y = Other_interactions, fill = grade))
+
  geom_boxplot() +
  scale_y_log10() +
  labs(
    title = "Total other modules viewed by grade and college",
    y = expression("Log10(Other interactions)")
  ) +
  labs(title = "Total other modules viewed by grade and college") +
  theme(axis.text.x = element_text(hjust = 0.5))+
  facet_wrap(~ college, scales = "fixed", ncol = 2)
dev.off()
# topic structure
create_college_plots <- function(d, column, title, filename) {
  png(file = paste(writePath, paste0(filename, '.png')), width = 1200, height = 900)  # Increase
file dimensions as needed
  print(
    ggplot(data = d, aes_string(x = column, fill = "college")) +
      geom_bar(stat = "count", position = "dodge") +
      labs(title = title, x = "Topic Shortcode", y = "Count", fill = "College") +
      theme_minimal(base_size = 14) +  # Use a minimal theme and increase base text size
      theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5),  # Rotate the x-axis labels
        plot.title = element_text(hjust = 0.5),
        legend.position = "bottom"
      ) +
      facet_wrap(~ college, scales = "free_x", ncol = 3) +  # Use free scales and adjust the
number of columns as needed
      guides(fill = guide_legend(title = "College"))
  )
  dev.off()
}
# topic structure per topic type
create_college_plots_types <- function(d, column, title, filename) {
  for (college_name in unique(d$college)) {
    college_data <- d[d$college == college_name, ]
    png(file = paste(writePath, paste0(filename, '_', college_name, '.png')))
    print(
      ggplot(data = college_data, aes(x = topic_shortcode, fill = college)) +
```

```
      geom_bar(stat = "count", position = "dodge") +
      labs(title = paste(title, college_name)) +
      theme(plot.title = element_text(hjust = 0.5)) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
      guides(fill = guide_legend(title = "College"))
    )
    dev.off()
  }
}
# Topic videos per college
topicdata_video <- topicdata[!(topicdata$lecture_videos == 0),]
create_college_plots(topicdata_video, "lecture_videos", "Topic videos per college",
"Topic_videos_col")
create_college_plots_types(topicdata_video, "lecture_videos", "Topic videos per college",
"Topic_videos")
# Topic quizzes per college
topicdata_quiz <- topicdata[!(topicdata$quiz_count == 0),]
create_college_plots(topicdata_quiz, "quiz_count", "Topic quizzes per college",
"Topic_quiz_col")
create_college_plots_types(topicdata_quiz, "quiz_count", "Topic quizzes per college",
"Topic_quiz")
# Topic assignments per college
topicdata_assignments <- topicdata[!(topicdata$assignment_count == 0),]
create_college_plots(topicdata_assignments, "assignment_count", "Topic assignments per
college", "Topic_assignments_col")
create_college_plots_types(topicdata_assignments, "assignment_count", "Topic
assignments per college", "Topic_assignments")
# Topic forum posts per college
topicdata_forum <- topicdata[topicdata$forum_activity >= 1,]
create_college_plots(topicdata_forum, "topic_shortcode", "Topic forum posts per college",
"Topic_forum_col")
create_college_plots_types(topicdata_forum, "topic_shortcode", "Topic forum posts per
college", "Topic_forum")
# Topic participation modules per college
topicdata_participation <- topicdata[!(topicdata$participation_activity_count == 0),]
create_college_plots(topicdata_participation, "participation_activity_count", "Topic
participation modules per college", "Topic_participation_col")
create_college_plots_types(topicdata_participation, "participation_activity_count", "Topic
participation modules per college", "Topic_participation")
# Topic support modules per college
topicdata_support <- topicdata[!(topicdata$support_matrials_count == 0),]
create_college_plots(topicdata_support, "support_matrials_count", "Topic support modules
per college", "Topic_support_col")
create_college_plots_types(topicdata_support, "support_matrials_count", "Topic support
modules per college", "Topic_support")
# Topic other modules per college
topicdata_other <- topicdata[!(topicdata$other_interactions_count == 0),]
```

```
create_college_plots(topicdata_other, "other_interactions_count", "Topic other modules per college", "Topic_other_col")
create_college_plots_types(topicdata_other, "other_interactions_count", "Topic other modules per college", "Topic_other")
##
```

7.5. Appendix E: Python scripts for decision tree manipulation and visualisation

7.5.1. Perform WEKA classification

```
# -*- coding: utf-8 -*-
"""
@author: Adam
Perform WEKA Classifications and generate .dot files
"""
import jpype
import jpype.imports
from jpype.types import *
import os
output_dir = "R:\Experiment_trees\dot\"
# Check if JVM is already started
if not jpype.isJVMStarted():
    jpype.startJVM(classpath=['C:\Program Files\Weka-3-8-6\weka.jar'])
# Import Java classes
from weka.classifiers.trees import REPTree
from weka.core.converters import ConverterUtils
# Define the operations
operations = [
    {
     'input_file': r'R:\dat\predictGrade_excCol_1S_mod.arff',
     'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
     'output_file': 'ex1_predict_grade_tree.dot',
    },
    {
     'input_file': r'R:\dat\predictCollege_excGra_1S.arff',
     'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
     'output_file': 'ex2_predict_college_tree.dot',
    },
    {
     'input_file': r'R:\dat\predictGrade_col1_1S_mod.arff',
     'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
     'output_file': 'ex3_predict_grade_col_1_tree.dot',
    },
    {
     'input_file': r'R:\dat\predictGrade_col2_1S_mod.arff',
     'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
     'output_file': 'ex3_predict_grade_col_2_tree.dot',
    },
    {
```

```
 'input_file': r'R:\dat\predictGrade_col3_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_3_tree.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col4_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_4_tree.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col5_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_5_tree.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col6_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L -1 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_6_tree.dot',
},
{
 'input_file': r'R:\dat\predictGrade_excCol_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex1_predict_grade_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictCollege_excGra_1S.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex2_predict_college_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col1_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_1_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col2_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_2_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col3_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_3_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col4_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
```

```
 'output_file': 'ex3_predict_grade_col_4_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col5_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_5_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col6_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 7 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_6_tree_redlvl7.dot',
},
{
 'input_file': r'R:\dat\predictGrade_excCol_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex1_predict_grade_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictCollege_excGra_1S.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex2_predict_college_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col1_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_1_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col2_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_2_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col3_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_3_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col4_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_4_tree_redlvl5.dot',
},
{
 'input_file': r'R:\dat\predictGrade_col5_1S_mod.arff',
 'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
 'output_file': 'ex3_predict_grade_col_5_tree_redlvl5.dot',
},
```

```
{
    'input_file': r'R:\dat\predictGrade_col6_1S_mod.arff',
    'options': '-M 20 -V 0.001 -N 3 -S 1 -L 5 -I 0.0 -num-decimal-places 0',
    'output_file': 'ex3_predict_grade_col_6_tree_redlvl5.dot',
   }
]
for op in operations:
   # Instantiate REPTree classifier
   tree = REPTree()
   # Python list of options
   options = op['options'].split()
   # Convert Python list to Java array
   java_options = jpype.JArray(jpype.java.lang.String)(options)
   # Set options for REPTree
   tree.setOptions(java_options)
   # Load dataset
   data_source = ConverterUtils.DataSource(op['input_file'])
   data = data_source.getDataSet()
   data.setClassIndex(data.numAttributes() - 1)
   # Build classifier
   tree.buildClassifier(data)
   # Save the model tree to a file (dot format)
   graph_content = tree.graph()
   # Construct the full path for the output file
   output_file_path = os.path.join(output_dir, op['output_file'])
   with open(output_file_path, 'w') as f:
      # Convert Java string to Python string and write to file
      f.write(str(graph_content))
   print(f"Completed operation for {op['input_file']}c, output saved to {output_file_path}")
# Shutdown the JVM when finished
#jpype.shutdownJVM()
```

7.5.2. Restructure trees

```
"""
@author: Adam
Cleaning trees, restructure and simplify
"""
import pydot
import re
paths = [
    "R:\Experiment_trees\dot\ex1_predict_grade_tree.dot",
    "R:\Experiment_trees\dot\ex1_predict_grade_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex1_predict_grade_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex2_predict_college_tree.dot",
    "R:\Experiment_trees\dot\ex2_predict_college_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex2_predict_college_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree_redlvl5.dot"
]

# Define color scheme for leaf nodes
color_scheme = {
    'F': '#FF5733',  # Red
    'P': '#FFD433',   # Yellow
    'CR': '#33FF57',  # Green
    'DN': '#8D33FF',  # Purple
    'HD': '#33C1FF',   # Blue
    'BGL': '#FF5733',  # Red
    'EPS': '#FFD433',  # Yellow
    'HAS': '#33FF57',  # Green
    'MPH': '#33FFD5',  # Cyan
```

```
    'NHS': '#8D33FF',  # Purple
    'S&E': '#33C1FF'   # Blue
}
# Graph attribute settings
graph_attributes = {
    'rankdir': 'TB',
    'nodesep': 0,
    'ranksep': 0.5
}
# Edge attribute settings (default for all edges)
edge_attributes = {
    'color': 'black',
    'arrowhead': 'normal',
    'arrowsize': 1
}

def format_label(label):
    # Remove leading sections of number and colon
    label = re.sub(r'^[0-9]+\s*:\s*', '', label)

    # Replace underscores with spaces and capitalise each word
    label = ' '.join(word.capitalize() for word in label.replace('_', ' ').split())
    # Break the label into multiple lines
    label = '\n'.join(label.split())
    return label
def convert_college_number(label):
    # Convert numbers to codes for colleges
    college_mapping = {
        '1': 'BGL',
        '2': 'EPS',
        '3': 'HAS',
        '4': 'MPH',
        '5': 'NHS',
        '6': 'S&E'
    }
    # Directly replace the number at the start of the string with the college code
    match = re.match(r'^(\d+)', label)
    if match:
        num = match.group(1)
        if num in college_mapping:
            label = re.sub(r'^\d+', college_mapping[num], label, 1)
    return label
```

```
def trim_leaf(label):
    # Remove the leading numeric prefix and spaces
    label = re.sub(r'^\d+\s*:\s*', '', label)
    # Apply college mapping (before handling the pattern)
    label = convert_college_number(label).upper()
    # Step 3: Perform calculation and replace text with newline
    def replace_with_calculation(match):
        c, d = map(int, match.groups())
        result = 1 - (d / c)
        return f"\n{result:.2f}"
    # Adjusted regex to ensure it captures the correct part for calculation
    label = re.sub(r'\(\d+/\d+\)\s*\[(\d+)/(\d+)\]', replace_with_calculation, label)
    # Remove any remaining pattern that was not intended for calculation
    label = re.sub(r'\(\d+/\d+\)', '', label).strip()
    return label
for path in paths:
    (graph,) = pydot.graph_from_dot_file(path)
    attribute_labels = set()  # Set to hold unique attribute labels
    # Apply graph-level attributes
    for attr, value in graph_attributes.items():
        graph.set(attr, value)
    # Create a set of all source nodes (nodes with outgoing edges)
    source_nodes = {edge.get_source() for edge in graph.get_edge_list()}
    # Initialize a set to store attribute labels
    attribute_labels = set()
    # Process nodes
    for node in graph.get_nodes():
        node_label = node.get_attributes().get('label', '').strip('"').strip()
        node_name = node.get_name().strip('"')
        # Modify the node label
        if node_name in source_nodes:
            # Non-leaf nodes
            formatted_label = format_label(node_label)
            attribute_labels.add(formatted_label.replace('\n', ' '))
            node_label = formatted_label
        else:
            # Leaf nodes
            node_label = trim_leaf(node_label)
            for pattern, color in color_scheme.items():
                if re.search(pattern, node_label, re.IGNORECASE):
                    node.set_fillcolor(color)
                    node.set_style('filled')
                    break
        node.set_label(node_label)
    # Apply edge-level attributes
    for edge in graph.get_edge_list():
        for attr, value in edge_attributes.items():
```

```
        edge.set(attr, value)
# Saving modified .dot file
modified_path = path.replace('.dot', '_mod.dot')
graph.write_dot(modified_path)
# Use PyDot to render the .dot file to a PNG
png_path = path.replace('.dot', '.png')
graph.write_png(png_path)
print(f"Rendered PNG saved to {png_path}")

# Write attribute labels to a .txt file
txt_path = path.replace('.dot', '_attributes.txt')
with open(txt_path, 'w') as f:
    for attribute in sorted(attribute_labels):
        f.write(f"{attribute}\n")
print(f"Attribute labels saved to {txt_path}")
```

7.5.3. Splitting trees

```
"""
@author: Adam
Splitting trees for paths
"""
import pydot
import re
import os
gpaths = [
    "R:\Experiment_trees\dot\ex1_predict_grade_tree.dot",
    "R:\Experiment_trees\dot\ex1_predict_grade_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex1_predict_grade_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_2_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_3_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_4_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_5_tree_redlvl5.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex3_predict_grade_col_6_tree_redlvl5.dot"
]
cpaths = [
    "R:\Experiment_trees\dot\ex2_predict_college_tree.dot",
    "R:\Experiment_trees\dot\ex2_predict_college_tree_redlvl7.dot",
    "R:\Experiment_trees\dot\ex2_predict_college_tree_redlvl5.dot"
]
def process_tree(path, class_labels):
    (graph,) = pydot.graph_from_dot_file(path)
    # Getting all nodes and edges
    nodes = graph.get_nodes()
    edges = graph.get_edge_list()
    # Mapping node names to nodes and extracting leaf nodes
    node_map = {node.get_name(): node for node in nodes}
    leaf_nodes = [node for node in nodes if node.get_name() not in [edge.get_source() for
edge in edges]]
```

```
    # Filtering leaf nodes by class
    for class_label in class_labels:
        subgraph = pydot.Dot(graph_type='digraph')
        added_edges = set()
        non_leaf_nodes = []
        # Add nodes and edges if the leaf node's label matches the class label
        for leaf_node in leaf_nodes:
            if class_label in leaf_node.get_label():
                current_node = leaf_node
                while current_node is not None:
                    subgraph.add_node(current_node)
                    if current_node not in leaf_nodes:
                        non_leaf_nodes.append(current_node)  # Add non-leaf node
                    parent_edge = next((edge for edge in edges if edge.get_destination() ==
current_node.get_name()), None)
                    if parent_edge and (parent_edge.get_source(), parent_edge.get_destination())
                        subgraph.add_edge(parent_edge)
                        added_edges.add((parent_edge.get_source(), parent_edge.get_destination()))
                        current_node = node_map.get(parent_edge.get_source())
                    else:
                        current_node = None
        # Saving the subgraph's non-leaf nodes to a text file
        class_label_clean = re.sub(r'\W+', '', class_label)
        subgraph_path = os.path.splitext(path)[0] + f"_{class_label_clean}"
        subgraph.write_dot(subgraph_path + '.dot')
        subgraph.write_png(subgraph_path + '.png')
        # Write non-leaf nodes to a text file
        with open(subgraph_path + '.txt', 'w') as txt_file:
            for node in non_leaf_nodes:
                txt_file.write(f'Node: {node.get_name()}, Label: {node.get_label()}\n')
for path in gpaths:
    class_labels = ['F', 'P', 'CR', 'DN', 'HD']
    process_tree(path, class_labels)
for path in cpaths:
    class_labels = ['BGL', 'EPS', 'HAS', 'MPH', 'NHS', 'S&E']
    process_tree(path, class_labels)

"""
@author: Adam
Fix tall trees
"""
import pydot
path = " R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_HD.dot"
out1 = "R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_HD_shape.dot"
out2 = " R:\Experiment_trees\dot\ex3_predict_grade_col_1_tree_HD_shape.png"
# Load the graph
(graph,) = pydot.graph_from_dot_file(path)
```

```
# This makes the graph layout horizontal. (TB = top to bottom, LR = left to right)
graph.set_rankdir('TB')
#graph.set_rankdir('LR')
graph.set('ranksep',0.0)
# Saving modified .dot and .png
graph.write_dot(out1)
graph.write_png(out2)
```

7.6. Appendix F: Ethics approval for research

From: Human Research Ethics
Sent: 20 July 2018 12:11:42
To: Adam Wilden; Denise de Vries; Anna Shillabeer
Subject: 7987 SBREC Final approval notice (20 July 2018)

Dear Adam,

The Chair of the Social and Behavioural Research Ethics Committee (SBREC) at Flinders University considered your response to conditional approval out of session and your project has now been granted final ethics approval. This means that you now have approval to commence your research. Your ethics final approval notice can be found below.

FINAL APPROVAL NOTICE

Project No.:	**7987**
Project Title:	E-learning in Higher Education
Principal Researcher:	Mr Adam Wilden
Email:	adam.wilden@flinders.edu.au
Approval Date:	20 July 2018
Ethics Approval Expiry Date:	**6 August 2021**

The above proposed project has been **approved** on the basis of the information contained in the application, its attachments and the information subsequently provided.

8. References

Abdul Razzak, N 2022, 'E-Learning: From an Option to an Obligation', *International Journal of Technology in Education and Science*, vol. 6, no. 1, pp. 86-110.

Abeysekera, L & Dawson, P 2015, 'Motivation and Cognitive Load in the Flipped Classroom: Definition, Rationale and a Call for Research', *Higher Education Research and Development*, vol. 34, no. 1, pp. 1-14.

Aboagye, E, Yawson, JA & Appiah, KN 2021, 'Covid-19 and E-Learning: The Challenges of Students in Tertiary Institutions', *Social Education Research*, vol. 2, no. 1, pp. 1-8.

Abu-Rasheed, H, Weber, C & Fathi, M 2023, 'Context Based Learning: A Survey of Contextual Indicators for Personalized and Adaptive Learning Recommendations – a Pedagogical and Technical Perspective', *Frontiers in Education*, vol. 8.

Agrawal, R & Srikant, R 1994, 'Fast Algorithms for Mining Association Rules', in *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB'94*, vol. 1215, pp. 487-99.

Agresti, A 2010, *Analysis of Ordinal Categorical Data*, 2nd edn, vol. 656, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ.

Aissaoui, N 2022, 'The Digital Divide: A Literature Review and Some Directions for Future Research in Light of Covid-19', *Global Knowledge, Memory and Communication*, vol. 71, no. 8/9, pp. 686-708.

Al-Roomy, MA 2023, 'The Relationship among Students' Learning Styles, Health Sciences Colleges, and Grade Point Average (Gpa)', *Advances in Medical Education and Practice*, vol. 14, no. null, pp. 203-13.

Allen, WC 2006, 'Overview and Evolution of the Addie Training System', *Advances in developing human resources*, vol. 8, no. 4, pp. 430-41.

Ally, M 2004, 'Foundations of Educational Theory for Online Learning', *Theory and practice of online learning*, vol. 2, pp. 15-44.

Almelhi, AM 2021, 'Effectiveness of the Addie Model within an E-Learning Environment in Developing Creative Writing in Efl Students', *English Language Teaching*, vol. 14, no. 2, pp. 20-36.

Altinpulluk, H, Kilinc, H, Firat, M & Yumurtaci, O 2019, 'The Influence of Segmented and Complete Educational Videos on the Cognitive Load, Satisfaction, Engagement, and Academic Achievement Levels of Learners', *Journal of Computers in Education*, vol. 7, no. 2, pp. 155-82.

Anderson, TW & Darling, DA 1952, 'Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes', *The annals of mathematical statistics*, vol. 23, no. 2, pp. 193-212.

Anitha, R, Cowsigan, SP, Vanitha, U, Niranjana, MI, Shadrach, FD & Raagavi, S 2022, 'Performance Analysis of E-Learning System Using Data Mining Techniques', in *Ieee 2nd International Conference On Mobile Networks And Wireless Communications (Icmnwc)*, pp. 1-8.

Arnold, KE & Pistilli, MD 2012, 'Course Signals at Purdue: Using Learning Analytics to Increase Student Success', in *2nd International Conference On Learning Analytics And Knowledge (Lak '12)*, Vancouver British Columbia Canada, pp. 267-70.

Assiry, KAH & Muniasamy, A 2022, 'Predicting Learning Styles Using Machine Learning Classifiers', in *International Conference on Electrical, Computer and Energy Technologies (Icecet)*, Prague, Czech Republic, pp. 1-7.

Avella, JT, Kebritchi, M, Nunn, SG & Kanai, T 2016, 'Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review', *Online Learning*, vol. 20, no. 2, pp. 13-29.

Aytekin, C 2022, 'Neural Networks Are Decision Trees', *arXiv preprint arXiv:2210.05189*.

Baek, C & Doleck, T 2022, 'Educational Data Mining: A Bibliometric Analysis of an Emerging Field', *IEEE Access*, vol. 10, pp. 31289-96.

Baherimoghadam, T, Hamedani, S, mehrabi, M, Naseri, N & Marzban, N 2021, 'The Effect of Learning Style and General Self-Efficacy on Satisfaction of E-Learning in Dental Students', *BMC medical education*, vol. 21, no. 1, p. 463.

Bai, X 2017, 'Promote Technology Self-Efficacy Via a Scorm-Based E-Learning Approach', *International Journal of Information and Education Technology*, vol. 7, no. 8, p. 575.

Bąkała, A & Bąkała, M 2020, 'E-Learning Course Design in Addie Methodology as a Process in Bpmn 2.0', *Informatyka Ekonomiczna*, vol. 4, no. 58, pp. 21-32.

Baker, R & Inventado, P 2014, 'Educational Data Mining and Learning Analytics', in JA Larusson & B White (eds), *Educational Data Mining and Learning Analytics*, Springer, New York, NY, pp. 61-75.

Bandura, A 1977, 'Self-Efficacy: Toward a Unifying Theory of Behavioral Change', *Psychological review*, vol. 84, no. 2, pp. 191-215.

Bandura, A 2002, 'Social Cognitive Theory: An Agentic Perspective', *Asian Journal of Social Psychology*, vol. 2, no. 1, pp. 21-41.

Bannan, B 2013, 'The Integrative Learning Design Framework: An Illustrated Example from the Domain of Instructional Technology', in T Plomp & N Nieveen (eds), *Educational Design Research Part A: An Introduction*, 2nd edn, Netherlands institute for curriculum development (SLO), Enschede, The Netherlands, pp. 114-33.

Barbara, M & Donna, V 2005, 'Learner-Centered Framework for E-Learning', *Teachers college record*, vol. 107, no. 8, pp. 1582-600.

Beare, PL 1989, 'Media: The Comparative Effectiveness of Videotape, Audiotape, and Telelecture in Delivering Continuing Teacher Education', *The American Journal of Distance Education*, vol. 3, no. 2, pp. 57-66.

Beatty, BJ, Merchant, Z & Albert, M 2017, 'Analysis of Student Use of Video in a Flipped Classroom', *TechTrends*, vol. 63, no. 4, pp. 376-85.

Beaudoin, MF 2002, 'Learning or Lurking?: Tracking the “Invisible” Online Student', *The Internet and Higher Education*, vol. 5, no. 2, pp. 147-55.

Becher, T 2001, *Academic Tribes and Territories*, 2nd edn, SRHE and Open University Press, Ballmore, Buckingham.

Beetham, H & Sharpe, R 2007, *Rethinking Pedagogy for a Digital Age: Designing for 21st Century Learning*, Routledge, Milton Park, Abingdon.

Behrens, JT 1997, 'Principles and Procedures of Exploratory Data Analysis', *Psychological methods*, vol. 2, no. 2, pp. 131-60.

Bertholdo, APO, Melo, CdO, Rozestraten, AS & Gerosa, MA 2018, 'Relations between Actions Performed by Users and Their Engagement', in A Rodrigues, B Fonseca & N Preguiça (eds), *Collaboration and Technology*, Cham, pp. 207-22.

Beşoluk, Ş, Önder, İ & Deveci, İ 2011, 'Morningness-Eveningness Preferences and Academic Achievement of University Students', *Chronobiology International*, vol. 28, no. 2, pp. 118-25.

Better Buys Staff 2023, *How Much Does an Lms Cost? 2023 Pricing Guide*, Better Buys, viewed 17 Feb 2024, <https://www.betterbuys.com/lms/lms-pricing-guide/>.

Biglan, A 1973a, 'The Characteristics of Subject Matter in Different Academic Areas', *Journal of applied Psychology*, vol. 57, no. 3, p. 195.

Biglan, A 1973b, 'Relationships between Subject Matter Characteristics and the Structure and Output of University Departments', *Journal of applied Psychology*, vol. 57, no. 3, p. 204.

Bonferroni, C 1936, *Teoria Statistica Delle Classi E Calcolo Delle Probabilita*, vol. 8, Pubblicazioni Del R. Istituto Superiore Di Scienze Economiche E Commericiali Di Firenze, Seeber, Florence, Italy.

Borba, MC, Askar, P, Engelbrecht, J, Gadanidis, G, Llinares, S & Aguilar, MS 2016, 'Blended Learning, E-Learning and Mobile Learning in Mathematics Education', *ZDM*, vol. 48, pp. 589-610.

Bouckaert, RR & Frank, E 2004, 'Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms', in *8th Pacific-Asia Conference, PAKDD*, Berlin, Heidelberg, pp. 3-12.

Branch, RM 2009, *Instructional Design: The Addie Approach*, vol. 722, Springer, Spring Street, New York.

Branch, RM & Dousay, TA 2015, *Survey of Instructional Design*, 5th edn, Association for Educational Communications and Technology, Bloomington, Indiana.

Breiman, L 1996, 'Bagging Predictors', *Machine learning*, vol. 24, no. 2, pp. 123-40.

Breiman, L 2001, 'Random Forests', *Machine learning*, vol. 45, no. 1, pp. 5-32.

Breiman, L, Friedman, J, Stone, CJ & Olshen, RA 1984, *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, FL.

Brusilovsky, P & Millán, E 2007, 'User Models for Adaptive Hypermedia and Adaptive Educational Systems', in *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer-Verlag, Berlin, Heidelberg, vol. 4321, pp. 3-53.

Bryant, J, Child, F, Dorn, E, Espinosa, J, Hall, S, Kola-Oyeneyin, T, Lim, C, Panier, F, Sarakatsannis, J, Schmautzer, D, Ungur, S & Woord, B 2022, *How Covid-19 Caused a Global Learning Crisis*, McKinsey & Company, viewed 17 Feb 2024, <https://www.mckinsey.com/industries/education/our-insights/how-covid-19-caused-a-global-learning-crisis#/>.

Castaño-Muñoz, J, Duart, JM & Sancho-Vinuesa, T 2014, 'The Internet in Face-to-Face Higher Education: Can Interactive Learning Improve Academic Achievement?', *British Journal of Educational Technology*, vol. 45, no. 1, pp. 149-59.

Castro, MDB & Tumibay, GM 2019, 'A Literature Review: Efficacy of Online Learning Courses for Higher Education Institution Using Meta-Analysis', *Education and Information Technologies*, vol. 26, no. 2, pp. 1367-85.

Cerezo, R, Lara, JA, Azevedo, R & Romero, C 2024, 'Reviewing the Differences between Learning Analytics and Educational Data Mining: Towards Educational Data Science', *Computers in Human Behavior*, vol. 154, p. 108155.

Chang, J 2024, *Costs and Pricing Models of Learning Management Systems*, FinancesOnline, viewed 17 Feb 2024, <https://learning-management.financesonline.com/costs-pricing-models-learning-management-systems/>.

Chaubey, A & Bhattacharya, B 2015, 'Learning Management System in Higher Education', *International Journal of Science Technology and Engineering*, vol. 2, no. 3, pp. 158-62.

Chicco, D & Jurman, G 2020, 'The Advantages of the Matthews Correlation Coefficient (Mcc) over F1 Score and Accuracy in Binary Classification Evaluation', *BMC Genomics*, vol. 21, no. 1, p. 6.

Clow, D 2013, 'An Overview of Learning Analytics', *Teaching in Higher Education*, vol. 18, no. 6, pp. 683-95.

Cohen, J 1960, 'A Coefficient of Agreement for Nominal Scales', *Educational and psychological measurement*, vol. 20, no. 1, pp. 37-46.

Cohen, J 1992, 'Statistical Power Analysis', *Current Directions in Psychological Science*, vol. 1, no. 3, pp. 98-101.

Cox, DR 1958, 'The Regression Analysis of Binary Sequences', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215-42.

Cuban, L 1986, *Teachers and Machines: The Classroom Use of Technology since 1920*, Teachers College Press, New York, NY.

Dabbagh, N & Bannan-Ritland, B 2005, *Online Learning: Concepts, Strategies, and Application*, 1st edn, Pearson/Merrill/Prentice Hall, Upper Saddle River, N.J.

Dagger, D, O'Connor, A, Lawless, S, Walsh, E & Wade, VP 2007, 'Service-Oriented E-Learning Platforms: From Monolithic Systems to Flexible Services', *IEEE Internet Computing*, vol. 11, no. 3, pp. 28-35.

Das, AK, Das, A & Das, S 2015, 'Present Status of Massive Open Online Course (Mooc) Initiatives for Open Education Systems in India–an Analytical Study', *Asian Journal of Multidisciplinary Studies*, vol. 3, no. 7, pp. 67-80.

Dash, G, Akmal, S, Mehta, P & Chakraborty, D 2022, 'Covid-19 and E-Learning Adoption in Higher Education: A Multi-Group Analysis and Recommendation', *Sustainability*, vol. 14, no. 14, p. 8799.

Davidoff, Y & Jayusi, W 2024, 'Effective Online Teaching and Learning Strategies: Interdisciplinary Research of Student Perceptions in Higher Education', *Education and Information Technologies*.

Davis, K, Christodoulou, J, Seider, S & Gardner, H 2011, 'The Theory of Multiple Intelligences', in *The Cambridge Handbook of Intelligence*, Cambridge University Press, New York, USA, pp. 485-503.

Deerwester, S, Dumais, ST, Furnas, GW, Landauer, TK & Harshman, R 1990, 'Indexing by Latent Semantic Analysis', *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407.

Denden, M, Tlili, A, Chen, N-S, Abed, M, Jemni, M & Essalmi, F 2024, 'The Role of Learners' Characteristics in Educational Gamification Systems: A Systematic Meta-Review of the Literature', *Interactive Learning Environments*, vol. 32, no. 3, pp. 790-812.

Diamond, J & Gonzalez, PC 2016, 'Digital Badges for Professional Development: Teachers' Perceptions of the Value of a New Credentialing Currency', in D Ifenthaler, N Bellin-Mularski & D-K Mah (eds), *Foundation of Digital Badges and Micro-Credentials: Demonstrating and Recognizing Knowledge and Competencies*, Springer International Publishing, Cham, pp. 391-409.

DiCarlo, SE 2009, 'Too Much Content, Not Enough Thinking, and Too Little Fun!', *Advances in Physiology Education*, vol. 33, no. 4, pp. 257-64.

Dick, W, Carey, L & Carey, JO 2014, *The Systematic Design of Instruction*, 8th edn, Pearson Education, Inc., New Jersey, NY.

Dietz-Uhler, B & Hurn, JE 2013, 'Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective', *Journal of Interactive Online Learning*, vol. 12, no. 1, pp. 17-26.

Djeki, E, Dégila, J, Bondiombouy, C & Alhassan, MH 2022, 'E-Learning Bibliometric Analysis from 2015 to 2020', *Journal of Computers in Education*, vol. 9, no. 4, pp. 727-54.

Docebo 2014, *E-Learning Market Trends & Forecast 2014 - 2016 Report*, Docebo, viewed 17 Feb 2024, <https://www.academia.edu/31610714/E_Learning_Market_Trends_and_Forecast_2014_2016_Report?auto=download>.

Domingos, P 2012, 'A Few Useful Things to Know About Machine Learning', *Communications of the Acm*, vol. 55, no. 10, pp. 78-87.

Dougiamas, M & Taylor, P 2003, 'Moodle: Using Learning Communities to Create an Open Source Course Management System', paper presented to EdMedia + Innovate Learning 2003, Honolulu, Hawaii, USA, <https://www.learntechlib.org/p/13739>.

Dringus, LP & Cohen, MS 2005, 'An Adaptable Usability Heuristic Checklist for Online Courses', in *Frontiers in Education, 2005. FIE'05. Proceedings 35th Annual Conference*, pp. T2H-6.

Dunn, OJ 1961, 'Multiple Comparisons among Means', *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52-64.

Dunn, OJ 1964, 'Multiple Comparisons Using Rank Sums', *Technometrics*, vol. 6, no. 3, pp. 241-52.

Eckerson, WW 2007, *Predictive Analytics: Extending the Value of Your Data Warehousing Investment*, The Data Warehouse Institute, Chatsworth, CA.

El-Sabagh, HA 2021, 'Adaptive E-Learning Environment Based on Learning Styles and Its Impact on Development Students' Engagement', *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, pp. 1-24.

Eljak, H, Osman, A, Saeed, F, Hashem, I, Abdelmaboud, A, Syed, H, Abulfaraj, A, Ismail, M & Elsafi, A 2023, 'E-Learning-Based Cloud Computing Environment: A Systematic Review, Challenges, and Opportunities', *IEEE Access*, vol. PP, pp. 1-.

Essa, SG, Celik, T & Human-Hendricks, NE 2023, 'Personalized Adaptive Learning Technologies Based on Machine Learning Techniques to Identify Learning Styles: A Systematic Literature Review', *IEEE Access*, vol. 11, pp. 48392-409.

Fahd, K, Miah, SJ & Ahmed, K 2021, 'Predicting Student Performance in a Blended Learning Environment Using Learning Management System Interaction Data', *Applied Computing and Informatics*, vol. ahead-of-print, no. ahead-of-print.

Fariani, RI, Junus, K & Santoso, HB 2022, 'A Systematic Literature Review on Personalised Learning in the Higher Education Context', *Technology, Knowledge and Learning*, vol. 28, no. 2, pp. 449-76.

Fauzi, MA 2022, 'E-Learning in Higher Education Institutions During Covid-19 Pandemic: Current and Future Trends through Bibliometric Analysis', *Heliyon*, vol. 8, no. 5.

Fazil, M, Rísquez, A & Halpin, C 2024, 'A Novel Deep Learning Model for Student Performance Prediction Using Engagement Data', *Journal of Learning Analytics*, vol. 11, no. 2, pp. 23-41.

Felder, RM & Silverman, LK 1988, 'Learning and Teaching Styles in Engineering Education', *Engineering education*, vol. 78, no. 7, pp. 674-81.

Ferguson, R, Coughlan, T, Egelandsdal, K, Gaved, M, Herodotou, C, Hillaire, G, Jones, D, Jowers, I, Kukulska-Hulme, A & McAndrew, P 2019, *Innovating Pedagogy 2019: Open University Innovation Report 7*, The Open University, Milton Keynes, UK.

Ferriman, J 2017, *6 Lms Pricing Models Explained*, LearnDash, viewed 17 Feb 2024, <https://www.learndash.com/6-lms-pricing-models-explained/>.

Fischer, C, Pardos, ZA, Baker, RS, Williams, JJ, Smyth, P, Yu, R, Slater, S, Baker, R & Warschauer, M 2020, 'Mining Big Data in Education: Affordances and Challenges', *Review of Research in Education*, vol. 44, no. 1, pp. 130-60.

Fisher, RA 1922, 'On the Interpretation of X 2 from Contingency Tables, and the Calculation of P', *Journal of the royal statistical society*, vol. 85, no. 1, pp. 87-94.

Fix, E & Hodges, JL 1989, 'Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties', *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238-47.

Flinders University 2018, *Topics Mapped to Colleges*, Flinders University, viewed 17 Feb 2024, <https://staff.flinders.edu.au/colleges-and-services/college-topics-staff>.

Frank, E 2015, *Re: [Wekalist] Description of Reptree Result*, The University of Waikato, viewed 17 Feb 2024.

Frank, E, Hall, MA & Witten, IH 2016, 'The Weka Workbench', in *Data Mining: Practical Machine Learning Tools and Techniques*, 4th edn, Morgan Kaufmann Publishers Inc.

Freund, Y & Schapire, RE 1996, 'Experiments with a New Boosting Algorithm', in M Kaufmann (ed.), *Proceedings of the 13th International Conference on Machine Learning (ICML 1996)*, Desenzano sul Garda, Italy, pp. 148-56.

Freund, Y & Schapire, RE 1997, 'A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting', *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-39.

Gaff, JG & Wilson, RC 1971, 'Faculty Cultures and Interdisciplinary Studies', *The Journal of Higher Education*, vol. 42, no. 3, pp. 186-201.

Gajwani, J & Chakraborty, P 2020, 'Students' Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms', in *International Conference on Innovative Computing and Communications Proceedings of ICICC 2020*, Singapore, vol. 1, pp. 347-54.

Gao, Y, Wong, SL, Khambari, MNM & Noordin, N 2022, 'A Bibliometric Analysis of the Scientific Production of E-Learning in Higher Education (1998-2020)', *International Journal of Information and Education Technology*, vol. 12, no. 5, pp. 390-9.

Gardner, H 1983, *Frames of Mind: The Theory of Multiple Intelligences*, Harper Colophon Books, Basic Books, New York, NY.

Garrison, DR, Anderson, T & Archer, W 1999, 'Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education', *The Internet and Higher Education*, vol. 2, no. 2, pp. 87-105.

Garrison, DR & Cleveland-Innes, M 2005, 'Facilitating Cognitive Presence in Online Learning: Interaction Is Not Enough', *The American Journal of Distance Education*, vol. 19, no. 3, pp. 133-48.

Gašević, D, Dawson, S & Siemens, G 2015, 'Let's Not Forget: Learning Analytics Are About Learning', *TechTrends*, vol. 59, no. 1, pp. 64-71.

Goldie, JGS 2016, 'Connectivism: A Knowledge Learning Theory for the Digital Age?', *Medical teacher*, vol. 38, no. 10, pp. 1064-9.

Grafinger, DJ 1988, *Basics of Instructional Systems Development, Info-Line Issue 8803*, American Society for Training and Development, Alexandria, VA.

Grinsztajn, L, Oyallon, E & Varoquaux, G 2022, 'Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?', in S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho & A Oh (eds), *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, New Orleans, LA, pp. 507-20.

Hamadneh, NN, Atawneh, S, Khan, WA, Almejalli, KA & Alhomoud, A 2022, 'Using Artificial Intelligence to Predict Students' Academic Performance in Blended Learning', *Sustainability*, vol. 14, no. 18, p. 11642.

Han, J, Kamber, M & Pei, J 2012, *Data Mining Concepts and Techniques*, 3rd edn, University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University, Morgan Kaufmann Publishers Inc., Waltham, MA.

Hanley, JA & McNeil, BJ 1982, 'The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve', *Radiology*, vol. 143, no. 1, pp. 29-36.

Harasim, L 2017, 'Learning Theories: The Role of Epistemology, Science, and Technology', in *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*, Springer International Publishing, pp. 1-39.

Hassanzadeh, A, Kanaani, F & Elahi, S 2012, 'A Model for Measuring E-Learning Systems Success in Universities', *Expert Systems with Applications*, vol. 39, no. 12, pp. 10959-66.

He, H & Garcia, EA 2009, 'Learning from Imbalanced Data', *IEEE transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263-84.

Heinrich, E & Carvalho, L 2022, 'Fostering Connections for Professional Identity Formation: Two Case Studies of Discord Discipline-Focused Communities', in *ASCILITE 2022 Conference Proceedings: Reconnecting relationships through technology*, pp. e22136-e.

Henrie, CR, Halverson, LR & Graham, CR 2015, 'Measuring Student Engagement in Technology-Mediated Learning: A Review', *Computers & Education*, vol. 90, pp. 36-53.

Hill, P 2022, *State of Higher Ed Lms Market for Us and Canada: Year-End 2021 Edition*, Phil Hill & Associates, viewed 17 Feb 2024, <https://philhillaa.com/onedtech/state-of-higher-ed-lms-market-for-us-and-canada-year-end-2021-edition/>.

Ho, AD, Blair Justin Fire Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Philip Mullaney, James H. Waldo & Chuang, aI 2014, 'Harvardx and Mitx: The First Year of Open Online Courses, Fall 2012-Summer 2013', *HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1)*, pp. 1-33.

Hochberg, Y 1988, 'A Sharper Bonferroni Procedure for Multiple Tests of Significance', *Biometrika*, vol. 75, no. 4, pp. 800-2.

Hodges, CB, Moore, S, Lockee, BB, Trust, T & Bond, MA 2020, *The Difference between Emergency Remote Teaching and Online Learning*, EDUCAUSE Review, viewed 17 Feb 2024, <https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning>.

Holm, S 1979, 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian journal of statistics*, vol. 6, no. 2, pp. 65-70.

Horne, JA & Ostberg, O 1976, 'A Self-Assessment Questionnaire to Determine Morningness-Eveningness in Human Circadian Rhythms', *International journal of chronobiology*, vol. 4, no. 2, pp. 97-110.

Huang, CL, Luo, YF, Yang, SC, Lu, CM & Chen, A-S 2019, 'Influence of Students' Learning Style, Sense of Presence, and Cognitive Load on Learning Outcomes in an Immersive Virtual Reality Learning Environment', *Journal of educational computing research*, vol. 58, no. 3, pp. 596-615.

Huang, T-C, Chen, M-Y & Hsu, W-P 2019, 'Do Learning Styles Matter? Motivating Learners in an Augmented Geopark', *Journal of Educational Technology and Society*, vol. 22, no. 1, pp. 70-81.

Iba, W & Langley, P 1992, 'Induction of One-Level Decision Trees', in D Sleeman & P Edwards (eds), *Machine Learning Proceedings 1992*, San Francisco, CA, pp. 233-40.

Imel, S 1998, *Distance Learning. Myths and Realities*, ERIC Clearinghouse on Adult, Career, and Vocational Education, Columbus, OH.

Ingwersen, H 2016, *The Top 8 Free/Open Source Lmss*, Capterra, viewed 17 Feb 2024, <https://medium.com/@CapterraLMS/the-top-8-free-open-source-lmss-4d6432486f8>.

Irwanto, I, Wahyudiati, D, Saputro, AD & Lukman, IR 2023, 'Massive Open Online Courses (Moocs) in Higher Education: A Bibliometric Analysis (2012-2022)', *International Journal of Information and Education Technology (IJIET)*, vol. 13, no. 2, pp. 223-31.

Ithriah, S, Ridwandono, D & Suryanto, T 2020, 'Online Learning Self-Efficacy: The Role in E-Learning Success', in *Journal of Physics: Conference Series*, vol. 1569, p. 022053.

Jayaprakash, SM, Moody, EW, Lauría, EJ, Regan, JR & Baron, JD 2014, 'Early Alert of Academically at-Risk Students: An Open Source Analytics Initiative', *Journal of Learning Analytics*, vol. 1, no. 1, pp. 6-47.

Jia, K, Wang, P, Li, Y, Chen, Z, Jiang, X, Lin, C-L & Chin, T 2022, 'Research Landscape of Artificial Intelligence and E-Learning: A Bibliometric Research', *Frontiers in psychology*, vol. 13, p. 795039.

Jia, Y, Song, Z, Bai, X & Xu, W 2017, 'Towards Economic Models for Mooc Pricing Strategy Design', in *International Conference on Database Systems for Advanced Applications*, pp. 387-98.

John, GH & Langley, P 1995, 'Estimating Continuous Distributions in Bayesian Classifiers', in *Proceedings of the 11th conference on Uncertainty in artificial intelligence*, Montréal, Qué, Canada, pp. 338–45.

Kahu, ER 2013, 'Framing Student Engagement in Higher Education', *Studies in Higher Education*, vol. 38, no. 5, pp. 758-73.

Kahu, ER, Thomas, HG & Heinrich, E 2022, ''A Sense of Community and Camaraderie': Increasing Student Engagement by Supplementing an Lms with a Learning Commons Communication Tool', *Active Learning in Higher Education*, vol. 0, no. 0, p. 14697874221127691.

Kaiss, W, Mansouri, K & Poirier, F 2023, 'Effectiveness of an Adaptive Learning Chatbot on Students' Learning Outcomes Based on Learning Styles', *International Journal of Emerging Technologies in Learning (Online)*, vol. 18, no. 13, pp. 250-61.

Kanchon, MKH, Sadman, M, Nabila, KF, Tarannum, R & Khan, R 2024, 'Enhancing Personalized Learning: Ai-Driven Identification of Learning Styles and Content Modification Strategies', *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 269-78.

Kaptelinin, V, Kuutti, K & Bannon, L 1995, 'Activity Theory: Basic Concepts and Applications: A Summary of a Tutorial Given at the East West Hci 95 Conference', in *International Conference on Human-Computer Interaction, Berlin*.

Keegan, D 1995, *Distance Education Technology for the New Millennium Compressed Video Teaching.* , Zentrales Institut für Fernstudienforschung, Hagen, 1435-9340, <https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00000342>.

Kelleher, JD, Mac Namee, B & D'arcy, A 2015, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, The MIT Press Cambridge, Massachusetts.

Khan, BH & Ally, M 2015, *International Handbook of E-Learning Volume 1: Theoretical Perspectives and Research*, Taylor & Francis, New York, NY.

Khan, FM & Gupta, Y 2022, 'A Bibliometric Analysis of Mobile Learning in the Education Sector', *Interactive Technology and Smart Education*, vol. 19, no. 3, pp. 338-59.

Khan, MA & Salah, K 2020, 'Cloud Adoption for E-Learning: Survey and Future Challenges', *Education and Information Technologies*, vol. 25, no. 2, pp. 1417-38.

Kika, A, Leka, L, Maxhelaku, S & Ktona, A 2019, 'Using Data Mining Techniques on Moodle Data for Classification of Student's Learning Styles', in *47th International Academic Conference*, Prague, Czech Republic, pp. 26-33.

King, FB, Young, MF, Drivere-Richmond, K & Schrader, P 2001, 'Defining Distance Learning and Distance Education', *AACE Review (Formerly AACE Journal)*, vol. 9, no. 1, pp. 1-14.

Kirkwood, A & Price, L 2013, 'Missing: Evidence of a Scholarly Approach to Teaching and Learning with Technology in Higher Education', *Teaching in Higher Education*, vol. 18, no. 3, pp. 327-37.

Kizilcec, RF, Piech, C & Schneider, E 2013, 'Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses', in *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, Leuven, Belgium, pp. 170-9.

Kohavi, R 1996, 'Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid', in E Simoudis, J Han & U Fayyad (eds), *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR, vol. 96, pp. 202-7.

Kolb, DA 1981, 'Learning Styles and Disciplinary Differences', *In A. Chickering (Ed.), The Modern American College. San Francisco: Jossey-Bass.*, vol. 1, no. January 1981, pp. 232-5.

Kolmos, A, Hadgraft, RG & Holgaard, JE 2016, 'Response Strategies for Curriculum Change in Engineering', *International Journal of Technology and Design Education*, vol. 26, no. 3, pp. 391-411.

Krizhevsky, A, Sutskever, I & Hinton, GE 2012, 'Imagenet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, vol. 25, pp. 1097-105.

Krouska, A, Troussas, C, Virvou, M & Fragkakis, CK 2018, 'Applying Skinnerian Conditioning for Shaping Skill Performance in Online Tutoring of Programming Languages', in *9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1-5.

Kruger, J-L & Doherty, S 2016, 'Measuring Cognitive Load in the Presence of Educational Video: Towards a Multimodal Methodology', *Australasian Journal of Educational Technology*, vol. 32, no. 6, pp. 19-31.

Kruskal, WH & Wallis, WA 1952, 'Use of Ranks in One-Criterion Variance Analysis', *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583-621.

Lang, C, Siemens, G, Wise, A & Gasevic, D 2017, *Handbook of Learning Analytics*, SOLAR, Society for Learning Analytics and Research New York.

Lange, C 2023, 'The Relationship between E-Learning Personalisation and Cognitive Load', *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 38, no. 3, pp. 228-42.

Latip, MSA, Tamrin, M, Noh, I, Rahim, FA, Nur, S & Latip, NA 2022, 'Factors Affecting E-Learning Acceptance among Students: The Moderating Effect of Self-Efficacy', *International Journal of Information and Education Technology*, vol. 12, no. 2, pp. 116-22.

Lave, J & Wenger, E 1991, *Situated Learning: Legitimate Peripheral Participation*, Learning in Doing: Social, Cognitive and Computational Perspectives, Cambridge University Press, Cambridge.

Leach, RJ 2018, *Introduction to Software Engineering*, 2nd edn, CRC Press.

Lebron, D & Shahriar, H 2015, 'Comparing Mooc-Based Platforms: Reflection on Pedagogical Support, Framework and Learning Analytics', in *Collaboration Technologies and Systems (CTS), 2015 International Conference on*, pp. 167-74.

Li, CL & Abidin, MJBZ 2024, 'Instructional Design of Classroom Instructional Skills Based on the Addie Model Education', *Technium Soc. Sci. J.*, vol. 55, p. 167.

Maatuk, AM, Elberkawi, EK, Aljawarneh, S, Rashaideh, H & Alharbi, H 2022, 'The Covid-19 Pandemic and E-Learning: Challenges and Opportunities from the Perspective of Students and Instructors', *Journal of Computing in Higher Education*, vol. 34, no. 1, pp. 21-38.

MacQueen, J 1967, 'Some Methods for Classification and Analysis of Multivariate Observations', in *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281-97.

Maheshwari, G 2021, 'Factors Affecting Students' Intentions to Undertake Online Learning: An Empirical Study in Vietnam', *Education and Information Technologies*, vol. 26, no. 6, p. 6629.

Mallillin, LLD, Mendoza, LC, Mallillin, JB, Felix, RC & Lipayon, IC 2020, 'Implementation and Readiness of Online Learning Pedagogy: A Transition to Covid 19 Pandemic', *European Journal of Open Education and E-learning Studies*, vol. 5, no. 2, pp. 71-90.

Marbouti, F, Diefes-Dux, HA & Madhavan, K 2016, 'Models for Early Prediction of at-Risk Students in a Course Using Standards-Based Grading', *Computers & Education*, vol. 103, pp. 1-15.

Margaryan, A, Bianco, M & Littlejohn, A 2015, 'Instructional Quality of Massive Open Online Courses (Moocs)', *Computers & Education*, vol. 80, pp. 77-83.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Witten, IH 2009, 'The Weka Data Mining Software: An Update', *SIGKDD Explorations*, vol. 11, no. 1, pp. 10-8.

Marras, A 1983, 'Review: [Untitled] - Reviewed Work: Language and Learning: The Debate between Jean Piaget and Noan Chomsky', review of Language and Learning: The Debate between Jean Piaget and Noan Chomsky, Massimo Piattelli-Palmarini, *Canadian Journal of Philosophy*, vol. 13, no. 2, pp. 277-91.

Mattar, J 2018, 'Constructivism and Connectivism in Education Technology: Active, Situated, Authentic, Experiential, and Anchored Learning', *Revista Iberoamericana de Educación a Distancia (RIED)*, vol. 21, no. 2, pp. 201-17.

Matthews, BW 1975, 'Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme', *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442-51.

Matzavela, V & Alepis, E 2021, 'Decision Tree Learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning Environments', *Computers & Education: Artificial Intelligence*, vol. 2, no. 6, p. 100035.

Mayer, RE 2009, 'The Promise of Multimedia Learning', in *Multimedia Learning*, 2nd edn, Cambridge University Press, New York, NY, p. 304.

Mayer, RE & Moreno, R 2003, 'Nine Ways to Reduce Cognitive Load in Multimedia Learning', *Educational psychologist*, vol. 38, no. 1, pp. 43-52.

Mayes, T & de Freitas, S 2007, 'Learning and E-Learning: The Role of Theory', in *Rethinking Pedagogy for a Digital Age*, Routledge, Abingdon, Oxon, pp. 33-45.

McCulloch, WS & Pitts, W 1943, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115-33.

Meland, PH & Gjære, EA 2012, 'Representing Threats in Bpmn 2.0', in *Seventh International Conference on Availability, Reliability and Security*, pp. 542-50.

Merrill, MD 2002, 'A Pebble-in-the-Pond Model for Instructional Design', *Performance Improvement*, vol. 41, no. 7, pp. 41-6.

Microsoft 2016, *Microsoft Excel*, 2016 edn, Redmond, Washington, Computer Software.

Microsoft 2018a, *Microsoft Sql Server*, 2017 edn, Microsoft, Redmond, Washington, Computer Software, <https://www.microsoft.com/en-us/sql-server/sql-server-downloads>.

Microsoft 2018b, *Sql Server Management Studio*, v17.8.1 edn, Microsoft, Redmond, Washington, Computer Software, <https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16>.

Microsoft 2022, *Datepart (Transact-Sql)*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/t-sql/functions/datepart-transact-sql?view=sql-server-2017>.

Microsoft 2023a, *Case (Transact-Sql)*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql?view=sql-server-2017>.

Microsoft 2023b, *Cast and Convert (Transact-Sql)*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/t-sql/functions/cast-and-convert-transact-sql?view=sql-server-2017>.

Microsoft 2023c, *Dateadd (Transact-Sql)*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/t-sql/functions/dateadd-transact-sql?view=sql-server-2017>.

Microsoft 2023d, *Editions and Supported Features of Sql Server 2017*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/sql-server/editions-and-components-of-sql-server-2017?view=sql-server-2017>.

Microsoft 2023e, *Left (Transact-Sql)*, Microsoft, viewed 17 Feb 2024, <https://docs.microsoft.com/en-us/sql/t-sql/functions/left-transact-sql?view=sql-server-2017>.

Mikić, V, Ilić, M, Kopanja, L & Vesin, B 2022, 'Personalisation Methods in E-Learning-a Literature Review', *Computer Applications in Engineering Education*, vol. 30, no. 6, pp. 1931-58.

MITx and HarvardX 2014, *Harvardx-Mitx Person-Course Academic Year 2013 De-Identified Dataset, Version 2.0*, <http://dx.doi.org/10.7910/DVN/26147>.

Mohebi, L 2021, 'Theoretical Models of Integration of Interactive Learning Technologies into Teaching: A Systematic Literature Review', *International Journal of Learning, Teaching and Educational Research*, vol. 20, no. 12, pp. 232-54.

Molenda, M 2003, 'In Search of the Elusive Addie Model', *Performance Improvement*, vol. 42, no. 5, p. 35.

Montgomery, DC 2017, *Design and Analysis of Experiments*, 9th edn, John Wiley & Sons, Inc., Hoboken, NJ.

Moore, JL, Dickson-Deane, C & Galyen, K 2011, 'E-Learning, Online Learning, and Distance Learning Environments: Are They the Same?', *The Internet and Higher Education*, vol. 14, no. 2, pp. 129-35.

Morris, LV, Finnegan, C & Wu, S-S 2005, 'Tracking Student Behavior, Persistence, and Achievement in Online Courses', *The Internet and Higher Education*, vol. 8, no. 3, pp. 221-31.

Morrison, GR, Ross, SJ, Kalman, HK & Kemp, JE 2019, *Designing Effective Instruction.*, 8th edn, John Wiley & Sons, Inc., Hoboken, NJ.

Moubayed, A, Injadat, M, Nassif, AB, Lutfiyya, H & Shami, A 2018, 'E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics', *IEEE Access*, vol. 6, pp. 39117-38.

Moule, P 2007, 'Challenging the Five-Stage Model for E-Learning: A New Approach', *ALT-J*, vol. 15, no. 1, pp. 37-50.

Mseleku, Z 2020, 'A Literature Review of E-Learning and E-Teaching in the Era of Covid-19 Pandemic', *International Journal of Innovative Science and Research Technology*, vol. 5, no. 10, pp. 588-97.

Munir, M & Waty, TK 2023, 'The Influence of Self Innovativeness and Self Efficacy on E-Learning Implementation Effectiveness', *International Journal of Service Science, Management, Engineering, and Technology*, vol. 3, no. 1, pp. 1-5.

Nichols, M 2003, 'A Theory for Elearning', *Journal of Educational Technology & Society*, vol. 6, no. 2, pp. 1-10.

Nicholson, P 2007, 'A History of E-Learning', in B Fernández-Manjón, JM Sánchez-Pérez, JA Gómez-Pulido, MA Vega-Rodríguez & J Bravo-Rodríguez (eds), *Computers & Education: E-Learning, from Theory to Practice*, Springer Netherlands, Dordrecht, The Netherlands, pp. 1-11.

Njenga, JK & Fourie, LCH 2010, 'The Myths About E-Learning in Higher Education', *British Journal of Educational Technology*, vol. 41, no. 2, pp. 199-212.

Oliveira, W, Hamari, J, Shi, L, Toda, AM, Rodrigues, L, Palomino, PT & Isotani, S 2023, 'Tailored Gamification in Education: A Literature Review and Future Agenda', *Education and Information Technologies*, vol. 28, no. 1, pp. 373-406.

Ouadoud, M, Rida, N & Chafiq, T 2021, 'Overview of E-Learning Platforms for Teaching and Learning', *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 9, no. 1, p. 50.

Papastergiou, M 2009, 'Digital Game-Based Learning in High School Computer Science Education: Impact on Educational Effectiveness and Student Motivation', *Computers & Education*, vol. 52, no. 1, pp. 1-12.

Pappas, C 2015, *The Insider's Guide to Learning Management Systems' Pricing Models*, eLearning Industry, viewed 17 Feb 2024, <https://elearningindustry.com/learning-management-systems-pricing-models-insiders-guide>.

Pardos, ZA & Heffernan, NT 2011, 'Kt-Idem: Introducing Item Difficulty to the Knowledge Tracing Model', in *User Modeling, Adaption and Personalization: 19th International Conference (UMAP 2011)*, Girona, Spain, pp. 243-54.

Patwari, MB, Dubey, S & Jagdale, MSV 2023, 'Impact and Effectiveness of Online Teaching Methods and Pedagogy', *The Online Journal of Distance Education and e-Learning*, vol. 11, no. 2, pp. 1417-24.

Pavlov, I & Thompson, WH 1902, *The Work of the Digestive Glands*, 1st edn, Scientific and Medical Knowledge Production, 1796-1918, Charles Griffin & Company, Ltd., London, UK.

Paz, FJ & Cazella, SC 2019, 'Academic Analytics: A Systematic Review of Literature', *International Journal of Development Research*, vol. 9, no. 11, pp. 31710-6.

Pearson, K 1895, 'Vii. Note on Regression and Inheritance in the Case of Two Parents', *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240-2.

Pearson, K 1900, 'On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling', *Philosophical Magazine*, vol. 50, no. 302, pp. 157-75.

Pearson, K 1901a, 'Liii. On Lines and Planes of Closest Fit to Systems of Points in Space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559-72.

Pearson, K 1901b, 'On the Mathematical Theory of Errors of Judgement with Special Reference to the Personal Equation', *Proceedings of the Royal Society of London*, vol. 68, no. 442-450, pp. 369-72.

Perner, P 2011, 'How to Interpret Decision Trees?', in *11th Industrial Conference, ICDM 2011*, New York, NY, pp. 40-55.

Peters, O 1973, *Die Didaktische Struktur Des Fernunterrichts: Untersuchungen Zu Einer Industrialisierten Form Des Lehrens Und Lernens*, Tübinger Beiträge Zum Fernstudium, Verlag Julius Beltz, Weinheim, Germany.

Piaget, J & Inhelder, B 1967, *The Child's Conception of Space*, Routledge & Kegan Paul, London, UK.

Picciano, AG 2002, 'Beyond Student Perceptions: Issues of Interaction, Presence, and Performance in an Online Course', *Journal of Asynchronous learning networks*, vol. 6, no. 1, pp. 21-40.

Pinner, R 2014, *What Is the Difference between an Lms and a Vle?*, eLearning Industry, viewed 17 Feb 2024, <https://elearningindustry.com/difference-between-lms-and-vle>.

Piskurich, GM 2015, *Rapid Instructional Design: Learning Id Fast and Right*, 3rd edn, John Wiley & Sons, Inc., Hoboken, NJ.

Porcaro, D 2011, 'Applying Constructivism in Instructivist Learning Cultures', *Multicultural Education & Technology Journal*, vol. 5, no. 1, pp. 39-54.

Porter, S 2015, 'The Economics of Moocs: A Sustainable Future?', *The Bottom Line*, vol. 28, no. 1/2, pp. 52-62.

Powers, DMW 2003, 'Recall and Precision Versus the Bookmaker', in *Proceedings of the International Conference on Cognitive Science (ICSC-2003)*, Sydney, Australia, pp. 529-34.

Powers, DMW 2011, 'Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness and Correlation', *International Journal of Machine Learning Technology*, vol. 2, no. 1, pp. 37-63.

Powers, DMW 2012a, 'The Problem with Kappa', in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 345-55.

Powers, DMW 2012b, 'Roc-Concert: Roc-Based Measurement of Consistency and Certainty', in *2012 spring congress on engineering and technology*, pp. 1-4.

Powers, DMW 2015, 'What the F-Measure Doesn't Measure: Features, Flaws, Fallacies and Fixes', *arXiv preprint arXiv:1503.06410*.

Prahani, BK, Alfin, J, Fuad, AZ, Saphira, HV, Hariyono, E & Suprapto, N 2022, 'Learning Management System (Lms) Research During 1991-2021: How Technology Affects Education', *International Journal of Emerging Technologies in Learning (Online)*, vol. 17, no. 17, p. 28.

Prioteasa, A-L, Ciocoiu, CN, Lazăr, L & Minciu, M 2023, 'E-Learning in Higher Education During the Covid-19 Pandemic: A Bibliometric Analysis', *Proceedings of the International Conference on Business Excellence*, vol. 17, no. 1, pp. 1858-72.

Provost, F & Fawcett, T 2013, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, 1st edn, O'Reilly Media, Inc., Sebastapol, CA.

Qiu, F, Zhang, G, Sheng, X, Jiang, L, Zhu, L, Xiang, Q, Jiang, B & Chen, P-k 2022, 'Predicting Students' Performance in E-Learning Using Learning Process and Behaviour Data', *Scientific Reports*, vol. 12, no. 1, p. 453.

Qiu, F, Zhu, L, Zhang, G, Sheng, X, Ye, M, Xiang, Q & Chen, P-K 2022, 'E-Learning Performance Prediction: Mining the Feature Space of Effective Learning Behavior', *Entropy (Basel)*, vol. 24, no. 5, p. 722.

Quinlan, JR 1986, 'Induction of Decision Trees', *Machine learning*, vol. 1, pp. 81-106.

Quinlan, R 1993, *C4.5: Programs for Machine Learning*, Representation and Reasoning Series, Morgan Kaufmann Publishers Inc., San Mateo, CA.

R Core Team 2023, *R: A Language and Environment for Statistical Computing*, 4.2.3 (2023-03-15 ucrt) edn, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.

Ragnedda, M & Muschert, GW 2013, *The Digital Divide: The Internet and Social Inequality in International Perspective*, Routledge, Abingdon, Oxon.

Ramaswami, G, Susnjak, T & Mathrani, A 2023, 'Effectiveness of a Learning Analytics Dashboard for Increasing Student Engagement Levels', *Journal of Learning Analytics*, vol. 10, no. 3, pp. 115-34.

Ramos, C & Yudko, E 2008, '"Hits"(Not "Discussion Posts") Predict Student Success in Online Courses: A Double Cross-Validation Study', *Computers & Education*, vol. 50, pp. 1174-82.

Ranjeeth, S, Latchoumi, TP & Paul, PV 2020, 'A Survey on Predictive Models of Learning Analytics', *Procedia Computer Science*, vol. 167, pp. 37-46.

Rankapola, ME & Zuva, T 2023, 'The effect of e-Learning quality, self-Efficacy and e-Learning satisfaction on the Students' Intention to Use the E-Learning System', in R Silhavy & P Silhavy (eds), *Artificial Intelligence Application in Networks and Systems*, Cham, pp. 640-53.

Rasheed, F & Wahid, A 2021, 'Learning Style Detection in E-Learning Systems Using Machine Learning Techniques', *Expert Systems with Applications*, vol. 174, p. 114774.

Rodrigues, MW, Isotani, S & Zárate, LE 2018, 'Educational Data Mining: A Review of Evaluation Process in the E-Learning', *Telematics and Informatics*, vol. 35, no. 6, pp. 1701-17.

Rodriguez, JJ, Kuncheva, LI & Alonso, CJ 2006, 'Rotation Forest: A New Classifier Ensemble Method', *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619-30.

Roenneberg, T, Wirz-Justice, A & Merrow, M 2003, 'Life between Clocks: Daily Temporal Patterns of Human Chronotypes', *Journal of biological rhythms*, vol. 18, no. 1, pp. 80-90.

Rogers, C 2023, *Impact of Covid-19 on the Edtech Sector*, Deloitte Australia, viewed 17 Feb 2024, <https://www.deloitte.com/au/en/Industries/government-public/analysis/australian-edtech-market-census.html>.

Romero, C & Ventura, S 2020, 'Educational Data Mining and Learning Analytics: An Updated Survey', *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, vol. 10, no. 3, p. e1355.

Romsi, A, Widodo, JP & Slamet, J 2024, 'Empowering Slow Learners: Gamification's Impact on Students' Engagement and Academic Performance in an Lms for Undergraduate Students', *International Journal of Information and Education Technology*, vol. 14, no. 2.

Rosenblatt, F 1957, *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, vol. 85-460-1, Report: Cornell Aeronautical Laboratory, Cornell Aeronautical Laboratory, Ithaca, New York.

Rotar, O 2022, 'Online Student Support: A Framework for Embedding Support Interventions into the Online Learning Cycle', *Research and Practice in Technology Enhanced Learning*, vol. 17, no. 1, p. 2.

RStudio Team 2023, *Rstudio: Integrated Development Environment for R*, Ocean Storm Release (33206f75, 2023-12-17) for windows edn, Posit Software, PBC, Boston, MA, <https://posit.co/>.

Rumelhart, DE, Hinton, GE & Williams, RJ 1986, 'Learning Representations by Back-Propagating Errors', *nature*, vol. 323, no. 6088, pp. 533-6.

Runkler, TA 2020, *Data Analytics: Models and Algorithms for Intelligent Data Analysis*, 3rd edn, Springer Vieweg, Wiesbaden, Germany.

Sandeen, C 2015, 'Balance Sheet on Moocs: Myth, Hype and Potential', *Asian Journal of the Scholarship of Teaching and Learning*, vol. 5, no. 1, pp. 9-22.

Sathe, MT & Adamuthe, AC 2021, 'Comparative Study of Supervised Algorithms for Prediction of Students' Performance', *International Journal of Modern Education & Computer Science*, vol. 13, no. 1, pp. 1-21.

Schniederjans, MJ, Schniederjans, DG & Starkey, CM 2014, *Business Analytics Principles, Concepts, and Applications: What, Why, and How*, Pearson Education, Inc., Upper Saddle River, NJ.

Setiawan, R, Arif, FAS, Putro, JO, Princes, E, Silalahi, FTR, Geraldina, I, Julianti, E & Safitri, J 2023, 'E-Learning Pricing Model Policy for Higher Education', *IEEE Access*, vol. 11, pp. 38370-84.

Sghir, N, Adadi, A & Lahmer, M 2023, 'Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012–2022)', *Education and Information Technologies*, vol. 28, no. 7, pp. 8299-333.

Shah, D 2014, *Moocs in 2014: Breaking Down the Numbers*, EdSurge News, viewed 17 Feb 2024, <https://www.edsurge.com/news/2014-12-26-moocs-in-2014-breaking-down-the-numbers>.

Sheeba, T & Krishnan, R 2019, 'Automatic Detection of Students Learning Style in Learning Management System', in A Al-Masri & K Curran (eds), *Smart Technologies and Innovation for*

a Sustainable Future, Proceedings of the 1st American University in the Emirates International Research Conference, Dubai, UAE, pp. 45-53.

Sherry, L 1996, 'Issues in Distance Learning', *International journal of educational telecommunications*, vol. 1, no. 4, pp. 337-65.

Siemens, G 2004, 'Connectivism: A Learning Theory for the Digital Age', *International Journal of Instructional Technology and Distance Learning*, vol. 2, no. 1, pp. 3-10.

Siemens, G & Long, P 2011, 'Penetrating the Fog: Analytics in Learning and Education', *Educause Review*, vol. 46, no. 5, p. 30.

Sims, R & Jones, D 2002, 'Continuous Improvement through Shared Understanding: Reconceptualising Instructional Design for Online Learning', in *Winds of Changing in the Sea of Learning, Proceedings of the 19th Annual Conference of the Australian Society for Computers in Tertiary Education (ASCILITE)*, Auckland, New Zealand, pp. 623-32.

Singh, P, Alhassan, I, Binsaif, N & Alhussain, T 2023, 'Standard Measuring of E-Learning to Assess the Quality Level of E-Learning Outcomes: Saudi Electronic University Case Study', *Sustainability*, vol. 15, no. 1, p. 844.

Singh, V & Thurman, A 2019, 'How Many Ways Can We Define Online Learning? A Systematic Literature Review of Definitions of Online Learning (1988-2018)', *The American Journal of Distance Education*, vol. 33, no. 4, pp. 289-306.

Skinner, BF 1965, *Science and Human Behavior*, 1st edn, Psychology, The Free Press, New York, NY.

Souza, VF, Cicalese, F, Laber, E & Molinaro, M 2022, 'Decision Trees with Short Explainable Rules', in S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho & A Oh (eds), *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, vol. 35, pp. 12365-79.

Spatioti, AG, Kazanidis, I & Pange, J 2022, 'A Comparative Study of the Addie Instructional Design Model in Distance Education', *Information*, vol. 13, no. 9, p. 402.

Spyropoulou, N, Pierrakeas, C & Kameas, A 2014, 'Creating Mooc Guidelines Based on Best Practices', in *EDULEARN14 Proceedings*, pp. 6981-90.

Srinivasa, KG, Kurni, M & Saritha, K 2022, 'Pedagogy for E-Learning', in *Learning, Teaching, and Assessment Methods for Contemporary Learners: Pedagogy for the Digital Generation*, Springer Nature Singapore Pty Ltd., Singapore, pp. 283-309.

Stewart, GW 1993, 'On the Early History of the Singular Value Decomposition', *SIAM Review*, vol. 35, no. 4, pp. 551-66.

Stobart, A & Duckett, S 2022, 'Australia's Response to Covid-19', *Health Economics, Policy and Law*, vol. 17, no. 1, pp. 95-106.

Student 1908, 'The Probable Error of a Mean', *Biometrika*, vol. 6, no. 1, pp. 1-25.

Subiyantoro, S, Degeng, INS, Kuswandi, D & Ulfa, S 2024, 'Developing Gamified Learning Management Systems to Increase Student Engagement in Online Learning Environments', *International Journal of Information and Education Technology*, vol. 14, no. 1.

Sun, P-C, Tsai, RJ, Finger, G, Chen, Y-Y & Yeh, D 2008, 'What Drives a Successful E-Learning? An Empirical Investigation of the Critical Factors Influencing Learner Satisfaction', *Computers & Education*, vol. 50, no. 4, pp. 1183-202.

Sweller, J, van Merriënboer, JJG & Paas, F 2019, 'Cognitive Architecture and Instructional Design: 20 years Later', *Educational Psychology Review*, vol. 31, no. 2, pp. 261-92.

Taneja, S & Goel, A 2014, 'Mooc Providers and Their Strategies', *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 222-8.

Thai-Nghe, N, Drumond, L, Krohn-Grimberghe, A & Schmidt-Thieme, L 2010, 'Recommender System for Predicting Student Performance', *Procedia Computer Science*, vol. 1, no. 2, pp. 2811-9.

The MathWorks Inc. 2022, *Matlab Version: 9.12.0.1975300 (R2022a)*, Natick, Massachusetts, <https://au.mathworks.com/products/matlab.html>.

Tukey, JW 1977, 'Exploratory Data Analysis', in *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, vol. 2.

Tukey, JW 1993, *Exploratory Data Analysis: Past, Present and Future*, Defense Technical Information Center, Princeton, NJ.

Turnbull, D, Chugh, R & Luck, J 2021, 'Transitioning to E-Learning During the Covid-19 Pandemic: How Have Higher Education Institutions Responded to the Challenge?', *Education and Information Technologies*, vol. 26, no. 5, pp. 6401-19.

Ujma, PP, Baudson, TG, Bódizs, R & Dresler, M 2020, 'The Relationship between Chronotype and Intelligence: The Importance of Work Timing', *Scientific Reports*, vol. 10, no. 1, p. 7105.

Vaicondam, Y, Sikandar, H, Irum, S, Khan, N & Qureshi, MI 2022, 'Research Landscape of Digital Learning over the Past 20 Years: A Bibliometric and Visualisation Analysis', *International journal of online and biomedical engineering*, vol. 18, no. 8, pp. 4-22.

Vaidya, R & Joshi, M 2018, 'Use of Learning Style Based Approach in Instructional Delivery', in X-S Yang, AK Nagar & A Joshi (eds), *Smart Trends in Systems, Security and Sustainability*, Singapore, pp. 199-209.

Valentine, D 2002, 'Distance Learning: Promises, Problems, and Possibilities', *Online Journal of Distance Learning Administration*, vol. 5, no. 3, pp. 1-11.

Valverde-Berrocoso, J, Garrido-Arroyo, MdC, Burgos-Videla, C & Morales-Cevallos, MB 2020, 'Trends in Educational Research About E-Learning: A Systematic Literature Review (2009–2018)', *Sustainability*, vol. 12, no. 12, p. 5153.

Van Merriënboer, JJ & Kirschner, PA 2017, 'Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design', in *21st Century Education: A Reference Handbook*, 1st edn, SAGE Publications Ltd., London, UK, vol. 1, pp. 244-53.

Vapnik, V 2000, *The Nature of Statistical Learning Theory*, 2nd edn, Statistics for Engineering and Information Science, Springer, New York, NY.

Viberg, O, Hatakka, M, Bälter, O & Mavroudi, A 2018, 'The Current Landscape of Learning Analytics in Higher Education', *Computers in Human Behavior*, vol. 89, pp. 98-110.

Vygotski, LS 1929, 'Ii. The Problem of the Cultural Development of the Child', *The Pedagogical seminary and journal of genetic psychology*, vol. 36, no. 3, pp. 415-34.

Wairooy, IK, Reynard, A, Octodhia, MR, Andanu, WM, Suri, PA & Syahputra, ME 2023, 'The Impactful Effect of E-Learning Study Method Towards Students Academic Achievements in General', *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 5, no. 1, pp. 11-4.

Wan Ali, WNA & Wan Yahaya, WAJ 2023, 'Waterfall-Addie Model: An Integration of Software Development Model and Instructional Systems Design in Developing a Digital Video Learning Application', *Asean Journal of Teaching and Learning in Higher Education*, vol. 15, no. 1, pp. 1-28.

Wang, FH 2017, 'An Exploration of Online Behaviour Engagement and Achievement in Flipped Classroom Supported by Learning Management System', *Computers & Education*, vol. 114, pp. 79-91.

Watson, W & Watson, SL 2007, 'An Argument for Clarity: What Are Learning Management Systems, What Are They Not, and What Should They Become', *TechTrends*, vol. 51, no. 2, pp. 28-34.

White, S & Liccardi, I 2006, 'Harnessing Insight into Disciplinary Differences to Refine E-Learning Design', in *Proceedings. Frontiers in Education. 36th Annual Conference*, pp. 5-10.

Wickham, H & Grolemund, G 2016, 'Exploratory Data Analysis', in *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media, Inc., Sebastopol, CA.

Wiggins, G & McTighe, J 2005, *Understanding by Design*, 2nd edn, Association for Supervision and Curriculum Development ASCD, Alexandria, VA.

Wijaya, A, Setiawan, NA & Shapiai, MI 2023, 'Mapping Research Themes and Future Directions in Learning Style Detection Research: A Bibliometric and Content Analysis', *Electronic Journal of E-learning*, vol. 21, no. 4, pp. 274-85.

Wilden, AJ, Shillabeer, A & deVries, D 2017, 'Predicting Success in E-Learning Courses', in *e-Proceeding of the 5th Global Summit on Education: Trends and Challenges in Education*, Kuala Lumpur, Malaysia, pp. 274-81.

Wittmann, M, Dinich, J, Merrow, M & Roenneberg, T 2006, 'Social Jetlag: Misalignment of Biological and Social Time', *Chronobiology International*, vol. 23, no. 1-2, pp. 497-509.

Woolley, DR 1994, *Plato: The Emergence of Online Community*, vol. 3, Social Media Archeology and Poetics, MIT Press, Cambridge, MA.

Wu, W & Plakhtii, A 2021, 'E-Learning Based on Cloud Computing', *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 10, pp. 4-17.

Xie, H, Chu, H-C, Hwang, G-J & Wang, C-C 2019, 'Trends and Development in Technology-Enhanced Adaptive/Personalized Learning: A Systematic Review of Journal Publications from 2007 to 2017', *Computers & Education*, vol. 140, p. 103599.

Yacobson, E, Fuhrman, O, Hershkovitz, S & Alexandron, G 2021, 'De-Identification Is Insufficient to Protect Student Privacy, or–What Can a Field Trip Reveal?', *Journal of Learning Analytics*, vol. 8, no. 2, pp. 83-92.

Yadav, S & Shukla, S 2016, 'Analysis of K-Fold Cross-Validation over Hold-out Validation on Colossal Datasets for Quality Classification', in *IEEE 6th International conference on advanced computing (IACC)*, Bhimavaram, India, pp. 78-83.

Yousef, AMF, Chatti, MA, Schroeder, U & Harald Jakobs, MW 2014, 'A Review of the State-of-the-Art', in *6th International Conference on Computer Supported Education (CSEDU)*, pp. 9-20.

Yu, Q, Yu, K & Li, B 2024, 'Can Gamification Enhance Online Learning? Evidence from a Meta-Analysis', *Education and Information Technologies*, vol. 29, no. 4, pp. 4055-83.

Yuniarti, WD, Winarko, E & Musdholifah, A 2020, 'Data Mining for Student Assessment in E-Leaming: A Survey', in *Fifth International Conference on Informatics and Computing (ICIC)*, pp. 1-6.

Zhang, D, Zhao, JL, Zhou, L & Nunamaker Jr, JF 2004, 'Can E-Learning Replace Classroom Learning?', *Communications of the Acm*, vol. 47, no. 5, pp. 75-9.

Zhou, L, Wu, S, Zhou, M & Li, F 2020, ''School's out, but Class' on', the Largest Online Education in the World Today: Taking China's Practical Exploration During the Covid-19 Epidemic Prevention and Control as an Example', *Best evid chin edu*, vol. 4, no. 2, pp. 501-19.