The Application of Data Mining Techniques to Learning Analytics and its Implications for Interventions with Small Class Sizes

EDWARD WAKELAM

A thesis submitted in partial fulfilment of the requirements of the University of Hertfordshire for the degree of Doctor of Philosophy

This programme of research was carried out in the School of Engineering and Computer Science, University of Hertfordshire

March 2020

Abstract

There has been significant progress in the development of techniques to deliver effective technology enhanced learning systems in education, with substantial progress in the field of learning analytics. These analyses are able to support academics in the identification of students at risk of failure or withdrawal. The early identification of students at risk is critical to giving academic staff and institutions the opportunity to make timely interventions.

This thesis considers established machine learning techniques, as well as a novel method, for the prediction of student outcomes and the support of interventions, including the presentation of a variety of predictive analyses and of a live experiment. It reviews the status of technology enhanced learning systems and the associated institutional obstacles to their implementation and deployment.

Many courses are comprised of relatively small student cohorts, with institutional privacy protocols limiting the data readily available for analysis. It appears that very little research attention has been devoted to this area of analysis and prediction. I present an experiment conducted on a final year university module, with a student cohort of 23, where the data available for prediction is limited to lecture/tutorial attendance, virtual learning environment accesses and intermediate assessments. I apply and compare a variety of machine learning analyses to assess and predict student performance, applied at appropriate points during module delivery. Despite some mixed results, I found potential for predicting student performance in small student cohorts with very limited student attributes, with accuracies comparing favourably with published results using large cohorts and significantly more attributes. I propose that the analyses will be useful to support module leaders in identifying opportunities to make timely academic interventions.

Student data may include a combination of nominal and numeric data. A large variety of techniques are available to analyse numeric data, however there are fewer techniques applicable to nominal data. I summarise the results of what I believe to be a novel technique to analyse nominal data by making a systematic comparison of data pairs.

In this thesis I have surveyed existing intelligent learning/training systems and explored the contemporary AI techniques which appear to offer the most promising contributions to the prediction of student attainment. I have researched and catalogued the organisational and non-technological challenges to be addressed for successful system development and implementation and proposed a set of critical success criteria to apply.

This dissertation is supported by published work.

Declaration

I declare that no part of this work is being submitted concurrently for another award of the University or any other awarding body or institution. This thesis contains a substantial body of work that has not previously been submitted successfully for an award of the University or any other awarding body or institution.

The following parts of this submission have been published previously and/or undertaken as part of a previous degree or research programme:

1. Chapter Three: Sections 3.2, 3.3, Chapter Seven: Sections 7.2.1, 7.2.3, 7.2.5, 7.2.9, 7.2.10, 7.2.11 and Appendix I, contain previously published material from: Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2015. *The potential for using artificial intelligence techniques to improve E-learning systems.* In ECEL 2015 Conference proceedings, pp. 762-770.

2. Chapter Two: Section 2.4.2, Chapter Six: Sections 6.2.3, 6.3.1 and Chapter Seven: Sections 7.3, 7.4.3, contain previously published material from: Wakelam, E., Davey, N., Sun, Y., Jefferies, A., Alva, P. and Hocking, A., 2016, May. *The Mining and Analysis of Data with Mixed Attribute Types*. In Proceedings: IMMM 2016: Sixth International Conference on Advances in Information Mining and Management. IARIA, pp 32-37.

3. Chapter Four, Sections 4.2, 4.4, Chapter Five, 5.2.1, 5.4, Chapter Six: Sections 6.2.5, 6.3.1and Chapter Eight, all sections except section 8.7.1.3, contain previously published material from: Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2020. *The potential for student performance prediction in small cohorts with minimal available attributes*. British Journal of Educational Technology, 51(2), pp. 347-370.

Except where indicated otherwise in the submission, the submission is my own work and has not previously been submitted successfully for any award.

Acknowledgements

I'm very grateful indeed to my lead supervisor Amanda Jefferies for her amazingly positive response to my original contact, an unexpected out of the blue email wondering if there were opportunities for a retired chap like me to follow my research ideas. Amanda's enthusiasm for the subject of my work and her professionalism has steadied me throughout my studies.

My supervision team of Amanda, Neil Davey and Yi Sun has provided me with unfaltering support and encouragement throughout my studies. Their positive coaching helped me to make what was a tricky transition from a rewarding but frenetic 40 year career in the computer industry where there was little time to think, to a path where taking time to reflect and fully understand topics was essential. Amanda, Neil, Yi, thank you so much.

The continuous support of the School of Engineering and Computer Science has been a constant reminder of how over 40 years ago the efforts of their predecessors helped me to get my first degree, setting me on course for a great career in the computer industry.

The opportunity to meet and interact with researchers and experts in the learning analytics field at a variety of institutions has been a very rewarding one. Moreover, as I see steady interest and traction being developed by even the more cautious institutions, particularly where there are champions to evangelise and communicate the huge potential benefits of LA to students and staff. In parallel, it has been very noticeable how public awareness of AI has steadily increased during my studies and it has been fun passing on some understanding of the subject to friends and family.

A very welcome surprise to me has been the opportunity to mix with fellow-researchers in other fields such as Bio-computation, Astrophysics and Robotics. This has enriched my study experiences throughout. It is a privilege to be given such opportunities to learn new things.

Most importantly has been my wife's support throughout my studies, without which it would have been impossible for me to follow this course of study. Sheila's patience with such an alternative retirement activity has been outstanding.

I feel very strongly that the application of learning analytics to support students and academics will become the norm in all levels of education in the relatively near future. I do not believe that there are any insurmountable challenges and I hope to continue to take a small part in supporting any opportunities to encourage next steps.

Table of Contents

Appendices **207**

List of Tables

List of Figures

Publications and Presentations

During the course of this work three peer reviewed documents directly related to my doctoral thesis were published.

1. Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2015. *The potential for using artificial intelligence techniques to improve e-learning systems*. In ECEL 2015 Conference proceedings, pp. 762-770. (Appendix J)
2. Wakelam, E., Davey, N., Sun, Y., Jefferies, A., Alva, P. and Hocking, A., 2016, May. *The Mining and Analysis of Data with Mixed Attribute Types.* In Proceedings: IMMM 2016: Sixth International Conference on Advances in Information Mining and Management. IARIA, pp 32-37. (Appendix K)
3. Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2020. *The potential for student performance prediction in small cohorts with minimal available attributes*. British Journal of Educational Technology, 51(2), pp. 347-370. (Appendix L).

Each of publications 1 and 2 were presented by me at their respective conferences. These publications are included, in publication format, as Appendices I, J and K respectively.

CHAPTER ONE

Introduction

1.1 Background to Study and Motivation

The development of intelligent learning systems for deployment in both education (Johnson et al., 2016) and the commercial world (Perrotta & Williamson, 2016) has the potential to provide both students and educators with a step change in the way we acquire and disseminate knowledge and skills.

My personal motivation has developed from my Computer Science degree in 1976 through a forty year career in software implementation and management roles in three global corporations, where I directly experienced a rapidly increasing trend in the need for and development of asynchronous training. In addition, I have had experience of remote learning through the Open University (OU) in the pre and post internet eras. My Mathematics degree in 1981 relied upon written material supplemented by a modest amount of television programs, whereas my Italian language module in 2011 made the maximum use of on-line materials, fully interactive on-line tutorials with electronic whiteboards and real time verbal and written dialogue, and on-line assessment and examination. More recently, in 2014, as preparation for registering for my PhD I completed four MOOCs (Massive Open Online Courses): The Open Course in Technology Enhanced Learning (ocTEL); Stanford University, Introduction to Artificial Intelligence; University of Washington, Machine Learning; John Hopkins University, Prediction and Machine Learning. These experiences led to my growing curiosity as to why significant advances in technology and in particular artificial intelligence (AI) had not been exploited in the support of academics and students, and equally for commercial organisations. Early in my research this curiosity developed into a passion to evangelise how it may be possible that modest but effective steps could be taken to improve student success rates, with consequential institutional benefits.

My role as a visiting lecturer over the past five years has given me valuable insights into the challenges faced by academic staff and students, providing me with continuous practical experience of many aspects of my research. During this time I have lectured and conducted tutorials on three different Level 5 and Level 6 BSc modules.

My research has shown that there has been clear progress in the development of techniques to deliver more effective e-learning systems in both education and commerce. However, I have identified very few examples of comprehensive learning systems that fully exploit contemporary AI and in particular, Machine Learning (ML) techniques to be adaptive to the student's learning experience. I have surveyed existing educational and commercial intelligent learning/training systems and explored the contemporary

AI techniques which appear to offer the most promising contributions to e-learning. I have considered the non-technological challenges to be addressed and considered those factors which will allow step change progress in e-learning systems. With the convergence of several of the required components for success increasingly in place I believe that the opportunity to make valuable progress is now much stronger (Wakelam et al, 2015).

A number of the training system developments and prototypes are so-called Adaptive Learning Systems (ALS). These systems adjust the learning experience based upon the student's progress, increasing the level of difficulty or accelerating progress when the student is progressing well, and slowing down if they need further support/instruction. In addition, the systems can dynamically select from alternative learning paths to determine the optimum one(s) based upon continuous assessment of the learner. My research is seeking to deploy leading AI input into tailoring the support which can be delivered to the student.

Key to the further development of such systems is the growing collection of static and dynamic student data available for exploitation by learning analytics research (Clow, 2013). My research has included both the identification of static data e.g. age, parent's education and internet access and dynamically assessable student attributes, which are measurable during the learning activity such as speed of progress through learning objects or performance in exercises, which have the potential to be useful in predicting student outcomes. The latter are important in order that academics may then be able to identify students at risk and make appropriate, timely supporting interventions.

1.2 Research Questions

Following my preliminary investigations, I constructed the research questions for my study as explained below.

1.2.1 Research Question 1: Small Student Cohorts and Limited Student Attributes

How accurately can we predict student performance on courses comprising relatively small student cohorts, where a very limited set of student attributes are readily available for analysis?

While there is evidence to show that predictions based upon large cohorts with multiple student attributes can provide educators with useful support in identifying students at risk (Heuer & Breiter, 2018), there is little research evidence to date of the value that can be derived where cohorts are small and very limited attributes are available for analysis. What are the relative predictive accuracies that may be achieved in the analysis of student outcomes when the student cohort is small (23 in the case of my experiment) and student attributes are limited to lecture/tutorial attendance, Virtual Learning Environment (VLE) accesses and limited formal interim assessments?

1.2.2 Research Question 2: The Opportunity to make Interventions

How useful would these analyses be in order to provide course leadership with the opportunity to make timely supportive interventions at appropriate points during a module?

The value of the implementation of learning analytics is directly related to their success in consequent application to support students and institutions through appropriate timely interventions. What are the methods and timeliness of such interventions which are critical to their success, and which methods are preferred by students and therefore most likely to be successful? What ethical, moral and privacy issues relating to students that must be taken into consideration?

1.2.3 Research Question 3: Data Mining Techniques

Which data mining techniques are suitable for predicting student performance?

Which data mining techniques are available for the prediction of student performance and how do their respective predictive accuracies compare when applied to differing student cohort sizes and differing varieties of student attributes? Which of these techniques are applicable to each of numeric and nominal data? What are the student attributes which may be available to learning analytics and how might students and institutions view their respective sensitivity to privacy issues and therefore present potential restrictions of their use in a learning analytics context?

1.2.4 Research Question 4: Current Intelligent Educational Technologies

What progress has been made in the development and deployment of intelligent learning/training systems and prototypes and what are the institutional barriers to the adoption of learning analytics, alongside corresponding approaches to their resolution?

What intelligent learning/training systems and prototypes, including adaptive learning and intelligent tutoring systems, are currently available in the education and commercial sectors? What are the institutional barriers which must be overcome in order to successfully implement learning analytic and intervention systems, the corresponding critical success criteria and alternative approaches to their resolution?

1.3 Contribution of Study

I have established and published the potential for predicting individual student interim and final assessment marks in small student cohorts with very limited attributes and show that these predictions could be useful to support module leaders in identifying students potentially at risk during the course of their studies. I have demonstrated how through the analysis of these limited attributes: attendance, VLE

accesses and intermediate assessments, useful intervention guidance may be provided to academic leadership. Chapter Eight is devoted to this contribution.

I have established a novel technique for the analysis of nominal data, an important subset of student attribute data alongside numeric attributes. I have published the application of this new method applied to the nominal attributes of a freely available student dataset and compared its results with those generated by two existing, established methods of analysing nominal data. This contribution is detailed in Chapter Four, sections 4.3 and 4.4.3 which describe the technique, its application to a student dataset and comparison with the results from an alternative method.

1.4 Supporting Activities

In support of the practical application and development of these contributions (see Section 1.3), I present a combination of syntheses of analysis of existing research and where appropriate, directly applicable experience from my previous career into:

- a comprehensive analysis of institutional barriers to the adoption of learning analytics, alongside corresponding approaches to their resolution (Chapter Three, sections 3.3 and 3.4).
- a comprehensive review and comparison of global prototypes and deployed implementations of adaptive learning and intelligent tutoring systems (Chapter Three, section 3.2).
- Experimenting with alternative data mining analyses of large student datasets with a wide variety of student attributes using a variety of techniques, identifying the relative predictive importance of different attributes on student results' prediction accuracy (Chapter Seven, section 7.4.3).

1.5 Research Programme Approach

1.5.1 Regular and systematic review of relevant papers.

I conducted a regular and systematic review of relevant papers during the five years of my research studies. I have refined my Google Scholar alerts in line with search criteria relevant to my research (Table 1.1). These ensure that research paper summaries are delivered to me for review every 3 days. As a result, I have reviewed over 70,000 alert summaries from 18 alerts, leading to abstracts where they appear relevant and subsequently 700 papers of value to my research. These papers alongside other material identified have resulted in a core database of 800 papers supporting my research. Of these, 240 have been selected to support this dissertation. This has led to my identification of the status and best practice in each of the key areas of my research:

Table 1.1: Google Scholar Alerts

Alert search string	Alert search string
Adaptive learning and corporate training	Expert systems and education
Adaptive learning systems and pedagogy	Intelligent agents in adaptive learning systems
Artificial intelligence and corporate training	Knowledge based systems and education
Artificial intelligence and education	Knowledge based systems and training
Artificial intelligence and pedagogy	Learning analytics
Artificial intelligence and training	Machine learning and pedagogy
Ant colony learning or training	Machine learning in education
Cognitive tutor	Student success
Education and data mining	Academic Intervention

1.5.2 Networking

I have developed a variety of contacts in the rapidly growing network of institutions and organisations who are working effectively in the areas of learning analytics and E-learning, including Joint Information Systems Committee (Jisc, 2019a), Learning Analytics Community Europe (LACE, 2019), British Computer Society (Bcs.org, 2019), the Open University (Open University, 2019) and the University of Hertfordshire (University of Hertfordshire, 2019a).

Jisc (formerly the Joint Information Systems Committee) is a UK higher, further education and skills sectors' not-for-profit organisation for digital services and solutions, which provides UK universities and colleges with shared digital infrastructure and services including learning analytics, as well as acting as the forum for knowledge sharing, interaction and debate. In terms of learning analytics Jisc are focusing on the collection of data which enables researchers and educators to detect the need for remedial action

LACE is an EU funded project involving nine partner organisations across Europe, with the objective of connecting researchers in the fields of Learning Analytics (LA) and Educational Data Mining (EDM), promoting knowledge exchange and sharing best practices. The project delivered a large number of comprehensive documents and studies covering web analytics, learning analytics interoperability and future visions (Ferguson et al., 2015).

The Open University is pre-eminent in the successful delivery of distance learning and has steadily embraced all opportunities to exploit E-learning. With a cohort of over 170,000 students studying across several hundred courses, the opportunity for collecting student data and looking to exploit learning analytics is clear. The freely available OU Learning Analytics Dataset (OULAD) contains course data, student data (static and dynamic) and their interactions with the Virtual Learning Environment (VLE) for seven selected modules. This dataset provided me with an understanding of the types of student attributes proved valuable in the implementation of a successful LA implementation.

The University of Hertfordshire Chief Information Officer and members of the IT department were invaluable in their support to help me understand UH's priorities and their local focus on the collection and exploitation of learning analytics. UH is aiming to expand its collection of student data and in particular to investigate how learning analytics could help to identify students where intervention is needed and thus maximise retention. I have accepted their invitation of involvement in future team discussions on learning analytics.

In addition, I have identified and attended relevant conferences and reviewed their prospective proceedings to ensure that I am aware of work supporting my research and studies.

1.5.3 Experiment

To support my research and thesis I conducted preliminary experiments on several freely available student datasets (see Sections 1.6.8 and 6.2) and finally upon a current and live university module, where I was responsible for student data collection and presentation. In all cases I applied a variety of analysis methods, with an emphasis on machine learning techniques, to evaluate and compare prediction accuracies. In the case of my final study I was able to explore how the different analysis techniques could support academic interventions. Given the nature of the experiment I applied for and was given Ethics Approval by the University of Hertfordshire prior to commencement of any studies using personal data (Appendix B).

1.6 Research Journey

In order to pursue my study and develop my thesis it was necessary to follow my research across both the fields of education and of computing. The following sections provide a map of my research journey through the relevant components of these two fields leading ultimately to a focus upon student performance prediction and potential intervention approaches.

1.6.1 Development of my Research Techniques

Pre-enrolment, as preparatory study I completed four ‘remote’ online learning courses (MOOCs), see section 1.1.

In order to re-master the relevant computational tools and techniques I completed the UH School of Computer Science Level 6 Computer Science undergraduate module Constructive Artificial Intelligence and the Level 7 Master’s module Neural Networks and Machine Learning.

To develop my capabilities as a researcher I completed 24 directly relevant University of Hertfordshire Researcher Development Programme (RDP) courses (Appendix A).

1.6.2 The Importance of Pedagogy

Pedagogy is usually defined as “the theory and practice of education” (Lewthwaite & Sloan, 2016), however it may be considered as covering a wider range of topics such as the act of teaching and the associated policies and challenges (Papatheodorou & Potts, 2016). The development of a critical understanding of pedagogical research and the continuing adjustments being made to best practice has been a key component of my research activities. In particular, this understanding is critical to identifying its applicability to the deployment of Technology Enhanced Learning and the development of intelligent learning systems in particular.

An early part of my pedagogical research was the evaluation of conflicting evidence on the value of exploiting a learner’s cognitive style in improving student learning achievement (see Section 2.5.1).

My research included a continuous review of the increasing deployment of Technology Enhanced Learning (TEL) in both education and commercial sectors, including the key drivers for its use and the challenges and obstacles to be overcome. The term TEL is used to describe any or all applications of technology to teaching and learning, including what is sometimes referred to as e-learning. The focus of my work is on the branch of TEL that applies to the exploitation of computing techniques.

I briefly reviewed research into the pedagogical aspects of learning systems and approaches, including the assessment of individual learning styles and their usefulness (see Section 2.5.1).

1.6.3 Artificial Intelligence and Machine Learning Techniques

The Oxford Dictionary defines Artificial Intelligence (AI) as “The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages” (The Oxford Dictionary, 2019).

Machine learning (ML) is a branch of AI research focussing upon the study of computer algorithms that are capable of improving automatically through experience. In particular, ML aims to determine how to perform important tasks by generalizing from examples (Hastie et al., 2005). Supervised and unsupervised learning are two types of ML: Supervised Learning (Sammut & Webb, 2017), where the goal of the analysis is known and the training of the system can therefore be given feedback about how the learning is progressing i.e. the teacher provides the learner with the answers at training time, for example a chess game; Unsupervised Learning (Sammut & Webb, 2017) is where all that we have is data and the objective is to identify hidden structure, useful patterns and features of the data.

In recent years the level of media interest in the field of AI has noticeably increased with articles in the news such as: "2029, the year when robots will have the power to outsmart their makers" (Kurzweil 2014), "Driverless cars trialled on UK roads for first time in four towns and cities" (Dearden 2015) and "UK government plans for how AI may be used to prevent traffic jams months in advance" (Shale-Hester, 2019). A broad range of mainstream news outlets such as the BBC (BBC, 2019), the Daily Mirror (Daily Mirror, 2019) and The Guardian (The Guardian, 2019) publish sections devoted to AI news and developments. This steady increase in public awareness (albeit often in more populist topics) will facilitate a more open approach to considering AI as a practical tool in real life activities. I have explored a number of appropriate Artificial Intelligence (AI) and Machine Learning (ML) techniques, including Data Mining (DM), Principal Component Analysis (PCA), Growing Neural Gas (GNG), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Bayesian classifier, Classification and Regression Tree (CART) and Support Vector Machines (SVM). I also considered techniques used in some adaptive learning research including Knowledge Based Systems (sometimes referred to as Expert Systems), Fuzzy logic, Roulette Wheel algorithms, Ant Colony optimisation and Chi-square testing (see Section 7.2).

Using freely available student datasets (see Section 6.2) I conducted a variety of experiments and analyses into the application of Machine Learning techniques to explore their use.

Given that student numbers can range from small to large cohorts, I identified and experimented with small (10), medium (258) and large (1000+) student datasets, using Support Vector Machine techniques for the small dataset and Principal Components Analysis and a variety of Machine Learning techniques for the medium and large datasets.

As part of the investigation of techniques to analyse nominal data, I developed what I believe to be a novel nominal data analysis technique, applying this technique to establish interesting correlations in the

Portuguese student data (see Section 6.2.3) and comparing the results with chi-square testing (see Section 7.3).

1.6.4 Technology Enhanced Learning Systems

I surveyed the existing intelligent learning/training systems in each of education (68% of the systems identified) and commercial (9%) sectors, with systems applicable in both education and commercial sectors at 23% (see Section 3.2). These systems are categorised as ALSs or Intelligent Tutor Systems (ITS). Over half (58%) of those surveyed are ALSs.

I established that geographically, traction is highest in the US, followed by Europe and that just over 40% of these systems have been developed by universities or as collaborative projects between Higher Education Institutions (HEIs) and industry.

In order to understand student prediction and potential intervention points I developed an adaptive learning system conceptual framework (Wakelam et al., 2015), see Appendix I.

Using existing available system design and implementation research (Ferguson et al., 2014) and my own experience in the software industry I catalogued the organisational/non-technological challenges that must be addressed for successful system development. These ranged from organisational and political obstacles to academic staff and student concerns and needs (see Section 3.3).

I subsequently proposed a set of critical success criteria to apply to the development and use of e-learning systems based upon available research and my own experience in the systems and software development industry (see Section 3.4).

1.6.5 Learning Analytics

As a consequence of continuously identified and reviewed research into Learning Analytics along with the techniques being applied to determine student knowledge, predict student performance and to identify any need for intervention, I also explored the approaches taken to detect an individual's learning style in the development of learning systems, considering the conflicting evidence of their usefulness in the implementation of learning analytics.

Combining research into the value of different student attributes in predicting performance with my own experiments I have compiled a comprehensive list of potentially useful static and dynamic student attributes. This may allow a rigorous qualification and reduction of these to a mutually independent subset that may be exploited in future predictive and adaptive learning systems work (see Section 6.3.1).

1.6.6 Identification of students at risk

In addition to surveying and compiling a list of the factors recognised as having potentially negative effects on student performance, including social, institutional and pedagogical (see Section 4.3), I researched the impacts of student failure and withdrawal on students and their families and the impacts, financial and league table (e.g. rate of degree completion and satisfaction scores), upon the institutions themselves (see Section 4.2). I then researched each of traditional, non-computational and computational methods of the identification of students at risk (see Section 4.4).

1.6.7 Intervention Opportunities

I surveyed methods of student interventions, reviewing manual and automatically generated approaches (see Section 7.2), considering which are applicable to student monitoring through learning analytics and how such interventions might be timely in resulting in positive learning outcomes. The issue of how such interventions are made is critical to their success, supported by published research into which methods have the most beneficial reception from students (see Section 5.3). I present and consider their implementation in respect of student privacy and ethics issues (see Sections 6.3.1 and 5.4).

1.6.8 Experimentation

I initially performed a machine learning analysis of final grade classification on the open source Small Student Dataset for Higher Education Teachers, comprising 10 students and 11 mixed numeric and categoric attributes (see Sections 6.2.1 and 7.4.1)

I subsequently performed a machine learning analysis of students' knowledge levels on DC Electrical Machines, an on-line web based Electrical Engineering course of 258 students, comprised of 5 numeric performance attributes per student (see Sections 6.2.2 and 7.4.2).

I then undertook detailed analysis of student performance in a Portuguese student dataset of 1044 students, comprised of 16 numeric and 17 nominal attributes per student using machine learning techniques, Principal Components Analysis (PCA) and Growing Neural Gas (GNG), which resulted in establishing some interesting correlations (see Sections 6.2.3 and 7.4.3).

After that, I researched and reviewed part of the OU student dataset of 32,000 students across 22 courses and 28, mixed numeric and nominal, attributes per student (see Sections 6.2.4 and 7.4.4).

Each of the above are freely available open source datasets.

Finally, I designed and conducted an experiment on a Level 6 UK module student cohort of 23, where individual student data is limited to lecture/tutorial attendance, virtual learning environment accesses and

intermediate assessments. This experiment was conducted in real time, allowing academic leadership to consider and act upon student performance predictions, including the potential for interventions (see Sections 6.2.5 and 7.4.5).

1.6.9 Publications

During the course of my research I have published three peer reviewed papers: these were two conference papers in 2015 and 2016 respectively and a journal paper in 2019, each supporting key components of my research towards this dissertation. The full publications are included in Appendices J, K and L.

1.7 Thesis structure and overview of chapters

In Chapter Two, I review the literature for the relevant pedagogy, technology enhanced learning systems, learning analytics, artificial intelligence and machine learning approaches also including the literature for the identification of students potentially at risk and the variety of potential intervention approaches.

In Chapter Three, I survey existing intelligent learning/training systems in each of educational and commercial sectors and I then compare adaptive and non-adaptive learning systems. I present a survey of intelligent learning/training products and prototypes and discuss relevant E-learning system success criteria. I catalogue the organisational/non-technological challenges that must be addressed for successful system development.

In Chapter Four, I focus upon the identification of students at risk, describing the possible factors affecting student performance, and how we may identify them during the course of their studies and in time for positive academic intervention.

In Chapter Five, I survey and review student intervention methods, considering each of traditional, non-computer facilitated and computer facilitated generated approaches. I discuss alternative methods of academic staff interactions with individual students, aligned with student preferences identified in published research. I present and consider their implementation in respect of student privacy and ethics issues

In Chapter Six, I provide a detailed description of each of the datasets used in this research, including a general definition of each data type for measurement (quantitative) and categorical (nominal and ordinal). I catalogue the wide variety of student attributes I have encountered during my research and experiments. I then propose a list of the potentially useful static and dynamic student attributes, which may be of value in student performance prediction.

In Chapter Seven, I provide a description of a variety of relevant AI and ML techniques. I describe the results and conclusions of my own experiments using selected techniques applied to freely available datasets and including a brief description of an experiment conducted on a live student cohort (fully described in Chapter Eight). I describe a novel technique for the analysis of nominal data.

In Chapter Eight, I describe an experiment to establish the potential for student performance prediction in small cohort of 23 students, with the minimal available attributes of lecture/tutorial attendance, virtual learning environment accesses and intermediate assessments, using learning analytics techniques. I discuss how these analyses could be used to support educators in the identification of students at risk during module delivery and how the data may support timely intervention where appropriate.

Finally, in Chapter Nine, I summarise the conclusions arising from my experimental results and my research overall and provide recommendations for future work.

CHAPTER TWO

Literature Review

2.1 Introduction

I have structured the following literature review in line with my research questions and the key components of my research journey.

2.2 Small Student Cohorts and Limited Student Attributes

2.2.1 Learning Analytics

The objective of learning analytics is to offer tutors the opportunity to identify and support the need to make timely interventions where a student's success is potentially at risk. A learning analytics cycle is shown in Figure 2.1 (Ferguson & Clow, 2017).

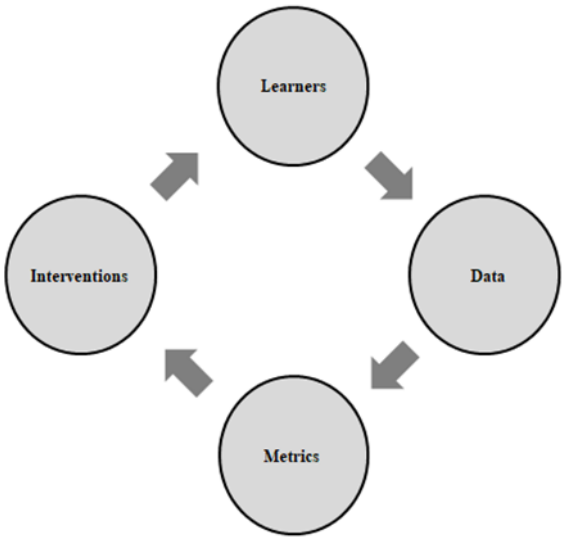


Figure 2.1: The learning analytics cycle (Ferguson & Clow, 2017, p7).

The deployment of learning analytics establishes the need and opportunity for student and module interventions (Clow, 2012). The study concludes that the faster the feedback loop to students, the more effective the outcomes.

Institutions routinely collect considerable amounts of data on each student, starting from their initial application forms and continuing throughout their studies. Given the very large quantity of data that is available to be captured and exploited and the level of complexity of the interdependencies of large numbers of data classes/attributes, requiring multi-dimensional analysis, these datasets are no longer capable of analysis, particularly in real-time, by using manual or orthodox IT techniques (Lang et al.,

2017). Of particular relevance to e-learning systems are continued developments in Machine Learning (Kubat, 2017) and Data Mining (Tan, 2018) methods.

There has been considerable progress in examining the potential of AI techniques in the analysis of student data for the benefits of students, staff and the institutions themselves. The 2019 Educause Horizon Report, Higher Education Edition (Alexander et al., 2019, p.10) shows positive progress in the deployment of predictive analytics in the 75 institutions surveyed (Figure 2.2).

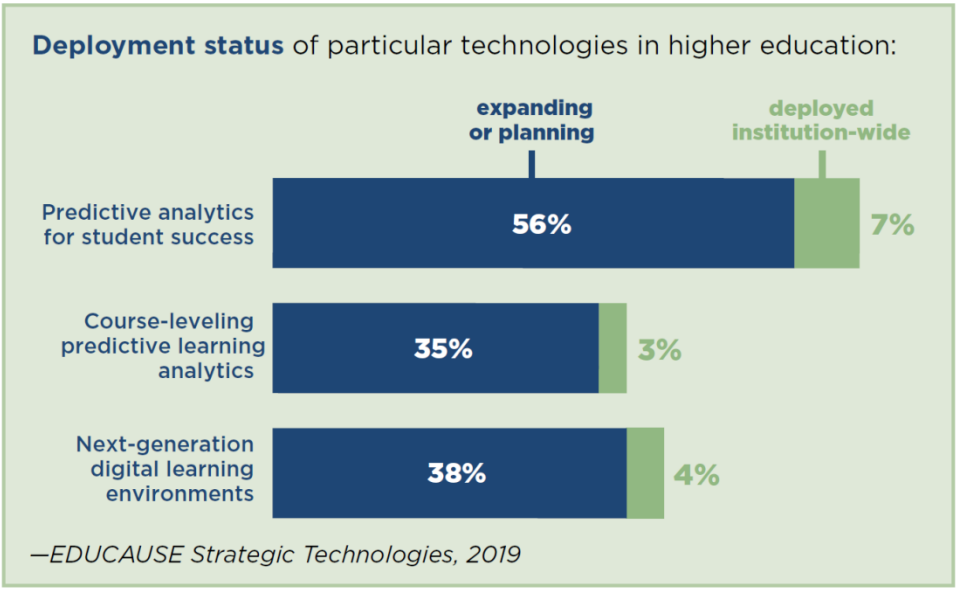


Figure 2.2: Deployment status of particular technologies in Higher Education (Alexander et al., 2019, p11)

The analysis of student performance and prediction of student outcomes is a core component of the field of Learning Analytics (LA). LA is defined as " the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Ferguson, 2012). There is growing evidence of the effectiveness of learning analytics initiatives, in particular its impact on student grades and retention, from Jisc summarising published evidence for the effectiveness of learning analytics initiatives (Sclater & Mullan, 2017) and more recently a systematic review of empirical studies conducted between 2013 and 2017 revealing evidence of LA reducing student dropout particularly in the US, Australia, and England (Yau et al., 2018). However, as discussed in the European Commission (Joint Research Centre (JRC) Science for Policy Report (Ferguson et al., 2016) the high expectations for LA have yet to be met.

In their analysis of learning analytics and interventions publications between 2007 and 2018, Wong and Li selected 23 case studies highlighting the measured benefits of learning analytics in distance learning institutions (Wong & Li, 2018). A study of Open University Analytics (OUA) usage by 189 teachers and 14,000 students across 15 undergraduate courses in the academic year 2017/18 (Herodotou et al., 2019) showed that teachers who made 'average' use of OUA were found to benefit their students the most, as measured by significantly better performance than their peers in the previous year's course presentation (where the same teachers did not use OUA).

The case for interventions based upon learning analytics is a strong one. Evidence from several institutions demonstrates reductions in student drop-out rates (Sclater et al, 2017) from 18% to 12% at the University of New England and a 14% reduction in the number of D and F grades at Purdue University (Purdue, 2019) using their SIGNALS project (Arnold & Pistilli, 2012). Purdue University (Indiana, USA) comprises over 40,000 students including both on-campus and on-line study (Purdue University, 2019). In 2007, the Purdue Course Signals project was initiated with the objective of applying learning analytics to provide students with real time feedback on their progress and academic staff with the opportunity to identify students at risk. The Signals project collects and analyses student' grades performance, demographics, previous academic performance and their VLE activity (Arnold & Pistilli, 2012).

However, it is important to note the limitations identified in the Purdue Course Signals research (Ferguson & Clow, 2017). For example, between 2007 and 2009, retention on courses that didn't employ course signals had also risen substantially, suggesting that other university-wide factors were having an effect on retention. In addition, it is unclear whether the research had explored whether student retention improvement could be explained because students had taken more courses using Course Signals, or whether they took more of those courses because they had been retained.

The overwhelming focus on learning analytics in Higher Education has been devoted to the analysis of "big data" (Ashraf et al., 2018) where the data comprises very large student cohorts and a large number of student data attributes. These attributes often include personal and admission data as well as previous educational records (see Table 2.1) cited from Ashraf et al. (2018).

Table 2.1: Student Attributes (Ashraf et al., 2018)

Criteria	**Details**
Student demographic information	Age, gender, region, residence, guardian info
Previous results	Cleared certificates, scholarships and results
Grades	Recent assignment results, quizzes, final exam, CGPA, attendance
Social network details	Interaction with social media websites
Extra-curricular activities	Games partitions, sports, hobbies
Psychometric factor	Behaviour, absence, remarks

The measurement of student performance during their progress through university study provides academic leadership with critical information on each student's likelihood of success. Academics have traditionally used their interactions with individual students through classroom activities and interim assessments to identify those "at risk" of failure/withdrawal. However, modern university environments, offering easy on-line availability of course material, may see reduced lecture/tutorial attendance (Marburger, 2001; Mearman et al., 2014), making such identification more challenging. Modern data mining and machine learning techniques provide increasingly accurate predictions of student examination assessment marks (Ashraf et al., 2018), although these approaches have focussed upon large student populations and large numbers of data attributes per student.

In fact, many university modules comprise relatively small student cohorts. A recent study, based upon 67 UK universities, found average class sizes of approximately 20 students (Huxley et al., 2018)

In addition, institutional ethical, privacy and moral protection protocols limit the student attributes available for analysis (Sclater et al., 2016a). It appears that very little research attention has been devoted to this area of analysis and prediction of low student cohorts and very limited attributes.

In addition to the sensitivity of such attributes, despite their algorithmic accuracy intentions, there is growing research into the potential for machine learning approaches to introduce bias, such as class, gender and ethnicity (Wilson et al., 2017). It is essential that learning analytics implementations guard against this. Furthermore, research into students' autonomy in learning (Fazey & Fazey, 2001) exploring the potential for measuring learning related psychological characteristics such as motivation and self-esteem, could provide additional attributes in future systems.

Some attributes are routinely collected components of student data that are typically available to LA systems have been shown to be useful both as indicators of a student's performance and in the prediction of likely outcomes such as passing or failing. These include on-going student attendance at lectures/tutorials, virtual learning environment (VLE) accesses and interim assessment results. There is evidence that student attendance at lectures and tutorials is a useful predictor of likely student outcomes (Aziz & Awlla, 2019; Fike & Fike, 2008). There is some evidence that interim assessment as part of the overall course assessment is a strong predictor of student success (Sclater et al., 2016). Case studies included in this report also identify a student's VLE accesses as a more accurate predictor of success than their historical or demographic data. The usefulness of VLE accesses as a predictor of student performance is further supported by an experiment conducted on the data from over 30,000 students across 7 OU modules (Doijode & Singh, 2017) where students with the highest VLE accesses obtained the highest scores. As with the majority of research conducted, these case studies measured very positive impacts from resulting interventions. A recent study (Heuer & Breiter, 2018) analysing student VLE activity across 22 courses and 32,593 OU students found student VLE accesses to be an important indicator of student performance. Further support of the value of the analysis of VLE accesses in predicting student outcomes is provided by Wolff et al. (2013) which indicates that the use of even coarse-grain data about students' VLE activity is useful in predicting students at risk , and more so when combined with other student data.

Note that each of these three attributes (attendance at lectures/tutorials, VLE accesses and interim assessment results) is collected live as the course/module progresses and therefore the LA algorithms are making no judgements on a student's profile, background or past history. In this case, the results of learning analytics may be considered as a very "pure" approach in that a student is being judged as capable of any level of achievement, or otherwise, irrespective of history or other factors. However, where institutional protocols permit, there is evidence that previous academic performance is a valuable predictor of student outcomes (Honicke & Broadbent, 2016; McKenzie & Schweitzer, 2001).

In the UK, the Open University (OU) is a world leader in the collection, intelligent analysis and use of large scale student analytics. It provides academic staff with systematic and high quality actionable analytics for student, academic and institutional benefit (Rienties et al., 2017). Rienties and Toetenel's 2016 study (Rienties & Toetenel, 2016) identifies the importance of the linkage between learning analytics outcomes, student satisfaction, student retention and module learning design.

Institutions are naturally cautious in their consideration of the design and implementation of any new systems, and the case of LA is no exception. A variety of research-based material is available to support

LA champions from initial presentations and discussions with executive management through to all stakeholders. In their critical review of LA, Banihashem and colleagues present a comprehensive summary of the potential benefits of learning analytics to stakeholders (Banihashem et al., 2018). This summary (Table 2.2) provides a useful starting point for institutions considering the deployment of LA.

Table 2.2: Benefits of Learning Analytics to stakeholders (Banihashem et al., 2018, p7)

Stakeholders	**Benefits**
Learners	Enhance engagement of students
	Improve learning outcomes
	Personalization of learning
	Increase in students adaptivity
	Enrich personalized learning environments
	Increase self - reflection and self-awareness
Teachers	Assessment services
	Make efficient interventions
	Get a real - time feedback
	Get a real - time insight
	Understand students learning habits
	Modify content for students' desire
	Monitoring students' activities
	Get a deeper understand of teaching and learning
	Predicting student performance
	Provide warning signal
	Improve teaching strategy
	Improve instructor performance
	Sources recommendation

Stakeholders	Benefits
Institutions	Improve educational decision making
	Increase student success
	Student success modelling
	Monitoring students' activities
	Boost cost efficiency
	Increase retention rate
	Make evidence - based decisions
	Prevent student drop out
	Identify students at risk
	Curriculum improvement
	Improve accountability
Researchers	Increase efficiency of education and serious games
	Identify knowledge gaps
Course designers	Identifying target course
	Improve learning design
Parents	Monitoring students' activities

The LACE project have developed the DELICATE checklist (Figure 2.3) to focus upon the critical issue of institutional and stakeholder trust in LA implementations (Drachsler & Greller, 2016). This also provides a useful checklist for the consideration of learning analytics.

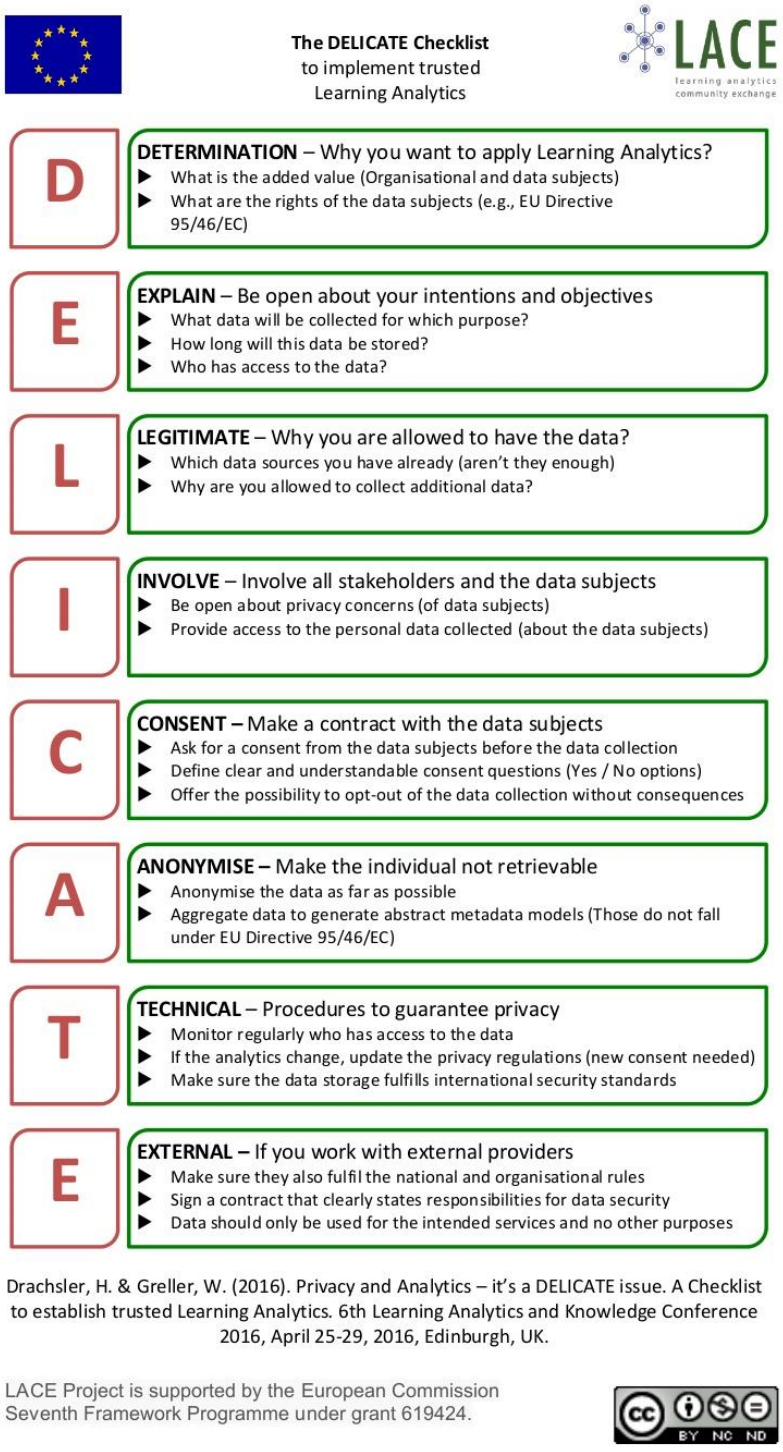


Figure 2.3: The DELICATE checklist (Drachsler & Greller, 2016, p8)

Legal, ethical and moral considerations in the deployment of learning analytics and interventions are key challenges to institutions. They include ensuring informed consent, transparency to students, the right to challenge the accuracy of data and resulting analyses and prior consent to intervention processes and their execution (Slade & Tait, 2019). These are well documented in a number of research papers, for example (Pardo and Siemens, 2014; DeFreitas et al., 2015; Corrin et al., 2019). In addition, a comprehensive literature review of 86 publications commissioned by Jisc discusses the challenges faced by institutions and provides the background for a future code of practice for LA (Sclater & Bailey, 2018). A discussion on ethical and data privacy issues in learning analytics based on three studies in Higher Education and Primary school contexts (Rodríguez et al., 2016), specifically focusses on tutor-led approaches. Legislation has been in place for over two decades, specifically the European Data Protection Directive 1995 (European Union, 1995) and the UK Data Protection Act 1998 (UK Data Protection Act, 1998).

A recent literature review of learning analytics (Banihashem et al., 2018) cited ethics and privacy (Figure 2.4) as one of the most important challenges of educational learning analytics (alongside what they describe as a "lack of attention to theoretical foundations and scope and quality of data").

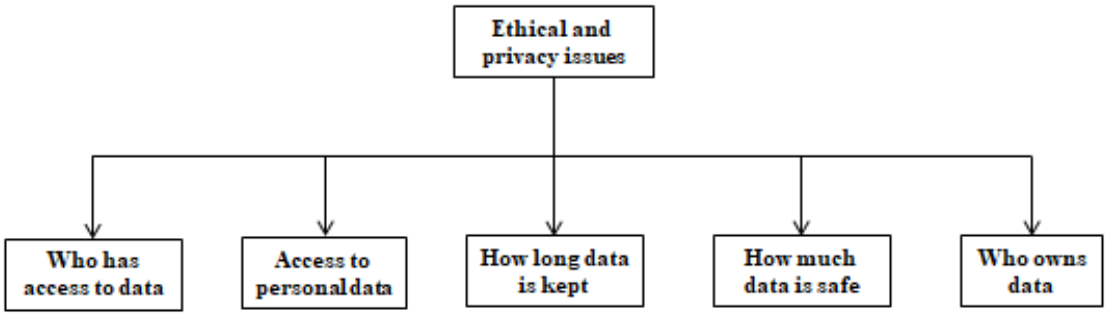


Figure 2.4: Ethical and Privacy Issues in the Use of Learning Analytics in Education (Banihashem et al., 2018, p6)

The International Council for Open and Distance Education 2019 report "Global guidelines: Ethics in Learning Analytics" (Slade, S. & Tait, 2019) identifies a number of what it believes to be core globally relevant ethical issues. The report recognises the development of several guidelines, codes of practice and policies in recent years, highlighting the OU Policy on Ethical use of Student Data for Learning (OU, 2014), Jisc's Code of Practice for Learning Analytics (Sclater & Bailey, 2018) and the Learning Analytics Community Exchange (LACE) framework in 2016 (Drachsler & Greller, 2016). However, the report argues that these are in response to local geographical and legal requirements.

The topic of student consent is integral to discussions of ethical and privacy considerations and policies. The principles of medical (patient) consent are often referred to as a basis for policy development

(Prinsloo & Slade, 2018). However, as the authors point out, in the Higher Education context where decision-making power is not equally shared, consent is a more complex topic. For example, in the educational context, the institutional objective may focus more upon the achievement of organisational goals than the most favourable outcome for the student.

More recently, General Data Protection Regulation (GDPR), (UK Government, 2018) sets out the legal and data protection principles which institutions and organisations are responsible for adhering to. In addition, despite their algorithmic accuracy intentions, there is growing research into the potential for machine learning approaches to introduce bias, such as class, gender and ethnicity (Wilson et al., 2017). The topics of legal, ethical and moral issues are also discussed Chapter Seven, section 5.4.

2.2.2 Experiment

Considerable research has been published describing the experimental analyses of a variety of learning analytic approaches and techniques, exploring patterns of student behaviour, correlations between attributes and consequent usefulness of results to support students, academics and institutions. Of specific interest to this study are experimental results focussing upon the ability of different machine learning techniques to identify useful student attributes and their suitability to predict student outcomes in order to identify students at risk.

A comparison of various data mining techniques (Ashraf et al., 2018) to predict student module marks using regression methods demonstrates achieved student prediction accuracy levels ranging from 50% to 97%. Accuracy is measured as the percentage accuracy of the prediction versus the actual student result. Accuracy levels are shown by algorithm (Figure 2.5) and by summary attributes and algorithm (Figure 2.6) cited from Ashraf et al. (2018). These analyses included student numbers in excess of 10,000 and 77 attributes in some cases.

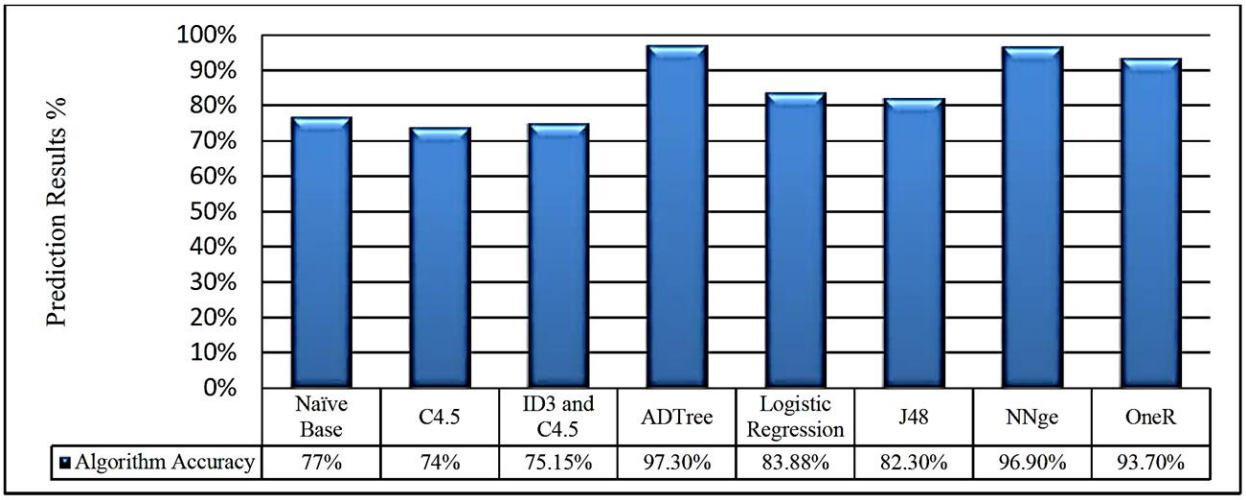


Figure 2.5: Prediction accuracy by algorithm (Ashraf et al., 2018, p134)

Note: Original paper included "Naïve Bayes" missspelt as "Naïve Base".

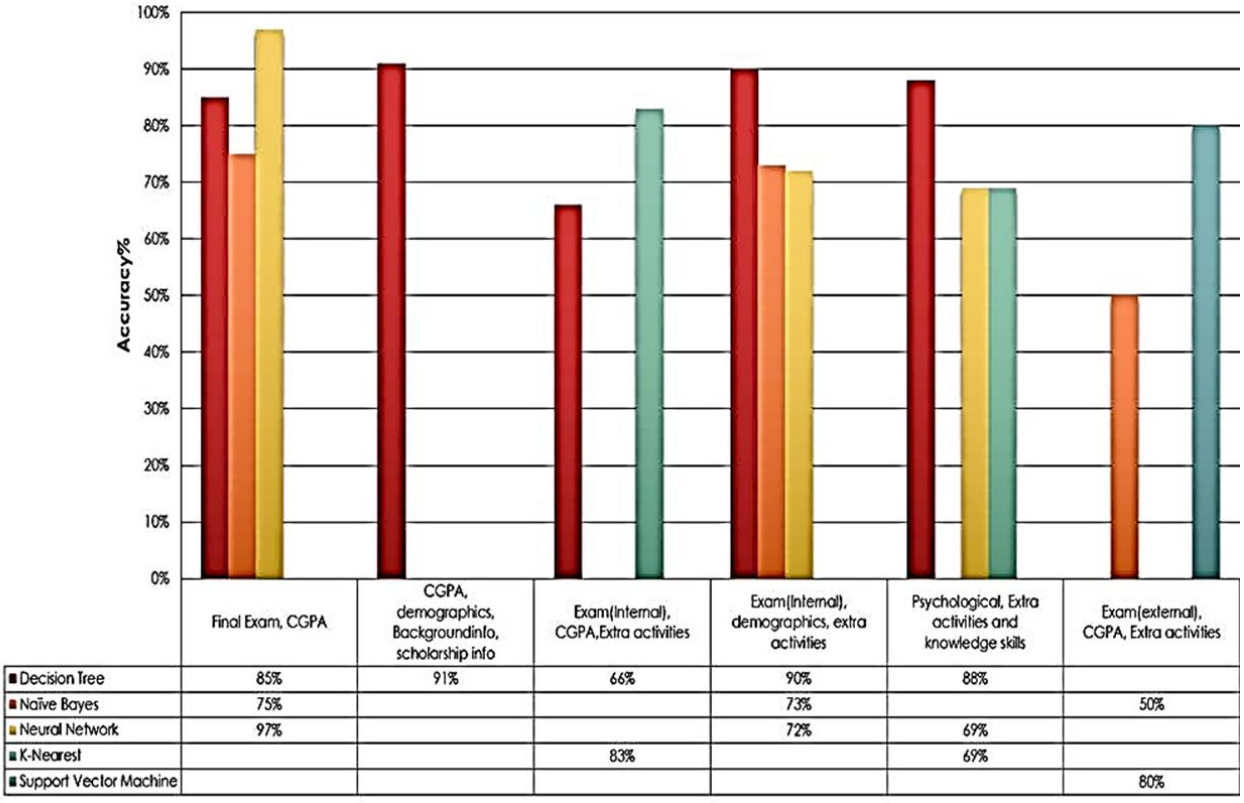


Figure 2.6: Prediction accuracy by summary attributes and algorithm (Ashraf et al., 2018, p131)

As an indication of the wide variety of student attributes potentially available to learning analytics when institutional privacy, ethical and moral considerations are not an issue, the dataset used in a Portuguese

student study provides a very good example (Cortez & Silva, 2008). Highly sensitive attributes such as alcohol consumption, romantic interest and parents' academic achievements and jobs are included in a 33 attribute dataset and over 1000 students. Experimentation on potential correlations between these attributes provided some interesting results, for example, potential correlations were evident between paid tutoring, the student's wish to take Higher Education and parent cohabitation, closely followed by educational support and Mother's job (Wakelam et al., 2016). This dataset is detailed in section 6.2.3.

In contrast, although the Open University learning analytics programme makes use of a similarly wide range of student attributes, the more sensitive demographic and more personal data are mostly excluded. In particular, of the 28 student attributes used, 23 are associated with prior academic performance. The 5 demographic attributes are age, gender, region of residence, disability and its associated Index of Multiple Deprivation (IMD) band (UK Government, 2015). Of these it is the IMD band which is the most sensitive. As a freely available dataset comprising over 32,000 students, OULAD provides researchers with excellent opportunities to perform a variety of experiments on a large dataset. This dataset is described in full in section 6.2.4.

2.3 The Opportunity to make Interventions

2.3.1 Identification of Students at Risk

An inability to identify and consequently successfully support students at risk of failure or withdrawal presents two serious threats to universities. Firstly, the consequences of already budgeted student fees disappearing from university revenues are significant as can be seen by the percentages of student withdrawals. The UK Higher Education Statistics Agency (HESA, 2018b) performance indicators show that the percentage of full time students not continuing after one year of study who started in 2015/16 was 6.4%. In the case of part-time students, the figure was 34.2%. In the case of American University students, Lin, Yu, and Chen (Lin et al., 2012) noted that predicted retention probability decreases from around 70% for a representative full-time student to 57% for a part-time student. In the case of open, distance environments retention and progression has been established to be a greater issue than for traditional full-time campus-based students according to Simpson (2006 and 2013). Secondly, student satisfaction scores are an integral part of the scoring mechanism that determines a university's place in national and global rankings. The impact of these scores on rankings has been shown to be greater for more able students, for universities with entry standards in the upper-middle tier, and for subject departments facing more competition from other universities (Gibbons et al., 2015).

The Open University Analytics4Action evaluation framework, analysing over 90 large-scale modules over a two year period, (Rienties et al., 2016b) identifies the importance of placing the power of evidence based learning analytics into the hands of academic staff to:

- accurately and reliably identify learners at-risk
- identify learning design improvements
- deliver (personalised) intervention suggestions that work for both student and teacher
- operate within the existing teaching and learning culture
- be cost-effective.

Lecturers and researchers at the OU have access to a substantial range of data pertaining to teaching and learning. The systems deployed monitor student VLE activity (Tempelaar et al., 2015), survey students (Ashby, 2004) and capture the pedagogic balances within a module (Cross et al., 2012). In addition, the OU have developed their own range of data interrogation and visualisation tools (Cross et al., 2012; Rienties & Rivers, 2014).

The OU Analyse project (Kuzilek et al., 2017) specifically aims to predict learners-at-risk (i.e., lack of engagement, potential to withdraw) in a module presentation as early as possible so that cost-effective interventions can be made. In OU Analyse, predictions are calculated in two steps:

- Predictive models are built by machine learning methods using legacy data recorded in the previous presentation of the same module
- Student performance is predicted weekly from these models and the other learner data of the current module presentation (Wolff et al, 2013; Wolff et al., 2014).

This includes VLE data representing students' interactions with on-line study material and these interactions are classified into activity types and actions. Each activity type corresponds to an interaction with a specific kind of study material (Rienties et al., 2016a; Wolff et al., 2013; Wolff et al., 2014). Student data is collected daily and provided to academic staff and students through a variety of methods including dash boards and emails.

Jisc has provided leadership in the research and deployment of learning analytics since 1993. Their focus is on the identification of students where interventional support may be needed (Jisc, 2019). Jisc is supporting a variety of UK universities and colleges in their development and deployment of learning analytics and through its networking and regular workshops valuable experience of LA and interventions

is shared. An example of Jisc knowledge sharing was at the 11th Jisc Learning Analytics Network event at Aston University where a presentation described how the University of New England, Australia, identifies three triggers used to identify students potentially at risk (Sclater, 2017). Firstly, no accesses to the VLE for more than 7 days during the first two weeks of the semester. Secondly, reminders sent for assessment tasks, followed by poor results or no-completion. Thirdly, limited or no access to major assessment information in the seven days prior to the due date.

2.3.2 Intervention Opportunities

A comprehensive review of learning analytics intervention case studies, from 23 institutions, published between 2007 and 2018 categorised interventions into four types: Direct message; Actionable feedback; Categorisation of students and Course redesign (Wong & Li, 2018). Direct messages and actionable feedback were the two most frequent intervention types (Figure 2.7).

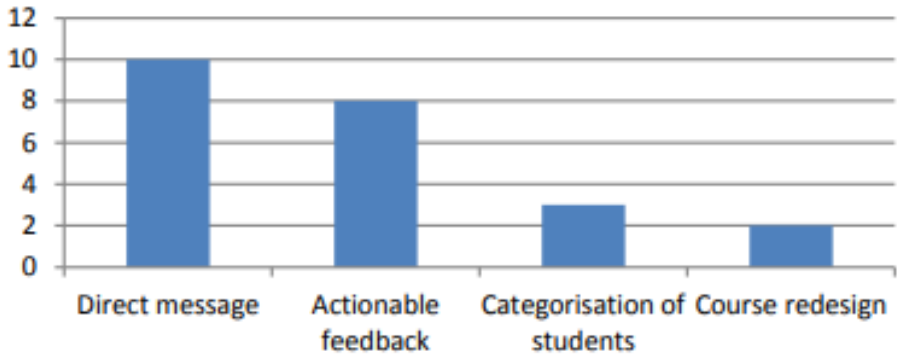
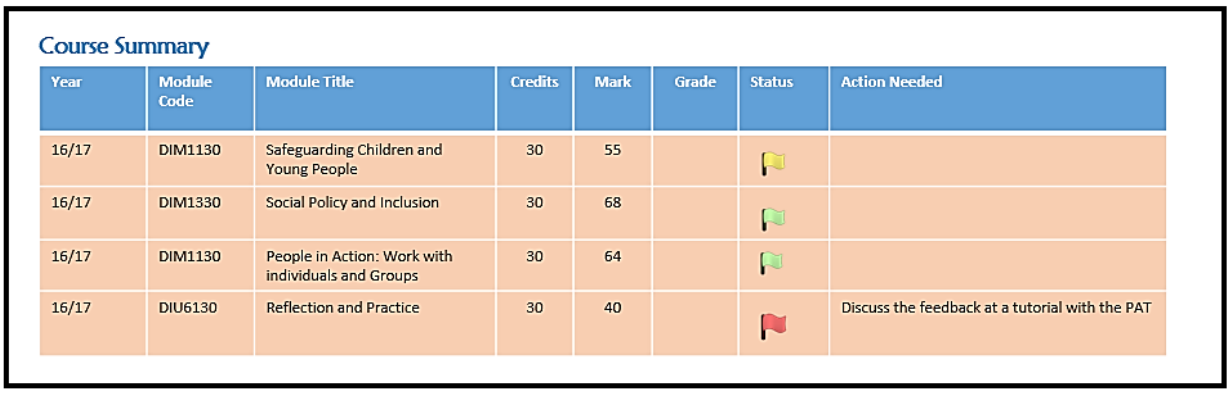


Figure 2.7: Frequency of different types of learning analytics intervention methods (Wong & Li, 2018, p179)

The paper gives a detailed description of the actual intervention methods used by each institution is described, highlighting institutional beliefs on the importance of personalised feedback. Choi and colleagues (Choi et al., 2018) summarise the pros and cons of alternative intervention methods (see Chapter Eight, Section 8.4) in their study highlighting the benefits to academic staff faced with limited time and resources.

An increasingly common method of providing interventional feedback to students is that of dashboards. A systematic literature review of learning analytics dashboard research presenting the results of 55 papers (Schwendimann et al., 2016) examines alternative methods most supportive to different educational stakeholders. Figure 2.8 (Bennett, 2019) shows an example student profile with RAG (Red, Amber and

Green) rating flags. Typically, red indicates an issue which must be addressed/requires action, amber indicates some concerns/early warning and green indicates on track. Implementers of dashboards are able to specify in detail how students and academic staff should interpret these indicators.



Course Summary

Year	Module Code	Module Title	Credits	Mark	Grade	Status	Action Needed
16/17	DIM1130	Safeguarding Children and Young People	30	55			
16/17	DIM1330	Social Policy and Inclusion	30	68			
16/17	DIM1130	People in Action: Work with individuals and Groups	30	64			
16/17	DIU6130	Reflection and Practice	30	40			Discuss the feedback at a tutorial with the PAT

Figure 2.8: Student's profile with RAG rating flags (Bennett, 2019, p12)

As in any intervention process, the earlier that learning analytics can identify students whose performance may not be on track and take first steps the better. It is important that initial alerts to students are sensitively made, whether automatically generated emails or direct staff contact. If LA does not recognise corrective action or improved progress then steadily escalating LA informed alerts may follow. Marist College New York (Marist College, 2019) gives a very simple example of messages which steadily increase in tone.

First message:

"I am reaching out to offer some assistance and to encourage you to consider taking steps to improve your performance. Doing so early in the semester will increase the likelihood of you successfully completing the class and avoid negatively impacting your academic standing"

Next message:

"Based on your performance on recent graded assignments and exams, as well as other factors that tend to predict academic success, I am becoming worried about your ability to complete this class successfully".

This attention to the consideration of how to make interventions in the way that will have the most positive effect on students is supported by a study of student preferences and attitudes to the use of alerts on their progress as shown in the survey results of 639 undergraduate students at Macquarie University, Sydney, Australia (Atif et. al, 2015). This is discussed in Section 5.3.

The OU also provides academic staff with a menu of potential intervention actions (Rienties et al., 2016b), based upon learning analytics data and visualisations (Table 2.2).

The process of intervention is often an iterative one. When an intervention is made, whether automatically generated or by staff contact with a student, the student's response in terms of corresponding changes in the activities or progress may require follow up. Similarly, identified issues may be seen to be wider than a single student and this may suggest that institutions must review and address systematic issues. In this case, multiple student intervention strategies such as revised tutorial topics may be appropriate, or a redesign of future occurrences of the module may be necessary.

This menu is based upon the Community of Inquiry (CoI), (Garrison & Arbaugh, 2007), initially developed by Garrison and colleagues In the CoI framework, three types of presence are identified: cognitive presence, social presence and teaching presence:

Cognitive presence is defined as "the extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication" (Garrison & Arbaugh, 2007).

Social presence is defined as "the ability of people to project their personal characteristics into the community, thereby presenting themselves to the other participants as ''real people'" (Garrison & Arbaugh, 2007).

Teaching presence is defined as the activity "to support and enhance social and cognitive presence for the purpose of realizing educational outcomes". This includes teaching design, facilitating discourse and direct instruction (Garrison & Arbaugh, 2007).

Table 2.3: Potential intervention options (learning design vs. in-action interventions) (Rienties et al., 2016b, p6),

	Learning design (before start)	**In-action interventions (during module)**
Cognitive Presence	Redesign learning materials Redesign assignments	Audio feedback on assignments Bootcamp before exam
Social Presence	Introduce graded discussion forum activities Group-based wiki assignment Assign groups based upon learning analytics metrics	Organise additional videoconference sessions One-to-one conversations Café forum contributions
Teaching Presence	Introduce bi-weekly online videoconference sessions Podcasts of key learning elements in the module Screencasts of "how to survive the first two weeks"	Organise additional videoconference sessions Call/text/skype student at-risk Organise catch-up sessions on specific topics that students struggle with
Emotional Presence	Emotional questionnaire to gauge students emotions Introduce buddy system	One-to-one conversations Support Emails when making progress

Recent research has suggested the need for a fourth category, that of emotional presence (Cleveland-Innes & Campbell, 2012; Cleveland-Innes et al., 2014), recognising the importance of emotional interactions between students and academic staff. A recent literature review (Rienties & Rivers, 2014) identified 100 different emotions that may have a positive, negative or neutral impact on learners in online environments.

Purdue University recommend a smaller number of personal student intervention methods (Sclater at al., 2016), just 5, given that a multiplicity of intervention methods deployed by different instructors may be confusing to students:

- Post coloured traffic signal on student's VLE home page
- Send email or SMS
- Refer student to an academic advisor

- Refer student to resource centre
- Schedule F2F meeting

A concern they raise is that traffic light systems which generally categorise student progress across a variety of measures as green, amber or red may not have the desired positive effect. For example, a green indicator may give a false sense of security and amber may be confusing. These concerns may be minimised by publishing a very visible and clear statement of both the meaning of the colour and recommended alternative actions to students. Similarly, their concern that alerts such as "next assignment in 2 weeks" could be distracting may be addressed by very clear classification and presentation under a reminders or timetable label.

An analysis of 522 intervention messages sent to Purdue students were analysed anonymously in conjunction with their results data (Sclater, 2017b) and showed:

- There was no correlation between student success and the *frequency* of feedback
- Instructional feedback appeared to be more effective than motivational feedback
- Explicit feedback which compared students to their peers appears to be more effective than comparing them to standards
- Succinct messages appeared to have a more positive impact than longer ones

Through regular reports and workshops, Jisc has provided case studies of learning analytics based interventions across a variety of international educational institutions. In its review of UK and international practice for Jisc, Sclater et al., (2016a) presented eleven institutional case studies are from five US, four UK and two Australian universities. In most cases the output from the learning analytics is a dashboard or other type of alert for academic staff use, although some dashboard data is provided to students. Some interesting conclusions were drawn from the case studies; for example, at the University of Maryland in the US, students who chose to view their VLE activity compared with their peers were almost twice as likely to achieve grade C or above compared with those who did not. At New York's Marist College in the US, at-risk students who were the subject of an intervention achieved 6% higher grades compared with a control group who were not.

Sclater's report for Jisc identified nine types of student intervention (Sclater, 2017) as follows

- Reminders sent to students about suggested progression through the task
- Questions to promote deeper investigation of the content
- Invitations to take additional exercises or practice tests

- Attempts to stimulate more equal contributions from participants in a discussion forum
- Simple indicators such as red/yellow/green traffic signals, giving students an instant feel for how they're progressing
- Prompts to visit further online support resources
- Invitations to get in touch with a tutor to discuss progress
- Supportive messages sent when good progress is being made
- Arranging of special sessions to help students struggling with a particular topic

A number of these, for example, reminders of suggested progression, traffic signals of student progress, prompts to exploit on-line resources (e.g. VLE accesses) and supportive messages may be automatically generated as a result of learning analytic processes.

2.4 Data Mining Techniques

2.4.1 Artificial Intelligence and Machine Learning Techniques

Data mining (DM) is a technique for analysing and extracting data, correlations and patterns from large datasets and turning it into useful information (Sammut & Webb, 2017). It has become a very important tool in recent years as huge volumes of data have become available for analysis (so called "Big Data").

AI, ML and DM techniques may be used to analyse student progress, predict potential outcomes to their studies and therefore support academic staff in timely interventions. There are a considerable number of such techniques (Kubat, 2017) and (Tan, 2018) each with its own suitability to differing situations, objectives, datasets and of data types. Chapter Seven, Relevant AI and ML Techniques, section 7.2, details the thirteen directly relevant to this research, including the methods they use and their comparative advantages and disadvantages. Section 4.4 describes the applications and results of these techniques to selected datasets.

2.4.2 General Definition of Data Types

Data types are described as either numeric, formally referred to as measurement (quantitative) or categoric (Figure 2.9), (Everything About Data Science, 2015).

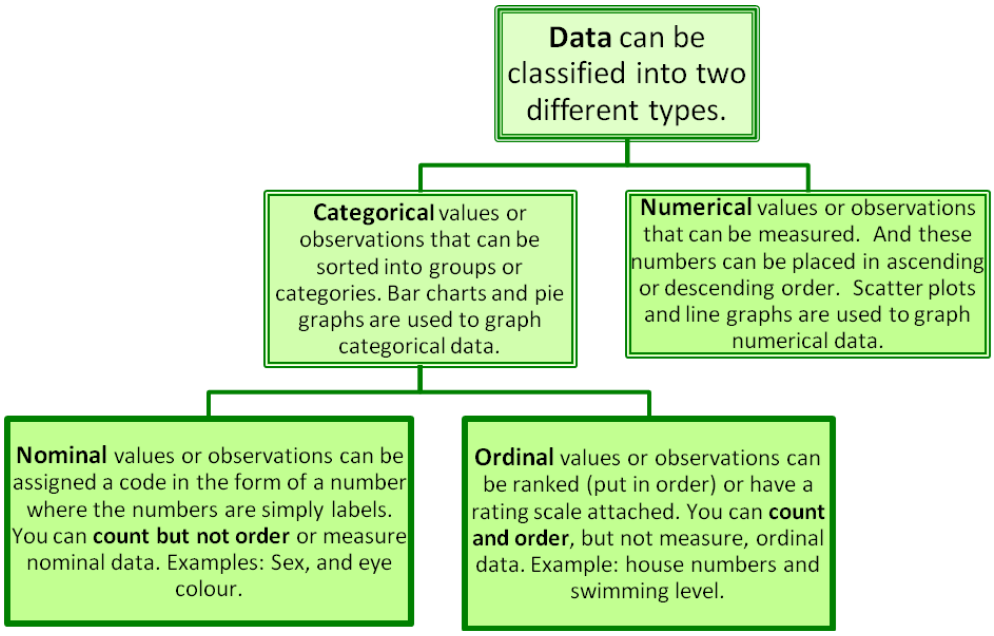


Figure 2.9: Types of Statistical Data: Numerical, Categorical, and Ordinal (Everything About Data Science, 2015, p1)

2.4.2.1 Measurement (Quantitative) Data

Quantitative data is defined as the value of data in the form of counts or numbers where each data-set has a unique numerical value associated with it. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis. This is often referred to as numeric data. The techniques suited to the analysis of numerical data include Support Vector Machine, Principal Components Analysis, Decision Trees, Random Forest and Neural Networks. These are discussed in Chapter Four Relevant AI and ML Techniques.

2.4.2.2 Categorical Data

Categoric data is defined as data which is identified as categories and for which no measurable value can be given, for example gender. The techniques suited to the analysis of categoric data include Contingency table and chi-square. Categoric data is comprised of two types, nominal and ordinal.

Nominal Data

Nominal data is data where the feature values are labels such as male/female or yes/no. There are a number of statistical techniques available to analyse nominal datasets, notably Chi-square (Agresti, 2002). These techniques are discussed in Chapter Seven Relevant AI and ML Techniques. Each has its own limitations, for example, sensitivity to sample size and a stronger than justified evidence of correlations (Bentler & Bonett, 1980).

In general, in the case of nominal data, it is not possible to compare attributes directly in order to search for correlations. However, we can compare the correspondence between groupings of attributes and we have explored the use of what we believe to be a novel technique to do so. In this case, we have chosen to compare correlations between pairs of attributes.

Ordinal Data

Ordinal data is a type of categorical data in which order is important, for example the Likert scale, where responses are typically "Like", "Like Somewhat ", "Neutral", "Dislike Somewhat", "Dislike". These are discussed in Chapter Four Relevant AI and ML Techniques.

2.5 The Importance of Pedagogy

Although pedagogy is conventionally defined as "the theory and practice of education" (Lewthwaite & Sloan, 2016), it usually includes the act of teaching itself and the associated policies and challenges (Papatheodorou & Potts, 2016).

Pedagogy continues to be an important area of research with significant on-going work into the field of Technology Enhanced Learning, alongside increased understanding of the behaviours and needs of both learner and tutor (Jenkins, et al., 2014). The Open University publish the series of Innovating Pedagogy reports identifying trends in education and AI (Ferguson et al., 2019).

The focus of pedagogy can be described as that of supporting positive student outcomes. The publication "What makes great teaching? Review of the underpinning research" (Coe et al., 2014) proposes six factors that they believe address the question with supporting evidence and an assessment of the strength of the evidence of impact upon student outcomes. The six factors that the authors have identified that support great teaching are shown in Table 2.4.

Table 2.4: Factors Supporting Great Teaching (Coe et al., 2014, p2)

Factor	**Evidence of impact on student outcomes**
(Pedagogical) content knowledge	Strong
Quality of instruction	Strong
Classroom climate	Moderate
Classroom management	Moderate
Teacher beliefs	Some
Professional behaviours	Some

This body of work, including a very wide variety of field trials and extensive data provides a firm foundation upon which to analyse existing TEL techniques, approaches and learning systems, and to identify the critical factors necessary for the successful definition, design and development of step-forward adaptive learning systems including subject matter knowledge classification. Modelling student performance and applying learning analytics is critical to the review of any application of pedagogical concepts as noted by Tempelaar, et al. (2015).

An exploration of the latest pedagogical research confirms the breadth and depth of formal understanding of the art and science of education available to the designers of learning systems, albeit with continuing adjustments being made to educational best practice.

An understanding of pedagogy is critical to effective teaching, with considerable recent and on-going research and experimentation into how to best exploit how people learn, including investigations into cognitive and learning styles. An interesting area of consideration is that of the value, or otherwise, of understanding an individual's learning style as a factor for exploitation in the development of learning systems. Learning styles can be defined as: "The composite of characteristic cognitive, affective, and physiological factors that serve as relatively stable indicators of how a learner perceives, interacts with, and responds to the learning environment" (National Association of Secondary School Principals (USA) and Keefe, 1979).

Many developers of learning systems consider that an understanding of the variety of individual learning styles is an important aspect (Graf, 2007). Graf's paper illustrates the considerable variety of research and opinion on an individual's learning approaches. A learner's cognitive style (the way an individual thinks, perceives and recalls information) is another key pedagogical concept where there is some

evidence that exploiting an understanding of these concepts has improved student learning achievement (Chipman, 2010).

This is an area for research and potential exploitation, although it is important to note that there is strong conflicting evidence on whether recognition of a student's learning style makes any difference when designing learning systems (as discussed by Mampadi et al., 2011). Despite the lack of evidence of the benefits of identifying and exploiting the student's learning style as part of technology enhanced learning systems implementations, an understanding of these styles may prove valuable in interventions. For example, a discussion between academic staff and student as part of an intervention process may provide an opportunity for the student to try an alternative mode of learning which may result in improved performance.

The usefulness of recognising and then looking to exploit learning styles has been the subject of considerable debate with recent research showing no evidence of benefits to learning from trying to present information to learners in their preferred learning style (Pashler et al., 2008; Geake, 2008; Riener & Willingham, 2010; Howard-Jones, 2014). This lack of any evidence contrasts strongly with the widely held view of practising teachers, where for example 93% of UK school teachers (The Netherlands 96%, Turkey 97%, Greece 96% and China 97%) as quoted by Howard-Jones (2014) appear to believe that individuals learn better when they receive information in their preferred learning style. While recognising that over 90% of teachers in various countries believe in the value of tuning teaching to learning styles Howard-Jones, (2014), Coe et al., (2014) cite research that shows that there is no evidence that this is the case.

Learning styles identifying a student's preferred way to learning has been an approach deployed in the development of adaptive e-learning systems (Truong, 2015). Using this knowledge, the system aims to adapt learning paths to best suit the student. Often, these systems rely upon a questionnaire approach rather than integrating machine learning/statistical detection methods into the system. Truong reviewed 51 studies (39 journal papers and 12 conference papers) which address different aspects of this integration process, including learning styles theories selection, online learning styles prediction, automatic learning styles classification and applications. The paper also provides discussion, recommendations and guidelines for future researches. Of the 51 studies reviewed, Felder–Silverman learning styles (Felder & Silverman, 1988) were the most popular theory applied. Here may be an opportunity for combining learning styles theories to achieve better results. A number of the papers point out that a learner's style may change over time, and therefore systems must recognise and respond to this.

In the last 30 years, over 70 theories (many overlapping) have been developed (Coffield et al., 2004), for example, Felder–Silverman's shares some dimensions with Kolb's (Kolb, 1981) and Riding's models. Secondly, according to Coffield et al. most learning styles' theories suffer some issues in terms of validity and reliability. Consequently, there is no single theory that can be shown to outperform others.

In recent years there have been a number of research papers casting doubt on the usefulness of tailoring teaching to the learning style of the student and in particular the absence of any evidence of correlation between learning style recognition and positive results from tailoring teaching accordingly. Pashler et al. (2008) describes it as "striking and disturbing" that the lack of evidence of the validity of teaching students based upon an assessment of their learning style has not been acknowledged by what they describe as the "widely held popular view". They cite several studies that used appropriate research designs which found evidence that contradicted the learning-styles hypothesis (Massa & Mayer, 2006; Constantinidou & Baker, 2002). In particular, they point out that the published research methods in favour of the hypothesis do not use the appropriate factorial randomised research designs essential to demonstrating evidence, for example, the classification of learners using clearly specified measures and then randomising the teaching approaches.

Riener and Willingham (2010) use the term "myth" in addressing the topic, while acknowledging the valid work of learning styles theorists in assessing how individuals learn, they attempt to make a logical case that this does not mean that the exploitation of the student's learning style in teaching has any benefit. However, research has shown that the consideration of learning style alternatives can provide students with the opportunity to reflect on how they learn, and to encourage them to adopt study strategies that may work better for them than their existing ones (Husmann & O'Loughlin, 2019).

A neuroscientific approach by Geake (2008), systematically deals with each of the common assumptions made in favour of learning style based teaching, for example, how the interconnectivity of brain functions such as working memory, decision making, emotional mediation etc. challenges the over-simplification of exploiting learning styles. He urges educators to seek independent validation before adopting what he describes as "brain-based" products in education.

Howard-Jones' 2014 paper in Neuroscience and Education echoes Geake's view, also pointing out that the brain's interconnectivity makes such an assumption unsound, and that reviews of the literature and controlled laboratory studies fail to support this approach to teaching. Geake likened this belief, despite the lack of evidence, to "cargo cult science" (Feynman, 1974) where popular hypotheses are adopted without rigorous scientific examination for evidential results.

The Learning & Skills research centre report “Learning styles and pedagogy in post-16 learning: A systematic and critical review” (Coffield et al., 2004) critically reviews the 13 most influential learning style models. Each of the 13 style models were reviewed against the minimal criteria of internal consistency, test-retest reliability, construct validity and predictive validity, with only one model meeting all four criteria, the Allinson and Hayes’ Cognitive Styles Index (Allinson & Hayes, 1996). Attention is drawn to the lack of diligent, independent investigation and hence evidence of the value of learning styles, and the report advises educators against pedagogical intervention based solely on any of single learning style instruments.

Husmann & O'Loughlin (2019) research provides further evidence that the conventional wisdom about learning styles should be rejected by educators and students alike.

2.6 Technology Enhanced Learning Systems

Technology enhanced learning (TEL) may be most simply defined as the support of teaching and learning through the use of technology (O'Donnell & O'Donnell 2015). It is often used synonymously with the term e-learning.

The commercial world is facing critical challenges in the training, development and retention of key skills, exacerbated by new, emerging technologies and business models, giving organisations business critical dependencies on the relevant subject matter experts (SMEs) and on leadership/talent development (Bhatia & Kaur 2014). These challenges are presenting a major threat in many organisations, limiting business opportunities and weakening their ability to compete (Schuler et al., 2011). Developments in TEL and in particular in the progress of adaptive learning systems have the potential to make a dramatic difference in addressing these challenges.

The field of TEL has been the subject of much research and practice, in a very wide range of techniques and approaches ranging from classroom management and collaborative learning to MOOCs and gamification. An analysis of TEL research in Higher Education published between 2009 and 2014 (Schweighofer & Ebner, 2015) recorded over 4500 papers, dealing with aspects from demographical differences to learner/teacher issues and technical infrastructure.

Commercial organisations are increasingly automating their training programmes to allow them to be delivered globally, asynchronously and electronically (Chang, 2016). This was my own experience (in executive roles} during the final 20 years of my career at two global corporations, Fujitsu and Unisys. These training modules can be stand-alone or part of a classroom based blended learning package and are ideal for situations where a large number of geographically separated leaners are targeted. Typically,

these modules are delivered as on-line question and answer based dialogues, presenting the learner with explanatory information, occasionally including video material, followed by marked exercises. The learner repeats the course until the pass level is reached and at each subsequent re-take the questions are varied from a set database.

In the UK Higher Education (HE) sector, progress in the numbers of on-line courses available to students has been modest in relatively recent years (see Table 2.5), giving rise to concerns that the investments in TEL are not addressing pedagogical needs (Jenkins, et al., 2014). Disappointingly, the table shows that the proportions of modules/units of study delivered in a TEL environment were broadly static between 2012 and 2014. More recent surveys by UCISA (Universities and Colleges Information Systems Association) of TEL and in particular VLE deployment in the UK, while promising in terms of some progress are not showing major changes in the way that technology is being used to support learning, teaching and assessment activities (Walker et al., 2018).

Table 2.5: Proportion of all modules or units of study in the TEL environment in use across the UK HE sector (Walker et al., 2014, p35)

Sector mean	**2014**	**2012**	**2010**	**2008**	**2005**	**2003**
Category A – web supplemented	39%	39%	46%	48%	54%	57%
Category Bi – web dependent, content	27%	29%	26%	24%	16%	13%
Category Bii – web dependent, communication	9%	10%	17%	13%	10%	10%
Category Biii – web dependent, content and communication	21%	18%	18%	13%	13%	13%
Category E – fully online	3%	3%	3%	4%	6%	5%

However, the 2014 summative HE Academy report (Barnett 2014) on flexible technologies observed that the drive towards greater flexibility was being influenced by a combination of the marketization of HE, including MOOCs, the demands of students as consumers, the potential of new technologies and the apparent potential for making HE available to a wider audience at lower unit costs.

Schweighofer & Ebner's (2015) recent analysis of 4567 TEL publications between 2009 and 2014 (recognises the breadth and depth of on-going research into TEL approaches, summarising key aspects to be taken account of in TEL implementation. These analyses show learner's aspects as the largest focus of research in the more technologically focused publications.

In the future it is likely that it will be the demands and imperatives of the students/learners that prove to be a major driver in TEL adoption, not only for its educational merit, but in order to enable them to support the stresses of combining work, study and personal life (Jefferies & Hyde 2010). Additionally, trends in social media, the integration of on-line, hybrid and collaborative learning alongside the rise of data driven learning and assessment are strong pressures for increasing the adoption of TEL in HE (Johnson et al., 2014).

Chapter Five presents a survey of existing intelligent learning/training systems in each of the education and commercial sectors, including those applicable to both, categorising each as Adaptive Learning Systems (ALS) or Intelligent Tutor Systems (ITS). Definitions of each of these systems are discussed below.

2.6.1 Adaptive Learning System

The field of adaptive learning has allowed these systems to develop a close relationship with the learner, monitoring and adjusting the teaching and creating idealised learning paths based upon a wide variety of analyses of their knowledge and performance (Marengo, et al., 2015).

This level of automated judgement is made by understanding the learner profile, their learning preferences and their base knowledge of the subject area (Marengo, et al., 2015).

In designing adaptive learning systems there are a significant number of potential techniques and models which can be deployed. Recent research into the prevalence of these show learner and domain knowledge modelling, adaptability and content presentation as the most prevalent in learning systems, with cognitive style almost the least characterised (Marković, et al., 2014).

In the US there is positive evidence of the increasing adoption of such systems. As discussed in section 3.3, the challenges are organisational and not technological (Oxman & Wong, 2014).

Additionally, some progress has been made in the area of adaptive learning systems in the commercial area, with research into the benefits and risk areas from the learner's point of view. The results indicated a positive response to the alignment of adaptive learning to job roles and career paths, while removing the time wasted on non-relevant learning material. The research also reinforced the criticality of the input and capture of expert knowledge (Höver & Steiner, 2009).

2.6.2 Intelligent Tutor System

The line between Intelligent Tutoring Systems (ITS) and Adaptive Learning Systems (ALS) has become increasingly blurred. In the past ITSs tended to be subject matter specific, developing from what can be described as "flowcharted learning" into increasingly sophisticated systems deploying AI techniques.

The typical major components of an ITS (Clement, et al., 2014; Nkambou, et al., 2010) are:

Cognitive model: This is sometimes referred to as the domain model. It contains the necessary subject matter knowledge (declarative knowledge) including the rules and processes that a subject matter expert (SME) will deploy in order to solve problems (procedural knowledge). Note that this subject matter knowledge is rarely static, particularly in commercial or emerging/developing subject areas, and provision must be made for periodic SME update.

Student model: This contains information on the individual learner including their base knowledge, cognitive skills, and progress. This model is dynamic, using real time and historic data to create an up to the minute representation of the learner's knowledge and learning process, which facilitates the choice of the appropriate pedagogical strategies to deploy in order to diagnose and consequently address knowledge gaps, to correct misconceptions/errors, and to elaborate partly complete learner understanding. The model will also predict the student's responses, initiate changes in the teaching strategy and evaluate the student's progress.

Tutoring model: This model exploits data from the cognitive and student models in order to make decisions on the learning paths, strategies and training activities to govern the learner.

User interface model: This model manages all interaction with the learner. It will deploy various different forms of content delivery and communication styles including simulations, hypermedia, and micro-worlds. A major body of related research is that of Natural Language Processing, however this field has yet to deliver the advances anticipated thirty years ago and is not included in this research.

Populating the cognitive model is traditionally an SME activity, however significant advances in the field of educational data mining (EDM) are providing opportunities for the data mining tools to be deployed in mining educational data, including student and institutional education data (Fatima D, et al., 2015) and to a modest extent to date in populating the subject matter itself. A good example of the latter is web data (content) mining which allows organisations to better link relevant information to their own web site (Kaur & Chawla 2014). The paper, Analysis & Survey of Different Data Mining Techniques for Predicting Student's Performance (Parmar & Khalpadacan, 2015) includes an informative table comparing the features, advantages and limitations of various data mining techniques.

Some of the first commercial successes in learning systems in the US came from cognitive tutoring systems which delivered high school mathematics to over 475,000 students in 2007 (Raley 2012), showing that students performed 15-25% and 50-100% respectively better than the control group on skill knowledge and problem solving.

2.7 Chapter Summary

In this chapter I have reviewed literature relevant to each of my research questions. In the following chapter I examine progress on the development of intelligent learning/training systems in education and commercial sectors and consider institutional challenges and barriers to the implementation of learning analytics systems, including critical success criteria. I present a survey of existing intelligent learning/training systems in each of education and commercial sectors, categorising each as Adaptive Learning Systems (ALS) or Intelligent Tutor Systems (ITS). I then compare these results with an equivalent survey conducted in 2015. I catalogue the organisational/non-technological obstacles and challenges that must be addressed for the development of such systems successful system development and summarise critical success criteria.

CHAPTER THREE

Intelligent Learning/Training Systems

3.1 Introduction

In this chapter I describe my research activities exploring intelligent learning/training systems in each of education and commercial sectors.

I survey existing intelligent learning/training systems in each of education and commercial sectors, including those applicable to both, categorising each as Adaptive Learning Systems (ALS) or Intelligent Tutor Systems (ITS). This 2019 survey is compared with the equivalent survey conducted in 2015 (Wakelam et al., 2015).

Using available research and analyses of system design and implementations and my own experience in the software industry I catalogue the organisational/non-technological challenges that must be addressed for successful system development. These range from organisational and political obstacles to academic staff and student concerns and needs.

I have proposed critical success criteria to apply to the development and use of e-learning systems based upon available research and my own experience in the systems and software development industry.

The following sections of this chapter are supported by previously published material:

> Section 3.2 Surveyed intelligent learning/training system products and prototypes (Wakelam et.al., 2015)
>
> Section 3.3 System challenges and barriers to success (Wakelam et.al., 2015)
>
> Section 3.4 System success criteria (Wakelam et al., 2016)

3.2 Surveyed Intelligent Learning/Training System Products and Prototypes

A number of systems, mostly niche, have been developed and are in place in the field, alongside a variety of prototypes. This research provided me with an understanding of which techniques are deployed by these systems to predict student progress. For example, SHERLOCK (Lesgold et al., 1988) and Realizeit (Realizeit, 2015) use Decision Trees and Realizeit also uses Fuzzy Logic. Cardiac Tutor (Cardiac Tutor, 2019) is a Knowledge Based System. In addition, it has provided me with insights into potential intelligent intervention methods, for example adaptive learning paths and reinforcement teaching material recommendations. A list of systems applicable to each of the education, commercial and combined

sectors is shown in Tables 3.2, 3.3 and 3.4. As can be seen in the summary metrics Table 3.1, systems in the education sector dominate. Respective home page/web links to each system are listed in Appendix H.

Table 3.1: Survey of Intelligent Learning/Training Systems Identified

Sector	**2019**		**2015 (Wakelam et al., 2015)**	
	Quantity	**Percentage**	**Quantity**	**Percentage**
Education sector	36	68%	32	78%
Commercial sector	5	9%	3	7%
Education & Commercial sector	12	23%	6	15%
Total	53	100%	41	100%

Of those surveyed, 30 (57%) have been developed by universities or as collaborative projects between university and industry. Over half (58%) are adaptive learning systems (highlighted in green), the details of which are shown in Tables 3.2, 3.3 and 3.4 (the definitions of each type of intelligent learning system are given in Chapter Two, Literature Review).

Comparing with the 2015 survey (Wakelam et al., 2015) we may observe a modest increase of intelligent learning/training systems in each of the commercial sector, by 2 percentage points, and combined education and commercial sectors, by 8 percentage points. This may indicate an increasing recognition in the commercial sector in the potential business value of further intelligent automation of their training systems.

Greatest progress appears to be where the knowledge base being addressed is embodied in comprehensively curated areas of knowledge, for example, STEMM subjects including mathematics and physics, and English education.

Table 3.2**:** Intelligent Learning/Training Systems in the Education Sector

No.	**System**	**Developed by**	**Type**	**Key words**
1	**ActiveMath [P, J, S]**	DFKI & Saarland University	Adaptive learning	Educational data mining. Natural Language Processing. Collaborative. STEMM.

No.	System	Developed by	Type	Key words
2	**ALEKS [P, J, S, U]**	New York University and the University of California, Irvine	Adaptive learning	Web based. Knowledge space theory. STEMM, Accounting.
3	**Algebra Tutor [S]**	Carnegie Mellon	Intelligent tutoring	Artificial intelligence, cognitive, human computer interaction. Computer programming, STEMM.
4	**Andes Physics Tutor [S, U]**	Arizona State University	Intelligent tutoring	Highly interactive. STEMM.
5	**Aplia [U, Po]**	Stanford university	Adaptive learning	On-line homework system. Multiple subjects - STEMM, accounting, English, history, finance.
6	**ASPIRE [J, U]**	University of Canterbury (New Zealand)	Intelligent tutoring	Authoring. Develops web tutoring systems.
7	**AutoTutor [U]**	University of Memphis	Intelligent tutoring	Natural language. Speech engine. Newtonian physics, Introductory computer literacy.
8	**Betty's Brain [P, S]**	Vanderbilt & Stanford Universities	Intelligent tutoring Learning by teaching	Metacognitive skills. STEMM.
9	**Carnegie Learning [S]**	Carnegie Mellon University	Adaptive learning	Pedagogy. Cognitive science. Research led. STEMM.
10	**CIRCSIM-Tutor [U]**	Sponsored by US Naval Research Office	Intelligent tutoring	Dialogue based, natural language. Medicine.
11	**COLLECT-UML [U, P, A]**	University of Canterbury (New Zealand)	Intelligent tutoring	Teaches object- oriented design using Unified Modelling Language (UML).
12	**DreamBox [P, J]**	DreamBox	Adaptive learning	Game-like environment based. STEMM.

No.	System	Developed by	Type	Key words
13	**EER-Tutor [U, P, A]**	University of Canterbury (New Zealand)	Intelligent tutoring	Teaches conceptual database design.
14	**ESC101-ITS [U]**	The Indian Institute of Technology, Kanpur, India	Intelligent tutoring	Programming.
15	**eSpindle [P, J, S]**	LearnThat	Personalised learning	US Spelling Bee system. Spelling.
16	**eTeacher [S, U]**	eTeacher	Adaptive learning	Intelligent agent. On-line assisted learning. System engineering course.
17	**Grockit [S]**	Kaplan	Adaptive learning	Collaborative. Game-like environment. STEMM.
18	**Knewton [S, U]**	Knewton	Adaptive learning	Content agnostic. Psychometrics and cognitive learning theory, Inference engine.
19	**Knowledge Sea II [U, Po]**	University of Pittsburgh	Adaptive learning	Computer programming.
20	**KnowRe [J, S]**	KnowRe	Adaptive learning	Game-like environment based. STEMM.
21	**LearnSmart [S]**	McGraw Hill	Adaptive learning	Classroom teaching tool. Science, Social Studies, Spanish
22	**Mathematics Tutor [J, S]**	University of Massachusetts	Adaptive learning	STEMM.
23	**Mathspring [P, J, S]**	Univ. of Massachusetts	Adaptive learning	Intelligent tutoring. Math.
24	**Memorangapp. [U, Po]**	MIT	Memory reinforcement.	Spaced repetition. Medicine.
25	**MyLab, Mastering [U, Po]**	Pearson	Adaptive learning	On-line learning. Multiple subjects.
26	**PlanetSherston [P]**	Sherston	Personalised learning	Game play learning.

No.	System	Developed by	Type	Key words
27	**PrepMe [S]**	Stanford, University of Chicago, CalTech	Adaptive learning	Virtual classroom. STEMM.
28	**PrepU [U, Po]**	PrepU, collaboration with UCLA	Adaptive learning	Quiz engine. STEMM.
29	**REALP [J, S]**	Worcester Polytechnic Institute, Carnegie Mellon	Personalised learning	Based upon a tool designed to investigate the development time for tutoring systems. Reading comprehension.
30	**Scootpad [P, J, S]**	Scootpad	Adaptive learning	Behaviour tracking. Prediction. STEMM.
31	**SmartTutor [A]**	University of Hong Kong	Adaptive learning	Personalised on-line distance learning. Generic.
32	**Snapwiz [U, Po]**	Wiley	Adaptive learning	Collaborative. STEMM, Languages, Business, Social Science.
33	**SpellBEE [P, J, S]**	Brandeis University	AI Machine learning Game theory	Education research tool.
34	**SQL-Tutor [U]**	University of Canterbury (New Zealand)	Intelligent tutoring	Computer programming. Web enabled.
35	**Why2-Atlas [U]**	UCLA	Intelligent tutoring	Textual analysis system. STEMM.
36	**ZOSMAT [J,S]**	Atatürk University	Intelligent tutoring	Classroom based. STEMM.

[Key: Primary, Junior, Secondary, University, Postgraduate, Adult

Adaptive Learning System (ALS) Intelligent Tutoring System (ITS)

Personalised Learning System (PLS) Other]

Table 3.3**:** Intelligent Learning/Training Systems in the Commercial Sector

No.	System	Developed by	Type	Key words
1	**3KEYMASTER**	Western Services Corporation	Intelligent tutoring	Simulator based training.
2	**Cerego Global**	Cerego	Adaptive learning	Domain independent. Corporate training.
3	**CODES**	Universidade Federal do Rio Grande do Sul	Personalised learning	Web-based. Musical prototyping specific for non-musicians.
4	**CogBooks**	CogBooks	Adaptive learning	Domain independent. Corporate training.
5	**SHERLOCK**	University of Pittsburgh	Intelligent Tutoring System	Decision trees. Student competence and performance model. USAF technician specific.

Table 3.4**:** Intelligent Learning/Training Systems in the Education & Commercial Sector

No.	System	Developed by	Type	Key words
1	**Adaptive 3.0 Learning Platform**	Fulcrum Labs	Adaptive learning	Domain independent
2	**Alelo**	University of Southern California	Virtual Role-Play simulations	Pedagogical agents as social actors. Multimedia. Cyber learning.
3	**aNewSpring**	aNewSpring	Adaptive learning	Corporate Learning Management System. Blended and hybrid learning
4	**Cardiac Tutor**	University of Massachusetts Medical School	Adaptive learning Intelligent tutoring	Real time simulation. Knowledge based. Medicine, cardiology specific.
5	**Desire2Learn, LeaP**	Brightspace	Adaptive learning	Predictive analytics.

No.	System	Developed by	Type	Key words
6	**ELM-ART**	Freiburg University of Education	Adaptive learning	Web-based. LISP programming specific
7	**Generalized Intelligent Framework for Tutoring (GIFT**	U.S. Army Research Laboratory	Intelligent tutoring	Open-source, domain independent, and can be downloaded online for free. Allows tutor to design domain specific tutoring program.
8	**Navigate 2**	Jones & Bartlett Learning	Adaptive learning.	Health, fitness and sport. STEM.
9	**OER & Competency Learning Platform**	LoudCloud	Adaptive learning	Domain independent.
10	**Oracle Intelligent Tutoring System (OITS)**	Al-Azhar University, Gaza, Palestine	Adaptive learning. Intelligent tutoring	Oracle programming training system.
11	**Realizeit**	CCKF/Realizeit	Adaptive learning	Content agnostic. Supervised & Unsupervised learning. Classification decision trees. Fuzzy Logic.
12	**Smart Sparrow**	University of New South Wales in Sydney	Adaptive learning Intelligent tutoring	Educational data mining. Content agnostic.

Although these systems are dominated by those focussed upon the education sector, we should expect increasing interest from the commercial world, since individuals may be faced with a number of different careers during their working life as industries are created, evolve and disappear. The development of new and more intelligent methods of supporting these aspirations will become very important to both individuals and organisations, presenting the opportunity to deliver significant value, in terms of reducing training and re-validation costs, in accelerating training delivery and in considerable enhancement of people's personal experience in learning.

In terms of organizational traction, analysis of existing systems shows that the field of education is leading the way in both research and in the development of learning/training systems, aimed at primary, secondary, and university education. STEMM is a popular subject area (Table 3.2).

Commercial research and learning/training systems traction is currently running a poor second (Tables 3.3 and 3.4) with Health Care, including diagnosis and training, appearing more often than others in the application of intelligent techniques to areas.

Many of the systems surveyed highlight their strengths in supporting distance learning, suggesting this to be an early TEL driver.

In terms of geographic traction, it is the highest in the US, followed by the UK, followed by Europe, with Australia and New Zealand showing up intermittently in searches.

3.3 System Challenges and Barriers to Success

While the adoption of TEL continues to gain traction, there are a number of organisational/non-technological challenges that must steadily be addressed and in particular kept in mind in the design, development and deployment of these systems (Table 3.5). I have compiled this list from the materials referenced and personal experience of systems design and implementation during a 40 year systems implementation career. In the following section I map these challenges against those identified for general systems development and implementation and describe potential mitigations. The awareness and investigation of institutional barriers to the large scale adoption of learning analytics have been identified since at least 1979, including the conservative culture of Higher Education institutions (Ferguson et al., 2014).

Table 3.5: E-learning Systems Challenges

	Challenges to success
Organisational	Systems can be expensive both to develop and to implement.
	Organisational conservatism – the prevailing attitude of "what we have works fine..", and the need to evidence benefits.
	Requires the cooperation and support of individuals across both organisations and organisational levels (Barnett 2014)
Administrative/political	Integration of TEL into the existing curriculum (Oxman & Wong 2014).
	Overcoming resistance from competing methods and their champions.
	Teacher/trainer resistance – the need for persistence while under significant pressure to deliver improved student grade performance dealing with high workloads (Wang & Hannafin 2005).
	Requires the cooperation and input of domain subject matter experts.
	Ensuring student/learner motivation and early identification of disenchantment (Oxman & Wong 2014).
	Continuous feedback to ensure the maintenance of a continuously accurate student model (progress measurement, learning rates, proven alternative learning paths).
Technical	The modelling of such a complex cognitive task.
	Incorporating the essential pedagogy. For example, effective feedback to the learner and very careful use of hints to ensure that deep learning is developed.
	Integration with all user platforms - mobile, fixed, on-line/off-line, social.
	Ability to exploit rapidly developing technologies/platforms.
	Necessity of systematic and regular update of domain subject matter.

3.4 System success criteria

Given the importance of this topic in the development of commercial systems there is a large body of material available. The DeLone and McLean model of information systems success is often drawn upon in research in this area (DeLone & McLean, 2003). This model defines 6 inter-related success measures (Table 3.6).

Table 3.6: Measure of Systems Success (DeLone & McLean, 2003, p17)

Measure	**Categories**
Technical	Systems quality
Semantic	Information quality
Effectiveness	Use User satisfaction Individual impacts Organisational Impacts

Note: (Shannon & Weaver, 1949) defined the *technical* level of communications as the accuracy and efficiency of the communication system that produces information. The *semantic* level is the success of the information in conveying the intended meaning. The *effectiveness* level is the effect of the information on the receiver.

A key conclusion of their work is that these components are highly interdependent (Figure 3.1).

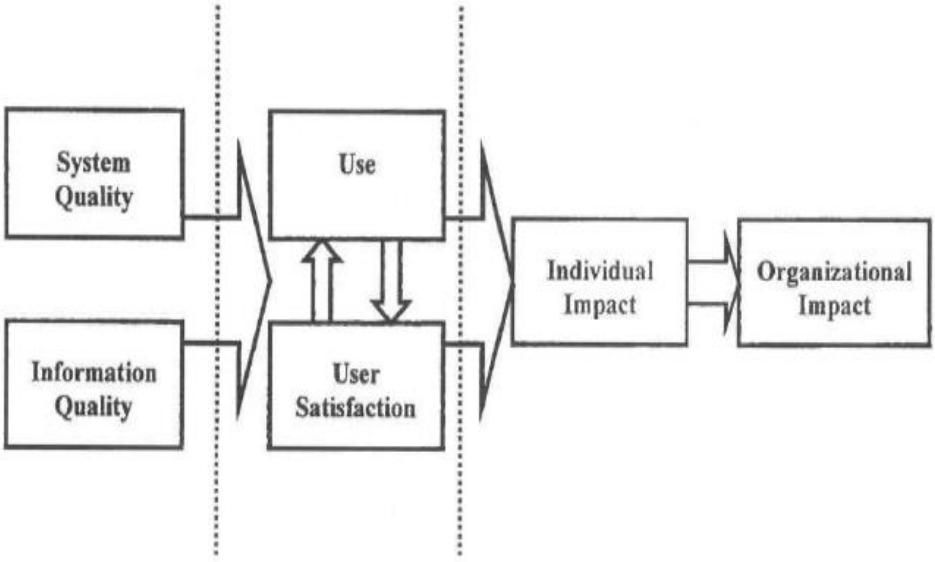


Figure 3.1: Interdependency of Components (DeLone & McLean, 2003, p12)

This model was extended (Wang et al., 2007; Wu & Wang, 2006) to encompass these six dimensions:

- Information quality
- System quality
- Service quality
- Use/intention to use
- User satisfaction
- Net benefits

This revised model is now regarded as one of the most widely used models of information systems success and has been used for various information systems (Hassanzadeh et al., 2012). The corresponding conceptual model is shown in Figure 3.2.

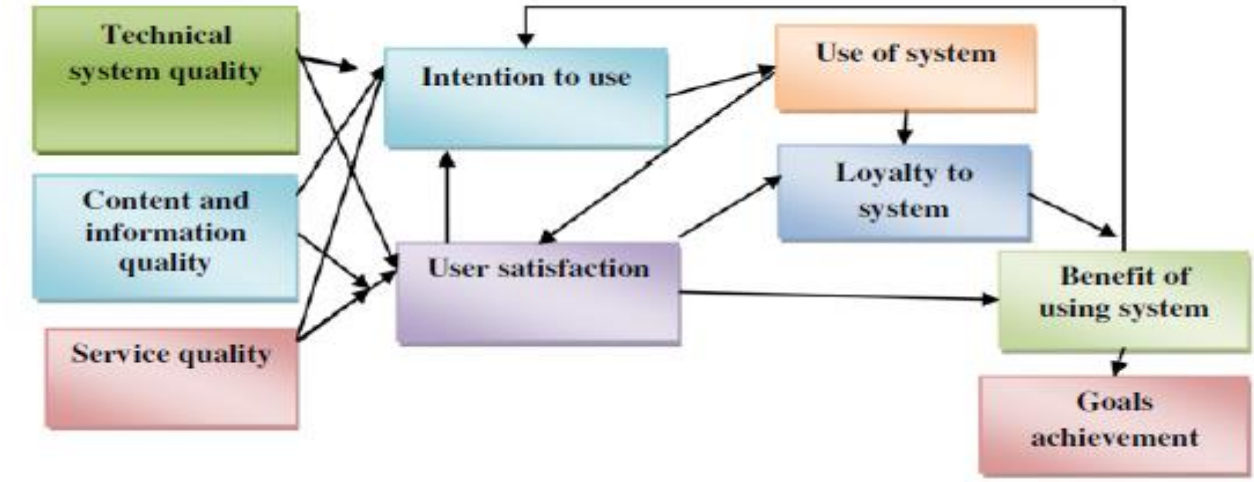


Figure 3.2: Conceptual Model (Hassanzadeh et al., 2012,p2)

I have mapped the nine Hassanzadeh success criteria against those identified by my own research work coupled with my own experience in the systems and software development industry (Table 3.7) as another method of validation. The results show a promising level of correspondence and may provide learning system designers, implementers and operational management with guidance.

Table 3.7: Mapping of e-learning System Challenges vs Success Criteria defined by (Hassanzadeh et al., 2012, p2)

E-learning Systems Challenges (Table 5.5 (Above))	**Success Criteria (Hassanzadeh et al., 2012)**	[1]
Organisational		
Systems can be expensive both to develop and to implement.	Benefits of using the system	
	Technical system quality	
	Service quality	
Organisational conservatism – the prevailing attitude of "what we have works fine..", and the need to evidence benefits.	Benefits of using the system	
	System operational quality	
	Goals achievement	

Requires the cooperation and support of individuals across both organisations and organisational levels.	Benefits of using the system	
	System quality	
	User satisfaction	
	Intention to use	
	Use of system	
	Loyalty to system	
	Goals achievement	
Administrative/political		
Integration of system into the existing curriculum (Oxman & Wong 2014).	Benefits of using the system	
	Operational system quality	
	Technical system quality	
Overcoming resistance from competing methods and their champions. The needs and concerns of users.	Benefits of using the system	
	Educational system quality	
	Content and information quality	
	User satisfaction	
	Goals achievement	
Management/staff/user resistance – the need for persistence while under significant pressure to deliver improved business performance dealing with high workloads	Benefits of using the system	
	Goals achievement	
	System quality	
	Content and information quality	
	User satisfaction	

	Intention to use	
	Loyalty to system	
Requires the cooperation and input of domain subject matter experts. The needs and concerns of the users	Benefits of using the system	
	Goals achievement	
	Content and information quality	
	Intention to use	
	Loyalty to system	
Ensuring user motivation and early identification of disenchantment (Oxman & Wong 2014).	Benefits of using the system	
	User satisfaction	
	Intention to use	
	Use of system	
Ensuring user motivation and early identification of disenchantment (Oxman & Wong 2014). (Continued)	Loyalty to system	
	Content and information quality	
	Goals achievement	
Continuous feedback to ensure the maintenance of a continuously accurate business model	Use of system	
	Loyalty to system	
	Goals achievement	
	System quality	
	Content and information quality	
Technical		
The modelling of such a complex cognitive task.	Technical system quality	
	System quality	

	Content and information quality	
Incorporating the essential business knowledge.	System quality	
Integration with all user platforms - mobile, fixed, on-line/off-line, social.	Technical system quality	
	Service quality	
Ability to exploit rapidly developing technologies/platforms.	Technical system quality	
	Service quality	
Necessity of systematic and regular update of domain subject matter.	Content and information quality	
	System quality	

[1] Colour coding to visually highlight matched Success Criteria. For example, all occurrences of "System quality" are coded yellow.

A comprehensive research review of success factors with specific focus upon E-learning systems is provided by Wang et al., (2007). The research gathered data from eight international organisations, including 206 individual e-learner responses ranging from top-level managers to general employees. The respondents completed a 37 question Likert scale questionnaire and after a wide analysis of research papers Wang selected the revised DeLone and McLean model as the basis for constructing a validated 34-item E-learning Systems Success (ELSS) measurement tool. Wang et al.'s paper aims to develop and validate a generic instrument for measuring e-learning systems success. ROMA (RAPID Outcome Modelling Approach) provides a very useful framework to support policy and strategy processes complex systems development, focussing upon evidence–based policy change (Young et al., 2014).

In my own systems experience these formal methods often struggle to take full account of the human factors in systems development, however, Wang et al., DeLone & McLean and ROMA provide valuable insights into useful techniques. Recent OU research into Human Centred Learning Analytics (HCLA) provides further supporting evidence of the challenges in the design and implementation of learning analytics systems (Buckingham et al., 2019).

3.5 Chapter Summary

In this chapter I have presented a survey of existing intelligent learning/training systems in each of the education and commercial sectors, comparing the results with the equivalent survey in 2015 in order to examine progress. Most notable is the increased percentage of system implementations or prototypes in the commercial sector, an increase of 10 percentage points to 32%. This trend of more investment in this sector may prove beneficial in the case of educational learning analytics in its likely cross fertilisation of ideas and techniques. These systems track student progress in real-time, applying learning analytic techniques to measure students' progress and personalise their teaching through reinforcement learning, modification of learning paths and tutor/trainer alert. The techniques and measurement of student attributes mirror and are directly relevant to research into learning analytics. As is the case in any major computer system design and implementation, the deployment of learning analytics in educational institutions must overcome a variety of challenges and barriers to success. Using available research and my own experience in the software industry I have catalogued these challenges and critical success criteria, including a mapping between the two. The successful deployment of any learning analytics and intervention system is critically dependent upon executive management, design and implementation management acknowledgement and implementation of these principles. In the following chapter I discuss student and institutional impacts of student withdrawals and explore the potential factors affecting student performance, followed by the variety of different techniques of identifying students at risk.

CHAPTER FOUR

Identification of Students at Risk

4.1 Introduction

4.1.1 Contributions to Knowledge Relevant to this Chapter

In this chapter I describe my research into the factors affecting student performance and the methods applied by academics to identify students at risk. These activities support my contribution demonstrating how the analysis of these limited attributes: attendance, VLE accesses and intermediate assessments, may provide potentially useful intervention guidance to academic leadership.

The following sections of this chapter are supported by previously published material:

Section 4.2 Problem to be addressed (Wakelam et al., 2020)

Section 4.4 Identification of Students at Risk (Wakelam et al., 2020)

4.1.2 Summary of Chapter Content

I explore and catalogue the human and financial impacts on institutions of students' failure to progress in their studies, collecting and contrasting the potential social, institutional and pedagogical factors affecting student performance. I consider each of traditional non-computational and then computational methods of the identification of students at risk.

4.2 Problem to be Addressed

The identification of students at risk has become increasingly important to academics, tutors, support staff and institutions, for a variety of reasons. For the students themselves, the failure to achieve their potential is a waste, as is the consequent limitations on their future career development. Worse is the personal stress and trauma they consequently face, alongside the potential impact on their families.

For institutions, the financial impacts can be very significant, compounded by the consequential effects of published statistical measures of student drop-out rates and student satisfaction scores. In the UK for example, in academic year 2015/16, 6.4% of UK domiciled full-time entrants did not continue in their studies after their first year (HESA, 2018a). In Australia and the US, these figures are worse with attrition rates of over 21% (Australian Government Department of Education and Training, 2016) and over 25% (Digest of Education Statistics, 2017).

Universities operate a sliding scale of refund levels to be applied should a student leave the course. In the case of the author's own university (University of Hertfordshire, 2019b), the cost of refunds for full time

UK and EU undergraduate student withdrawals based upon a 3 year, full time, undergraduate degree can be as high as £27,750 (Table 4.1). See Appendix C for extract from University of Hertfordshire student refund and liability dates.

Table 4.1: UK Student Refunds for Course Withdrawal during Semester A in Academic Year 2019/20

From Year 1 Semester A Commencement	Refund % of Semester Fee	Refund Value	Financial Impact on University if Year 1 Withdrawal[1]
Day 24	100%	£3,083	£27,750
Day 99	75%	£2,312	£26,980
Day 204	50%	£1,542	£26,209
Day 205 onwards	0%	£0	£24,667

[1] For simplicity, the effects of inflation over the course of a three year degree are excluded from the annual student fee calculations and therefore.

Should a UK/EU student withdraw during the initial 6 week period in Semester A, the financial impact on the university is £27,750. This assumes that the university place cannot be filled by a suitable replacement and is based upon current annual fees of £9,250. The financial impact of withdrawals in subsequent semesters and years ranges from £18,500 to £1,542.

To demonstrate the substantial financial impacts of first year undergraduate student failure to progress with their studies upon the budgets of a typical institution and the UK Higher Education system as a whole, we may apply some simple arithmetic (Table 4.2).

Table 4.2: Financial Impacts of First Year Student Withdrawals

	First year enrolment 2015/16[1]	% withdrawals during first year 2015/16[2]	Withdrawals during first year	Best and worst case total financial impact[3,4]
Typical University	6,280	6.4%	401	£7.2M - £10.8M
UK	542,575	6.4%	34,724	£625.0M - £937.5M

[1] (HESA, 2018b)

[2] (HESA, 2018a)

[3] Using University of Hertfordshire student fee for 2015/16 of £9,000 per annum as the UK average, assuming a 3 year undergraduate degree and UK or EU student

[4] Range calculated from "best case" of student withdrawal after first year fees paid in full to "worst case" of withdrawal with no fees due.

The financial impacts at institutional and UK level are very substantial indeed and suggests that the case for the identification of students at risk and positive intervention is a compelling one. If the successful application of modern analytical methods and consequent interventions were to result in even a modest reduction in student withdrawals, the financial benefits would be significant. For example, a 10% reduction in student withdrawals would improve institutional and UK budgets by between £0.7M - £1.1M and £63M - £94M respectively.

In comparison, fees for international (non UK/EU) students are £11,950 per annum, 21.6% higher than UK/EU. In this case, the financial impact on the university ranges from nil to £33,750 per student, depending upon withdrawal points.

Furthermore, universities operate in a very competitive environment, and pay considerable attention to their place in league tables and how they may improve their position. The student satisfaction score is an integral part of each institution's overall score and is therefore an area of focus for university management and policies. The prevalence of social media gives students at risk the opportunity to express their satisfaction/dissatisfaction at any time during their degree studies. In addition, student failures have a detrimental effect upon league table placings.

In the modern education system student non-attendance at lectures and tutorials remains high (Marburger, 2001; Mearman et al., 2014) as course material has increasingly become available on-line and accessible

to students 24/7. This reduction in face to face engagement between educators and students makes it increasingly difficult for tutors to identify students at risk who are struggling with the material or failing to engage. The use of learning analytics to support academic staff in identifying students at risk can provide some mitigation of this challenge.

4.3 Possible Factors Affecting Student Performance

In considering the identification of students at risk it is useful to gain an understanding of the potential factors which may affect student performance. A survey of 95 Computer Science, Physics and Mathematics first year undergraduate students in the Faculty of Science and Engineering at Macquarie University, Sydney, Australia provides interesting results (Atif et al., 2015). Their results are presented in Table 4.3 with the addition of a column averaging the factor across the three subject areas and presented in highest percentage order first.

Table 4.3: Possible Factors Affecting Student Performance (Atif et al., 2015, p7)

Factors	**Computing**	**Physics**	**Mathematics**	**Average**
Emotional health	37%	47%	60%	48%
Family responsibility/commitments	16%	29%	40%	28%
Financial issues	12%	18%	33%	21%
Problems with daily travel	12%	29%	20%	20%
Felt under-prepared for this unit	9%	12%	20%	14%
Physical health	14%	6%	20%	13%
Paid work commitments	16%	0%	20%	12%
Social coping skills/social life style	5%	12%	13%	10%
Lack of student academic support	5%	6%	13%	8%
Other	9%	0%	13%	7%
Communication skills	9%	0%	7%	5%
Issue with the convener/lecturer/tutor	0%	0%	7%	2%
Religious commitments/activities	0%	0%	0%	0%

It should be noted that the factors were specified by the researcher and students were asked to respond by Likert scale (see Section 2.4.2, General definition of data types, ordinal data). Consequently, respondents were required to match their challenges into the available categories. However, OU and MOOC research shown in Tables 4.4 and 4.5 below corroborates the selection of these categories.

By far the highest ranking factor, perhaps unsurprisingly given the heightened awareness of such issues in recent years, is emotional health, averaging 48%. Family responsibilities/commitments, financial issues and travel problems are the next three highest factors at 28%, 21% and 20% respectively. Interestingly, lack of student academic support ranks quite low in these results, which may be a positive reflection on the quality of academic provision and support at Macquarie University.

Given the shift of student study patterns to less lecture/tutorial attendance and increasing usage of VLEs including systematic availability of all lecture and supporting learning material, research into fully distance learning, for example the OU and MOOCS, is directly relevant. Both the OU and MOOC providers are faced with very high drop-out rates. The drop-out rate of OU first year degree undergraduates in 2015/16 was 45% (HESA, 2019c). MOOC drop-out rates measured in 2014 (Reich, 2014) were 78% (taking student intent on registration into account, 90% excluding intent).

In the case of the OU, research conducted on student cohorts into the factors affecting student performance (Castles, 2004) corroborates Atif's 2015 survey, albeit with a different categorisation of factors (Table 4.4).

Table 4.4: Factors Affecting OU Student Performance (Castles, 2004, p3; Atif et al., 2015, p7)

Category	**Factors**
Social and environmental	Time and space available for study
	Appropriate patterns of work
	Ability to take part in tutorials or other institutional offerings
	Support of significant others
	Accommodation of social activities and friendship
Traumatic	Illness
	Bereavement
	Unemployment
	Lack of support from partners
	Caring for children or the elderly
	Level of adaptation to the everyday stresses of living
I**ntrinsic** (Attitudes, motivation, qualities)	Persistence
	Hardiness
	Coping ability
	Approaches to study
	Methods of study

In the case of MOOCs, a great deal of research has been carried out into the reasons for the very high student drop-out rates. Dalipi et al. (2018) provides further corroboration of Atif's 2015 survey questions selection (Table 4.5).

Table 4.5: Factors Affecting MOOC Student Drop-out (Dalipi et al., 2018, p2)

Category	**Factor**
Student related	Lack of motivation
	Lack of time
	Insufficient background knowledge and skills
MOOC related	Course design
	Isolation and lack of interactivity
	Hidden costs

None of the factors addressed in this section are to do with the student's intellectual capability to complete the course of study. For the purposes of this research we have assumed that institutional study pre-requisites and admission procedures are achieving their goals of making offers to students who are capable of the chosen study. The only exception to this in the factors described above is insufficient background knowledge and skills, where the student's background knowledge is misunderstood. A common example is where an apparently non-technical course of study requires some mathematical skills, for example, a degree in Psychology is likely to require statistical work.

In general, academic factors reflect the student's application to their studies and methods of study such as approaches to study and study techniques, included in Tables 4.3, 4.4 and 4.5 above. These can lead to well documented issues of shallow vs deep learning (Dolmans et al., 2016) and knowledge gaps (Reyes, 2015). Shallow (sometimes referred to as surface) learning describes learning by rote, compared with deep learning which is learning with understanding of the topic. Knowledge gaps are self-explanatory. In these cases, good module design, interim assessments and academic contact with students at tutorials are the traditional methods of identifying these issues and making interventions.

4.4 Identification of Students at Risk

Prior to the advent of large scale computing support to academic staff, the identification of students at risk of withdrawal or failure could only be made by the academic staff (lecturer or tutor) themselves, perhaps with the analysis at departmental level of aggregated data by support teams. By its very nature, such identification depended on the experience, ability and motivation of academic staff. The advent of modern computer based methods to support this process have increasingly turned towards the application of intelligent techniques i.e. AI/ML, to recognise early signs of students who may be at risk. As

described in Chapters Three (Datasets used in this research and relevant student attributes), Four (Relevant AI and ML Techniques) and Five (Intelligent Learning/Training Systems), these techniques focus upon measurable student attributes such as student demographics, previous academic results, interim assessments, VLE access and attendance. While there is some research into the use of AI/ML techniques for the measurement of more esoteric student attributes such as motivation, ambition and level of anxiety, these are yet to be established as practically applicable analytic methods.

In order to consider approaches to intelligent support of institutional interventions (see Chapter Five), I have consolidated Tables 4.3 (Possible factors affecting student performance), 4.4 (Factors affecting OU student performance) and 6.3 (Factors affecting MOOC student drop-out) into Table 4.6 and for each factor I have identified it as usefully (for subsequent intervention purposes) categorised by intelligent techniques. In the majority of cases this categorisation is straightforward. Where judgement was necessary, I have either cited appropriate evidence or made it clear that in the particular case this was my own judgement, based upon conflicting or no clear evidence.

Table 4.6: Potential Factors Affecting Student Performance and Methods of Recognition

Factors	**Identifiable by**
Emotional health	Questionnaire or academic staff
Family responsibility/commitments	Questionnaire or academic staff
Financial issues	Questionnaire or academic staff
Problems with daily travel	Questionnaire or academic staff
Felt under-prepared for this unit	Questionnaire or academic staff
Physical health	Questionnaire or academic staff
Paid work commitments	Questionnaire or academic staff
Social coping skills/social life style	Questionnaire or academic staff
Lack of student academic support	Questionnaire or academic staff
Communication skills	Questionnaire or academic staff
Issue with the convener/lecturer/tutor	Questionnaire or academic staff
Religious commitments/activities	Questionnaire or academic staff

Factors	Identifiable by
Time and space available for study	Questionnaire or academic staff
Appropriate patterns of work	Questionnaire or academic staff [2]
Ability to take part in tutorials or other institutional offerings	AI/ML techniques
Accommodation of social activities and friendship	Questionnaire or academic staff
Bereavement	Questionnaire or academic staff
Unemployment	Questionnaire or academic staff
Lack of support from partners	Questionnaire or academic staff
Level of adaptation to the everyday stresses of living	Questionnaire or academic staff
Persistence	Questionnaire or academic staff [1]
Hardiness	Questionnaire or academic staff [1]
Lack of motivation	Questionnaire or academic staff [1]
Lack of time	Questionnaire or academic staff [2]
Insufficient background knowledge and skills	AI/ML techniques
Course design	AI/ML techniques
Isolation and lack of interactivity	AI/ML techniques

[1] Esoteric student attributes, yet to be established as identifiable by practically applicable analytic methods.

[2] Include some elements which may be identified by AI/ML techniques.

It is important to note that only 4 of the 27 identified potential factors affecting student performance, and hence with the potential to identify students at risk, is currently detectable by AI/ML analytical techniques. The assessment of a further 3 (Persistence, hardiness and motivation) identified by Castles (2004) may be the subject of AI/ML research in the future. In addition, there are 2 factors, appropriate patterns of work and lack of time, where AI/ML techniques are capable of providing useful information to learning analytics systems and academic staff. For example, VLE usage patterns may provide useful

information on students' appropriate patterns of work and on-line assessments on student's lack of time (from a time management perspective).

A very important consideration in the collection and use of this data are the legal, ethical and moral issues. These are discussed below in section 5.4.

4.5 Chapter Summary

In this chapter I have described the significant impacts upon students of a failure to progress and the very significant financial and reputational impacts upon institutions. The financial impact on the institutional budget of the withdrawal of a first year UK/EU undergraduate student can be as high as £27,750 across an anticipated 3 year study period. For international students this impact rises to £33,750. Using HESA statistics for the 2015/16 academic year, 6.4% of students withdrew during the first year of their studies, the consequential financial impacts on a typical institution and UK universities as a whole may be calculated as £7.2M - £10.8M and £625M to £938M respectively. I have presented a comprehensive list of the factors which potentially affect student performance, including how they may be identified. Only 4 of the 27 potential factors identified are detectable by current AI/ML techniques. It is also the case that almost none are concern the student's intellectual capability to complete the course of study. In the following chapter I present consequential non-computer facilitated and computer facilitated methods of student interventions, discussing their usefulness in achieving positive learning outcomes and research into how students prefer to receive interventions.

CHAPTER FIVE

Approaches to Intelligent Support of Institutional Interventions

5.1 Introduction

5.1.1 Contributions to Knowledge Relevant to this Chapter

In this chapter I describe the results of my research exploring student interventions. These activities support my contribution of demonstrating how the analysis of these limited attributes: attendance, VLE accesses and intermediate assessments, may provide potentially useful intervention guidance to academic leadership.

The following sections of this chapter are supported by previously published material:

5.2.1 Targeted individual student intervention - Individual student intervention methods (Wakelam et al., 2019)

5.4 Legal, ethical and moral considerations (Wakelam et al., 2019)

5.1.2 Summary of Chapter Content

I survey and review student intervention methods, considering each of traditional, non-computer facilitated and computer facilitated approaches. I consider which are applicable to student progress monitoring through LA and how such interventions might be timely in resulting in positive learning outcomes. Drawing upon published research, I discuss how the method and timeliness of such interventions is critical to their success, and which methods are preferred by students and most likely to succeed. I then discuss student privacy and ethics issues.

5.2 Intervention Methods

In order to understand the context of student interventions in respect of LA it is important to recognise institutional objectives for their corresponding investment. A recent review of 389 Higher Education institutions (Parnell et al., 2018) in the US (Figure 5.1) showed that 96% cited the improvement of student outcomes as their primary goal from using student analytics, with the goal of improved delivery of programmes and services in second place at 71%. The goal of eliminating or reducing programmes as their third goal was cited by 39% (Parnell et al., 2018).

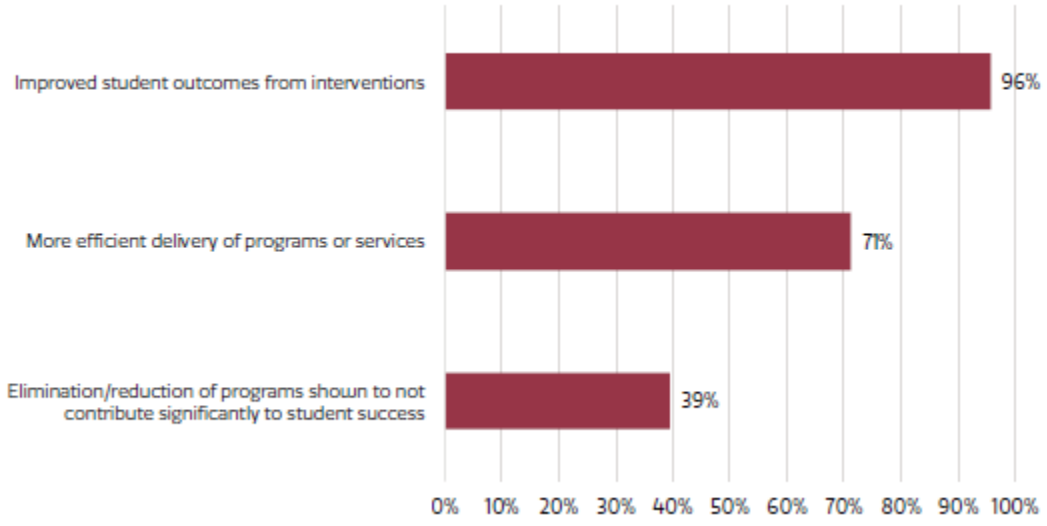


Figure 5.1: Institutions' Goals for Conducting Student Success Studies (Parnell et al., 2018, p5)

Two types of intervention resulting from the deployment of learning analytics may be considered, those targeted at an individual student during the progress on a module/course, and those which may be taken into account to re-design the module in question for the following occurrences or in the creation of future courses.

5.2.1 Targeted Individual Student Intervention

The first is the recognition of challenges to an individual student and a consequential opportunity for intervention. It is worth noting here that while the highest priority is likely to be given to students at risk of failure or dropping out, learning analytics may also provide tutors with the opportunity to provide interventional support to students whose performance is below expectations.

Before the introduction of computer facilitated methods a variety of intervention approaches were available to academic staff (Table 5.1).

Table 5.1: Non-Computer Facilitated Intervention Approaches

Method	**Contact type**
Regular progress meetings	Face to face
Personalised coaching	Face to face
Additional lectures or tutorials on selected topics for one or more students	Face to face
Pre-arranged drop-in sessions	Face to face
Appointment of a "peer" or an additional academic counsellor/mentor	Face to face
Provision of reinforcement learning material	Face to face (or via internal/external post)
Suspension of studies while a particular set of circumstances (perhaps illness or family issues) are resolved	Face to face
Consideration of an alternative, perhaps more appropriate, course.	Face to face

Clearly, all of the above are based upon face to face contact, telephone or paper communication. This corresponds with the requirement for students in this pre-computer facilitated era to attend lectures and tutorials in order to receive their teaching.

A more recent exception to this has been the establishment of the Open University in 1969. The vast majority of course material was delivered through the post directly to student's homes and intermediate summative assessments (cumulative value 50% of the overall mark) being submitted to the OU by the student by post. This material was supplemented by short television programmes and the opportunity to visit local and regional centres to meet with a tutor. The final examination (value 50% of the overall mark) was sat at a designated centre.

Under these circumstances opportunities to identify students at risk or to make interventions were very limited. This may have been a contributory factor to the high drop-out rates (see Section 4.2).

A wide variety of student intervention methods are now available to academic leadership with traditional pre-computing methods now supplemented by computer facilitated ones, including some which may be automatically (system) generated. A variety of methods are listed below (Choi et al., 2018) including the pros and cons of each (Table 5.2).

Table 5.2: Individual Student Intervention Methods (Choi et al., 2018; Rienties et al., 2016a)

Method	Pros	Cons
Email	Least expensive Allows automatically generated messages on attendance or concerning interim assessment result (seen as less confrontational by some students) Allows personalisation via mail merge May also be used for encouragement of students making post intervention progress	Students may easily overlook the message due to too many spam emails
Phone call	Good for emergency matters – two-way synchronous communications	Students may not be available and sometimes feel offended
Skype call	Provides face to face discussion Flexible on student/instructor location and timing	Often requires pre-arrangement.
Instant messaging	Preferred communication channel for many students Allows automatically generated messages on attendance or concerning interim assessment result (seen as less confrontational by some students)	More costly than email as it requires one-to-one communications
LMS post & news	Facilitates many-to-many asynchronous communications Allows automatically generated messages on attendance or concerning interim assessment result (seen as less confrontational by some students) Dashboards allow comparison with other student's progress	Requires students to login to the LMS and may overlook the posts and news
Group consultation	Effective communication Good for timid students	Usually needs making appointments in advance and expensive for instructors

Method	Pros	Cons
Face-to-face consultation	Effective communication One-to-one consultation	Most expensive and usually needs to make appointments in advance
Video recording	Effective instruction Not restricted by time	Substantial initial effort to record the instructions
Peer review	Encourages critical evaluation Students can learn from each other	Requires good question design Often conducted in class
Audio feedback on assessments	More informative than written feedback	Time expensive to instructors
E-tutorial	Supplementary instructions available 24/7 (e.g. MyMathLab and MyStatLab developed by Pearson Publishing) Suitable for highly motivated students	May incur a price for students or instructors
Organise catch-up tutorials on specific topics that student(s) are struggling with	Can be organised as face to face or videoconference/skype and include multiple students Ability to invite groups identified by similar LA metrics	Identified student may not attend
Podcasts of specific learning activities in the module	Supplements course material Focused upon specific selected topic	Time expensive to instructors
Schedule drop-in sessions	Face to face coaching on student identified topics Voluntary but targeting identified students Ability to invite groups identified by similar LA metrics	Identified student may not attend
Boot camps	Supplements course material Focused upon specific selected topics Face to face contact	Costly in time for instructors and students. May be difficult to schedule

Learning analytics can be deployed to automatically initiate first step interventions, either via email, SMS or VLE communications (notifications sent to appropriate academic staff in parallel). System generated

interventions may be seen by some students as less confrontational/stressful than personal academic staff contact, even by Email (see Section 5.3 Students' intervention preferences). Examples of potentially automatically generated intervention triggers include less than benchmark attendance at lectures/tutorials, notifications of below expectation interim assessments results (formative or summative) or a below defined thresholds of VLE accesses. In each case, the communication may include a summary of the rest of the cohort's performance to give the student some context. More advanced automatic communications may provide links to recommended additional subject material based upon knowledge gaps identified through interim assessment results. The techniques developed by modern adaptive learning systems (see Section 2.5.2.1) provide future potential for more sophisticated system generated personalised support to students. These include adapting the learning path to be more suited to an individual student's progress and the exploitation of continued progress in recommender systems (Hoic-Bozic et al., 2015). In general, such automatically generated messages are tailorable to deliver escalating messages over time depending on student progress, with an identified point at which the direct action of an appropriate member of academic staff is triggered.

However accurate and valuable the learning analytics data is developed, the methods of presentation of data to both students and academic staff are critical to their effectiveness. The presentation of this material, in a way in which information and trends are clearly understood, must in turn aim to encourage or provoke appropriate action. This is a major topic in its own right and is out of scope of this dissertation, however, as noted in Chapter Two, Literature Review, research into the continually developing field of dashboards is worthy of mention. A recent review of learning dashboard research (Schwendimann, et al., 2016) observes that despite substantial research on information visualisation, research on the resulting value of learning dashboards is still in its early stages.

5.2.2 Systematic Interventions to the Module

The second type of intervention is where through analysis of the challenges academic leadership identifies issues which require a wider view to be taken of the module/course as a whole, so called systematic issues. This may result in an intervention directed at the cohort as a whole during the module and/or lead to a review and potential re-design of the module in time for future execution. My research focus is upon individual and timely student interventions during a student's course of study, however given their importance to future executions of modules, a variety of resulting module/course opportunities are shown in Table 5.3 (Rienties, 2016b, p6).

Table 5.3: Module/Course Design and Execution Interventions

Action type
Re-design learning material
Redesign assignments
Introduce graded discussion forum activities
Group-based wiki assignment
Assign groups based upon learning analytics metrics
Introduce bi-weekly online videoconference sessions
Podcasts of key learning elements in the module
Screencasts of "how to survive the first two weeks"
Emotional questionnaire to gauge students emotions
Introduce buddy system

Any of the actions in Table 5.3 may be implemented during the module itself where academic staff considers it necessary, as well as incorporating the changes in a redesign of the module for future occurrences. Issues which may be relevant to other modules may be identified at departmental/school level for consideration on appropriate other modules.

It is also the case that academic staff have the opportunity to pursue a very thorough review and implementation of multi-student/course intervention strategies, including measuring the results of interventions with previous or future deliveries of the module/course, piloting (proof of concept) of implementation changes or random trials (Rienties et al., 2016a). These are discussed in Chapter Two, Literature Review.

5.3 Students' Intervention Preferences

Key to the success of student interventions in supporting students' at risk is an understanding of how these interventions are made, in particular whether they result in the desired positive effect on recipients. Simply put, an intervention method or message which is perceived as threatening instead of supportive may have a negative effect on the student. Research on student psychology could prove useful, however, an appropriate approach consistent with the principle of consent described later in this chapter (see Section 5.4) is to provide students with a selection of options aligned with a description of their benefits.

The success or otherwise of intervention methods will always be dependent upon the reaction of the student and their willingness to act upon the intervention. Clearly, a one size fits all approach is unlikely to work for all students and therefore intervention design must include early engagement, before course commencement, with students to establish their preferences. Students should be given the opportunity to change their preferences as they progress through their studies.

An interesting example of student preferences and attitudes to the use of alerts on their progress is shown in a survey of undergraduate students at Macquarie University, Sydney, Australia (Atif et. al, 2015). A total of 639 students responded to the survey, of which 62% were first years. The results provide useful data on areas to focus upon when considering interventions design. They include student preferences of the timing of intervention contacts (Figure 5.2), the specific behaviours they would like to be contacted about (Figure 5.3) and how they would like to receive intervention messages (Figure 5.4).

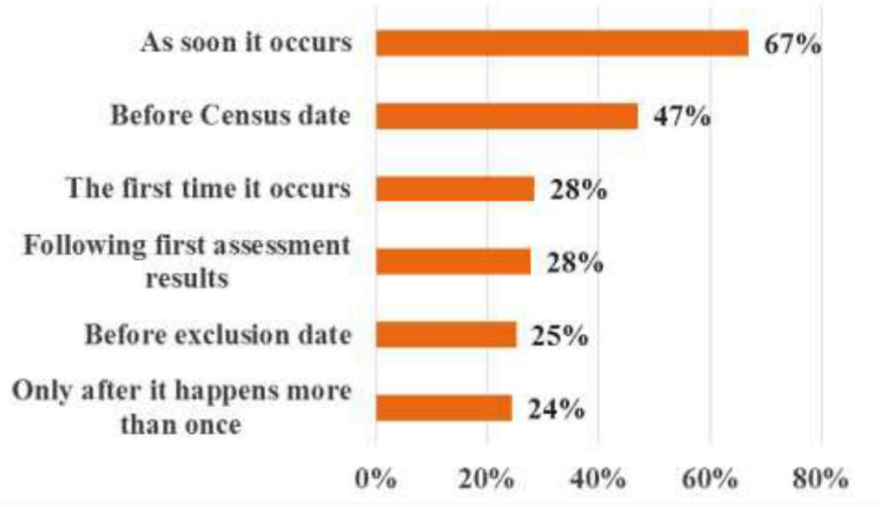


Figure 5.2: When Students' Like to be Contacted (Atif et. al, 2015 , p38)

An interesting observation of these results is that although two thirds of the students preferred intervention contact to happen immediately the recognition of an issue or the event occurred, a significant proportion of students opted for later contact. For example, almost a quarter of students wanted a "second strike" approach (only after it happens more than once) and a quarter appear to be content with being alerted to issues closer to the point at which exclusion was imminent (before exclusion date).

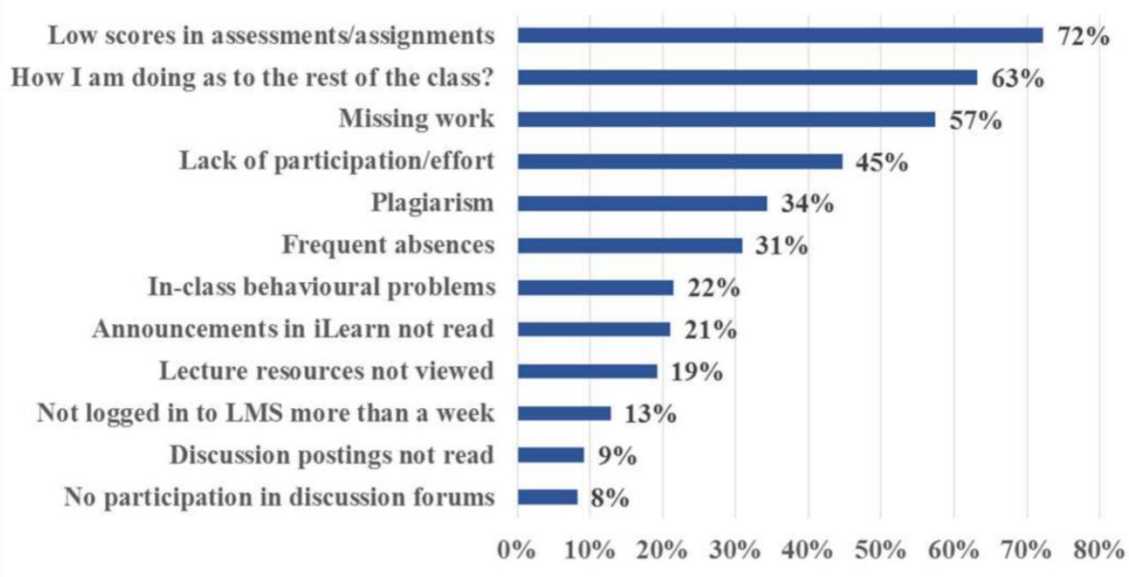


Figure 5.3: For What Specific Behaviours Students' Like to be Contacted (Atif et. al, 2015, p39)

Not unexpectedly, the almost three quarters of the student's surveyed prefer notifications on low scores in assessments/assignments. Perhaps surprising is that only just over half of the students wished to be notified of missing work. The high proportion of students (63%) interested in awareness of how they were performing in comparison with other members of their class is supported by research into student dashboards (Schwendimann et al., 2016).

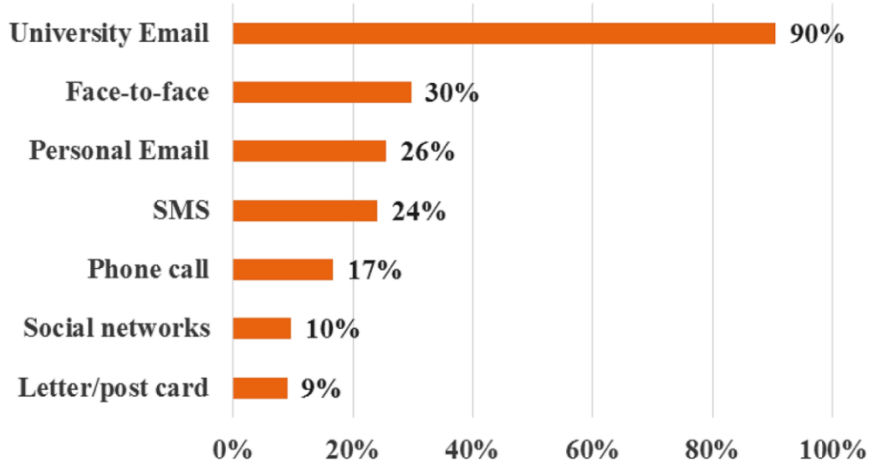


Figure 5.4: How Students would Like to Receive Intervention Messages (Atif et. al, 2015, p40)

A very strong student preference for intervention contact to be made via university Email (90%) would appear to suggest that students find indirect contact as opposed to an unexpected phone call (17%) either

less stressful or more comfortable. Given this, the relatively high proportion of students opting for face to face contact (30%) may be seen as surprising, however, such contact is more likely to be beneficial in providing positive intervention guidance.

Another good example of the variety of student preferences they may opt for is shown in Figure 5.5 (Atif et al., 2015).

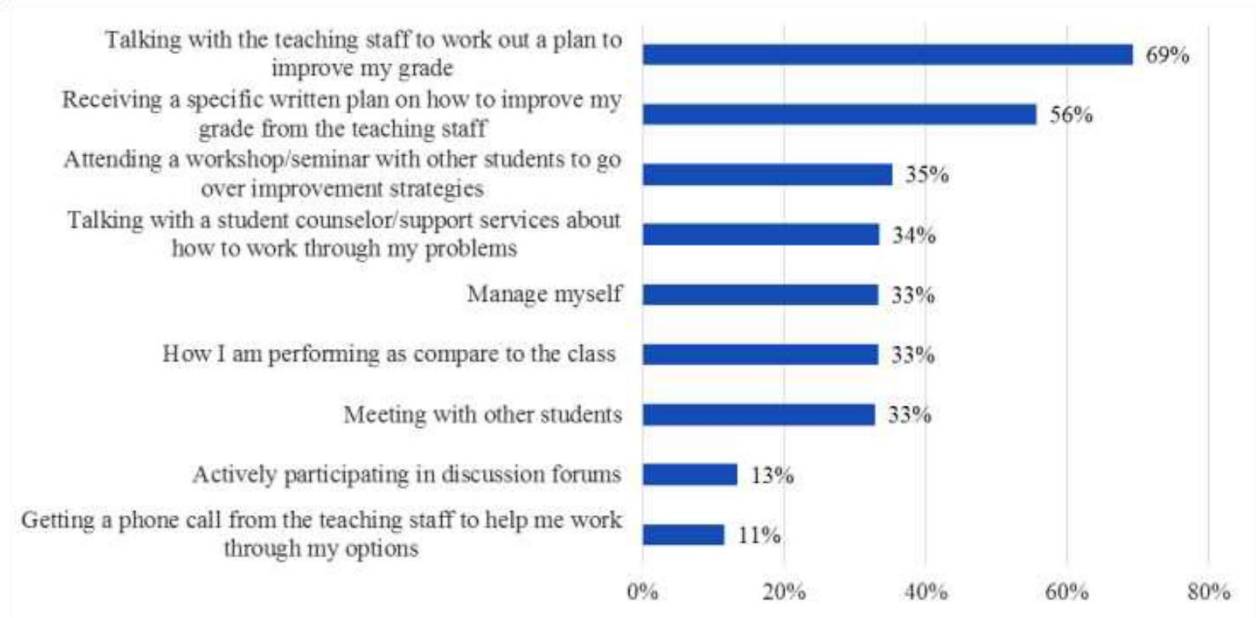


Figure 5.5: Student Preferences for Motivational Intervention Actions (Atif et al., 2015, p41)

5.4 Legal, Ethical and Moral Considerations

Legal, ethical and moral considerations in the deployment of learning analytics and interventions are key challenges to institutions. They include informed consent, transparency to students, the right to challenge the accuracy of data and resulting analyses and prior consent to intervention processes and their execution (Slade & Tait, 2019). It is also the case that universities are confused as to whether in providing this data students are in fact giving prior (and legally supportable) approval for their inclusion in learning analytics, and furthermore whether this entitles the institutions to categorise students and to be the catalyst/basis for interventions (Sclater & Bailey, 2018). These challenges are well documented in a number of research papers including Pardo and Siemens, (2014); DeFreitas et al., (2015). In addition, a comprehensive literature review of 86 publications was commissioned by Jisc (formerly the Joint Information Systems Committee), who provide UK universities and colleges with shared digital infrastructure and services including learning analytics, to discuss the challenges faced by institutions and provide the background for a future code of practice (Sclater & Bailey, 2018). A discussion on ethical and data privacy issues in

learning analytics based on three studies in Higher Education and primary school contexts (Rodríguez-Triana et al., 2016), specifically focusses on tutor-led approaches. Legislation has been in place for over two decades, specifically the European Data Protection Directive 1995 and the UK Data Protection Act 1998. More recently, General Data Protection Regulation (UK Government, 2018) sets out the legal data protection principles which institutions and organisations are responsible for adhering to. In addition, despite their algorithmic accuracy intentions, there is growing research into the potential for machine learning approaches to introduce bias, such as class, gender and ethnicity (Wilson et al., 2017).

5.5 Chapter Summary

In this chapter I have presented a comprehensive description, including their advantages and disadvantages, of non-computer and computer facilitated student intervention methods aimed at improving student success. I identify opportunities for student-personalised automatically system-generated intervention messages based upon learning analytics techniques described in the previous chapter. I have detailed three types of interventions, those personalised to individual students during a module, those which may require academic staff to deploy an intervention to a group of students and those which learning analytics systems prompt academic staff to consider re-design of the module for future occurrences. Key to their success in a positive learning outcome, I have presented and discussed research into students' preferences for the timing, reasons and methods of interventions. Interestingly, although two thirds of students preferred intervention contact to be made immediately an issue was identified, a significant portion opted for later contact. In terms of the reasons for intervention contact, the highest percentages of intervention preferences were direct achievement related, including low assessment scores (72%) and missing work (57%), with an interesting 63% of students wanting to know their relative performance to the rest of the class. A very high percentage (90%) of students surveyed have a very strong preference for initial intervention contact to be made via email, compared with an unexpected phone call (17%). I discuss the legal, ethical and moral considerations key to the deployment of learning analytics based interventions. In the following chapter I describe the datasets used in my research and I catalogue a variety of student attributes relevant to learning analytics and student performance prediction.

CHAPTER SIX

Datasets Used in this Research and Relevant Student Attributes

6.1 Introduction

In this chapter I describe each the four datasets used in my research and experimentation and I discuss the variety of student attributes encountered.

I catalogue the wide variety of student attributes I have encountered during my research and experiments. I then propose a list of static and dynamic student attributes, of potential use in student performance prediction. I present these, in each case considering how students and institutions may view their respective sensitivity to student privacy and therefore consequent restriction of their use in a learning analytics context.

The following sections of this chapter are supported by previously published material:

Section 6.2.1 General definition of data types (Wakelam et.al., 2016)

Section 6.2.3 Portuguese secondary school student achievement (Wakelam et.al., 2016)

Section 6.2.5 The University of Hertfordshire Strategic IT Management module (Wakelam et al., 2020)

Section 6.3.1 Potentially useful student attributes (Wakelam et al., 2016), (Wakelam et al., 2020)

6.2 Datasets used in this Research

6.2.1 Small Student Dataset for Higher Education Teachers

Data mining techniques focus upon delivering satisfactory analyses when dealing with large datasets (Andonie, 2010). However, academics are often faced with comparatively small numbers of students and therefore only small datasets. The work of Natek & Zwilling (2014) investigates the application of data mining techniques to a small dataset.

This dataset comprises 10 students and 11 mixed numeric and categoric attributes. The selection of two key numeric attributes (Activities Points and Exam Points) provided an opportunity to investigate the analysis of a very small dataset. In this case the Support Vector Machine (SVM) classification technique was used.

This dataset was extracted from research conducted on the 2010/11 Slovenia International School for Social and Business Studies degree program student cohort, studying Informatics – Economy in Contemporary Society (Natek & Zwilling, 2014). The overall cohort size was 42, with attributes as follows (Table 6.1):

Table 6.1: Economy in Contemporary Society Student Attributes (Natek & Zwilling, 2014, p2)

Attribute	**Values**	**Data type**
Study year	2010/11	Nominal
Student number	1 - 42	Ordinal
Gender	Female/male	Nominal
Student year of birth	1988	Numeric
Employment	No/yes	Nominal
Status (e.g. Sport etc.)	No/yes	Nominal
Registration	First/repeat	Nominal
Type of study	Full time/part time	Nominal
Exam condition	No/yes	Nominal
Activities points	0 - 50	Numeric
Exam points	0 - 50	Numeric
Final points	0 - 50	Numeric
Final grade	1 - 10	Numeric

The following extract of 10 students was chosen as the dataset to be used as the basis for analysis (Table 6.2):

Table 6.2: Small Student Dataset for Higher Education Teachers (Natek &Swilling, 2014, p2)

Study year	Student	Gender	Year of Birth	Employ -ment	Status (sport…)	Registr -ation	Type of Study	Exam Cond -ition	Activitie s Points (50)	Exam Points (50)	Final Points (100)	Final Grade (10)
2010-11	1	Female	1988	No	No	First	Full time	Yes	46	46	92	10
2010-11	2	Male	1990	No	No	First	Full time	Yes	38	33	71	7
2010-11	3	Female	1990	No	No	First	Part time	Yes	39	30	69	7
2010-11	4	Female	1990	No	No	First	Full time	Yes	47	35	82	8
2010-11	5	Female	1989	No	No	First	Full time	Yes	39	36	75	7
2010-11	6	Male	1990	No	No	First	Full time	Yes	38	30	68	7
2010-11	7	Female	1990	No	No	First	Full time	Yes	39	36	75	7
2010-11	8	Female	1990	No	Yes	First	Full time	Yes	39	33	72	7
2010-11	9	Male	1990	No	No	First	Full time	Yes	39	38	77	8
2012-13	10	Female	1990	No	No	First	Full time	Yes	44	30	74	

For the purposes of experimenting with a very small dataset, only Activities points and Exam points attributes were used.

6.2.2 Students' Knowledge Levels on DC Electrical Machines

This dataset was obtained from the research conducted into the creation of an efficient user knowledge model for adaptive learning systems (Kahraman et al., 2013), freely available from the University College Irvine (UCI) Machine Learning Repository. This dataset comprises 258 students in an on-line web based Electrical Engineering course (the full dataset is included in Appendix D) each with 5 numerical

attributes, providing the opportunity to investigate a modest sized dataset with minimal student attributes. In this case the application of Principal Component Analysis (PCA) was used. Data was measured against 5 attributes (Table 6.3):

Table 6.3: DC Electrical Machines Student Dataset (Kahraman et al., 2013)

Attribute	**Values**	**Data Type**
STG (The degree of study time for goal object materials),	0 – 1 (normalised)	Numeric
SCG (The degree of repetition number of user for goal object materials)	0 – 1 (normalised)	Numeric
STR (The degree of study time of user for related objects with goal object)	0 – 1 (normalised)	Numeric
LPR (The exam performance of user for related objects with goal object)	0 – 1 (normalised)	Numeric
PEG (The exam performance of user for goal objects)	0 – 1 (normalised)	Numeric
UNS (The knowledge level of user)	High/Middle/Low/Very Low	Ordinal

6.2.3 Portuguese Secondary School Student Achievement

The Portuguese student dataset is open source published data (Cortez & Silva, 2008). The data was taken from a set of students from a Portuguese study. It consists of information taken from two Portuguese secondary schools and each student has a surprising variety of 33 attributes. The data includes three labels: first period grade, second period grade and final grade. The subjects are Mathematics (395 students) and Portuguese Language (649 students) and the data was collected during the 2005-2006 academic year. The attributes comprise 16 numeric (including the labels: first period, second period and final performance grades) and 17 nominal (Table 6.4). This dataset provided the opportunity to investigate a large dataset, with a very large number of student attributes. In this case, for the 16 numeric attributes PCA was used to reduce the dimensionality of the data followed by Growing Neural Gas (GNG) to identify potentially useful clusters of data. For the 17 nominal attributes a novel technique was used, followed by PCA cluster analysis.

Table 6.4: Portuguese Student Dataset (Cortez & Silva, 2008, p3)

No.	Attribute Name	Description	Data Type	Values
1	School	Student's school	Nominal	"GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
2	Gender	Student's Gender	Nominal	"F" - female or "M" - male
3	Age	Student's age	Numeric	15 to 22
4	Address	Student's home address	Nominal	"U" - urban or "R" - rural
5	Famsize	Family size	Nominal	"LE3" - less or equal to 3 or "GT3" - greater than 3)
6	Pstatus	Parent's cohabitation status	Nominal	"T" - living together or "A" - apart)
7	Medu	Mother's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – Higher Education)
8	Fedu	Father's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – Higher Education)
9	Mjob	Mother's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
10	Fjob	Father's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
11	Reason	Reason to choose this school	Nominal	close to "home", school "reputation", "course" preference or "other"
12	Guardian	Student's guardian	Nominal	"mother", "father" or "other")
13	Traveltime	Home to school travel time	Numeric	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
14	Studytime	Weekly study time	Numeric	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours

No.	Attribute Name	Description	Data Type	Values
15	Failures	Number of past class failures	Numeric	n if 1<=n<3, else 4)
16	Schoolsup	Extra educational support	Nominal	Yes or no
17	Famsup	Family educational support	Nominal	Yes or no
18	Paid	Extra paid classes within the course subject (Math or Portuguese)	Nominal	Yes or no
19	Activities	Extra-curricular activities	Nominal	Yes or no
20	Nursery	Attended nursery school	Nominal	Yes or no
21	Higher	Wants to take Higher Education	Nominal	Yes or no
22	Internet	Internet access at home	Nominal	Yes or no
23	Romantic	With a romantic relationship	Nominal	Yes or no
24	Famrel	Quality of family relationships	Numeric	From 1 - very bad to 5 - excellent
25	Freetime	Free time after school	Numeric	From 1 - very low to 5 - very high
26	Goout	Going out with friends	Numeric	From 1 - very low to 5 - very high
27	Dalc	Workday alcohol consumption	Numeric	From 1 - very low to 5 - very high)
28	Walc	Weekend alcohol consumption	Numeric	From 1 - very low to 5 - very high
29	Health	Current health status	Numeric	From 1 - very bad to 5 - very good
30	Absences	No. of school absences	Numeric	From 0 to 93
31	G1	First period grade	Numeric	From 0 to 20
32	G2	Second period grade	Numeric	From 0 to 20)
33	G3	Final grade	Numeric	From 0 to 20

6.2.4 Open University

The Open University Learning Analytics Dataset (OULAD) is open source published data (Kuzilek et al., 2017; Open University, 2017). The dataset contains information about 22 courses from 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries). In total, there are 28 mixed numeric and nominal attributes per student. The dataset contains demographic data together with aggregated clickstream data of students' interactions with the OU Virtual VLE, as shown in the schema (Figure 6.1). As a subset of the 2013/14 academic year data, it provides a detailed insight into the data which supports the OU's institutional analysis of student progress which is systematically provided to academics via dashboards. This dataset is noteworthy as an extract from a successful, live operational learning analytics system delivering value to the institution, which has overcome a number of the common ethical and privacy barriers to the collection and exploitation of student data. The OU applies Bayesian classifier, Classification and Regression Tree (CART) and K-Nearest Neighbour (KNN) techniques. Given the substantial development and analysis carried out by the OU to develop a working system to support academics, it was not helpful to perform my own analyses of this data using alternative techniques. However, examination and reflection upon the data and analyses techniques used by the OU provides valuable insights into the benefits of very large datasets, driving my focus on what may be achieved with small datasets.

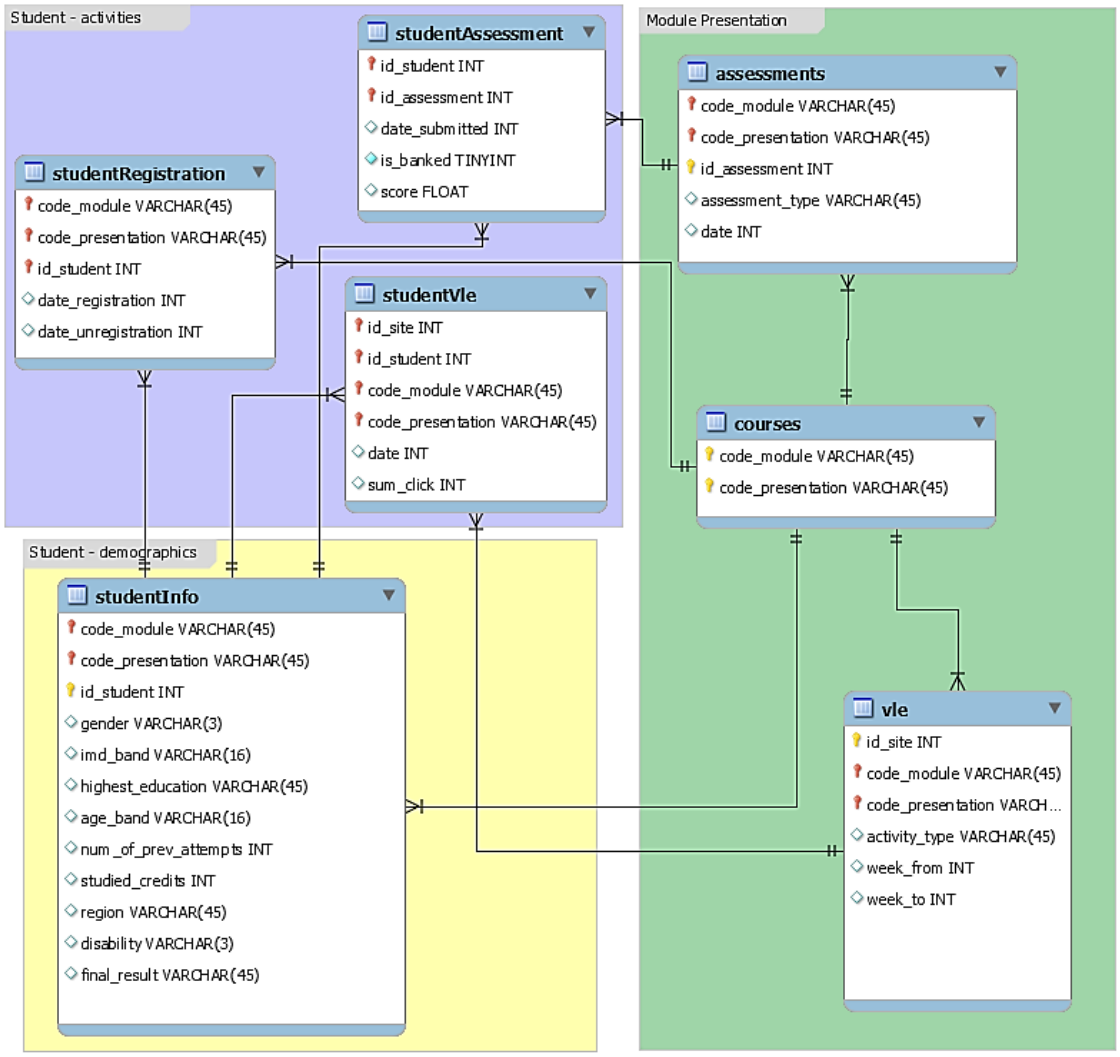


Figure 6.1: OULAD Schema (Kuzilek et al., 2017)

The following description of the attributes for each of the data files shown in Figure 3.1 above has been extracted from (Kuzilek et al., 2017). Table 6.5 contains the list of all modules and their presentations.

Table 6.5: Courses.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Code_module	Code name of the module, which serves as the identifier.
Code_presentation	Code name of the presentation, consisting of the year and "B" for the presentation starting in February and "J" for the presentation starting in October.
Length	Length of the module-presentation in days.

Table 6.6 contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam.

Table 6.6: Assessments.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Code_module	Identification code of the module, to which the assessment belongs.
Code_presentation	Identification code of the presentation, to which the assessment belongs.
Id_assessment	Identification number of the assessment.
Assessment_type	Type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
Date	Information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
Weight	Weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

Table 6.7 contains information about the available materials in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded.

Table 6.7: Vle.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Id_site	Identification number of the material.
Code_module	Identification code for module.
Code_presentation	Identification code of presentation.
Activity_type	The role associated with the module material.
Week_from	The week from which the material is planned to be used.
Week_to	Week until which the material is planned to be used.

Table 6.8 contains demographic information about the students together with their results.

Table 6.8: StudentInfo.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Code_module	An identification code for a module on which the student is registered.
Code_presentation	The identification code of the presentation during which the student is registered on the module.
Id_student	A unique identification number for the student.
Gender	The student's gender.
Region	Identifies the geographic region, where the student lived while taking the module-presentation.
Highest_education	Highest student education level on entry to the module presentation
Imd_band	Specifies the Index of Multiple Deprivation band of the place where the student lived during the module-presentation (UK Government, 2015).
Age_band	Band of the student's age.
Num_of_prev_attempts	The number times the student has attempted this module.
Studied_credits	Total number of credits for the modules the student is currently studying.
Disability	Indicates whether the student has declared a disability.
Final_result	Student's final result in the module-presentation.

Table 6.9 contains information about the time when the student registered for the module presentation. For students who have been unregistered, the date of their unregistration is also recorded.

Table 6.9: StudentRegistration.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Code_module	An identification code for a module.
Code_presentation	The identification code of the presentation.
Id_student	A unique identification number for the student.
Date_registration	The date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
Date_unregistration	Date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course, have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

Table 6.10 contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system.

Table 6.10: StudentAssessment.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Id_assessment	The identification number of the assessment.
Id_student	A unique identification number for the student.
Date_submitted	The date of student submission, measured as the number of days since the start of the module presentation.
Is_banked	A status flag indicating that the assessment result has been transferred from a previous presentation.
Score	The student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

Table 6.11 contains information about each student's interactions with the materials in the VLE.

Table 6.11: StudentVle.csv (Kuzilek et al., 2017)

Attribute identifier	Description
Code_module	An identification code for a module
Code_presentation	The identification code of the module presentation.
Id_student	A unique identification number for the student.
Id_site	An identification number for the VLE material.
Date	The date of student's interaction with the material measured as the number of days since the start of the module-presentation.
Sum_click	The number of times a student interacts with the material in that day.

6.2.5 The University of Hertfordshire, Strategic IT Management module

The is a Level 6 (Final Year undergraduate) Computer Science module, duration 15 weeks (including a 3 week vacation period and 2 weeks allocated for submission and review of each of the two final assessments) comprising 23 students, 5 intermediate summative assessments and no final examination.

A detailed description of this dataset is included in Chapter Eight, Experiment to establish the potential for student performance prediction in small cohorts with minimal available attributes using learning analytics techniques.

In this case Decision Tree, Random Forest and K-Nearest Neighbour techniques were used and their predictive results compared.

6.3 Relevant student attributes

6.3.1 Potentially useful student attributes

The following list has been compiled from details of the student datasets described in Chapter Three supplemented by presentation feedback from colleague researchers and staff. It is hoped that future networking with organisational/institutional stakeholders, such as UH, JISC and the OU, may refine this list, which is intended to be a tailorable starting point for institutions considering the deployment of learning analytics systems.

I have organised these potentially useful student attributes under the following categories:

Fixed Static (Table 6.12) – these are attributes that can be collected in advance of the first active learning session itself, either by extraction from the student information system or via a questionnaire approach as a course pre-requisite. They include attributes which are unlikely to change during the course of the learning period. Examples are gender, address, internet access.

Evolving static (Table 6.13) – these are attributes that are collected/updated at the start of each module, via a Q&A approach. They represent information that the learning system cannot generate automatically, instead requiring student input. Examples are independent study time, level of non-course work load, student self-assessment of progress.

Dynamic (Table 6.14) – these are attributes that are determined in real-time by the learning system itself, designed to evaluate progress and provide live information to the adaptive engine. Examples are performance in quizzes, speed of response to questions, number of repeats of learning components.

In each case, I have identified the source and the typical method of data collection, which can be expected to vary by institution. It should be noted that where questionnaire is identified as the data collection method this may be an on-line student activity at the institutional/course/module joining point.

I have also assigned a subjective "sensitivity" indicator based upon my own experience and judgement of student and institutional considerations of privacy and ethical behaviour: Likely to be readily available for analysis (Green); potentially sensitive (Amber); sensitive (Red). In general, I have defined the majority of student personal (not directly related to study) data as sensitive.

Table 6.12: Fixed Static

Attribute	Source	Method of collection	Type[4]	Sensitivity
Gender	Portuguese student dataset	Student information system	Nominal	
Age	Portuguese student dataset	Student information system	Numeric	
Address	Portuguese student dataset	Student information system	Nominal	
Travel time (zero if on-line course)	Portuguese student dataset	Questionnaire	Numeric	
Employment during course	Small Student Dataset for HE Teachers	Questionnaire	Nominal	

Attribute	Source	Method of collection	Type[4]	Sensitivity
Type of study (full/part time)	Small Student Dataset for HE Teachers	Student information system	Nominal	
Ethnicity	Presentation feedback	Student information system	Nominal	
Family size	Portuguese student dataset	Questionnaire	Numeric	
Parent's status	Portuguese student dataset	Questionnaire	Nominal	
Mother's job	Portuguese student dataset	Questionnaire	Nominal	
Father's job	Portuguese student dataset	Questionnaire	Nominal	
Deprivation index[3]	OU	System generated	Numeric	
Reason for choosing course	Portuguese student dataset	Questionnaire	Nominal	
Guardian (For under 18s)	Portuguese student dataset	Questionnaire	Nominal	
Extra educational support	Portuguese student dataset	Questionnaire	Nominal	
Paid support	Portuguese student dataset	Questionnaire	Nominal	
First Language	Presentation feedback	Student information system	Nominal	
Extracurricular activities	Portuguese student dataset	Questionnaire	Nominal	
Course pre-requisites completed	Presentation feedback	Student information system	Nominal	
Next course choice	Presentation feedback	Student information system	Nominal	
Personal internet access	Portuguese student dataset	Questionnaire	Nominal	

Attribute	Source	Method of collection	Type[4]	Sensitivity
Romantic interest	Portuguese student dataset	Questionnaire	Nominal	
Mother's education	Portuguese student dataset	Questionnaire	Nominal	
Father's education	Portuguese student dataset	Questionnaire	Nominal	
Failures (course(s) re-taken)	OU	Student information system	Nominal	
Family relationships	Portuguese student dataset	Questionnaire	Nominal	
Position in family (eldest/youngest)	Presentation feedback	Questionnaire	Nominal	
Intermediate module/course grades		Student information system	Numeric	
Previous educational results		Student information system	Numeric	
Previous work experience	Presentation feedback	Student information system	Nominal	
Disability	OU	Student information system	Nominal	

Table 6.13: Evolving Static

Attribute	Source	Method of collection	Type[4]	Sensitivity
Level of other course's/module's work-load		Student information system	Numeric	
Employment during course	Small Student Dataset for HE Teachers	Questionnaire	Nominal	
Student self-assessment of progress:				

Attribute	Source	Method of collection	Type[4]	Sensitivity
Level of difficulty of topic	Research	Questionnaire	Ordinal	
Level of understanding	Research	Questionnaire	Ordinal	
Desire to go faster/slower/no change	Research	Questionnaire	Ordinal	
Changes in:				
Extra educational support	Portuguese student dataset	Questionnaire	Nominal	
Paid support	Portuguese student dataset	Questionnaire	Nominal	
Extracurricular activities	Portuguese student dataset	Questionnaire	Nominal	
Personal internet access	Portuguese student dataset	Questionnaire	Nominal	
Romantic interest	Portuguese student dataset	Questionnaire	Nominal	
Study time	Portuguese student dataset	Questionnaire	Numeric	
Free time	Portuguese student dataset	Questionnaire	Numeric	
Level of social activity		Questionnaire	Nominal	
Alcohol consumption	Portuguese student dataset	Questionnaire	Numeric	
Health		Questionnaire	Nominal	

Table 6.14: Dynamic

Attribute	Source	Method of collection	Type[4]	Sensitivity
VLE accesses (per VLE section e.g. learning materials, news, study groups)	OU	LA system generated	Numeric	
Lecture/tutorial attendances		Student information system	Numeric	
Absences	Portuguese student dataset	Student information system	Numeric	
Speed of progress (measured by per knowledge item, per course component…)		LA system generated	Numeric	
Performance in interim assessments		LA system generated	Numeric	
Speed of response to questions		LA system generated	Numeric	
Number of repeats of learning components		LA system generated	Numeric	
Assessment of:				
Perception, receiving, processing and understanding of student[1]	Research	LA system generated	Numeric	
Abilities – verbal comprehension, word fluency, computational, spatial visualisation, associate memory, perceptual speed, reasoning[1]	Research	LA system generated	Ordinal	
Level of understanding of knowledge items[1]	Research	LA system generated	Ordinal	

Attribute	Source	Method of collection	Type[4]	Sensitivity
Performance evaluated against level of difficulty of course component	Research	LA system generated	Ordinal	
Concentration[1]	Research	LA system generated	Ordinal	
Motivation[1]	Research	LA system generated	Ordinal	
Ambition[1]	Research	LA system generated	Ordinal	
Self esteem[1]	Research	LA system generated	Ordinal	
Level of anxiety[1]	Research	LA system generated	Ordinal	
Locus of Control[1,2]	Research	LA system generated	Ordinal	
Open mindedness[1]	Research	LA system generated	Ordinal	
Impetuosity[1]	Research	LA system generated	Ordinal	
Perfectionism[1]	Research	LA system generated	Ordinal	

[1] Research has been conducted aimed at potentially measuring these student attributes against carefully defined criteria, see Chapter Two Literature review (Fazey & Fazey, 2001).

[2] Locus of control is defined as "a psychological concept that refers to how strongly people believe they have control over the situations and experiences that affect their lives. In education, locus of control typically refers to how students perceive the causes of their academic success or failure in school. Students with an "internal locus of control" generally believe that their success or failure is a result of the effort and hard work they invest in their education. Students with an "external locus of control" generally believe that their successes or failures result from external factors beyond their control, such as luck, fate, circumstance, injustice, bias, or teachers who are unfair, prejudiced, or unskilled" (The Glossary of Education Reform, 2013).

[3]Deprivation index is a measure of adversity faced by students as a result of their personal lives/background. In the UK this is measured by the Index of Multiple Deprivation (UK Government, 2015).

[4]Attribute type: Numeric, Nominal, Ordinal. See section 2.4.2 for definitions.

6.3.2 Discussion

Potentially available student data for use by learning analytics is substantial, a total of 66 attributes are listed in section 6.3. These student attributes comprise a very wide variety of different types of data which may be collected by university systems, application forms, questionnaires or lecturer/tutor assessment. It should be noted that in a number of cases dynamic attributes may be expanded to include cumulative measures of each as well as measures over time and event intervals (for example, the VLE accesses or lecture/tutorial attendances). A summary of the student attributes compiled in section 6.3.1 is given below (Table 6.15).

Table 6.15: Student Attribute Summary

Category	**No. of attributes**	**Sensitivity**		
Static	31	7	5	19
Evolving static	15	1	7	7
Dynamic	20	11	0	9
Total	66	19	12	35

Over half (53%) of the student attributes are cautiously categorised as sensitive and 18% potentially sensitive, almost 30% may be considered as readily available for analysis by learning analytics processes. The majority of these 19 attributes are directly related to the student's academic background and performance during the module itself. As described in Chapter Two, Literature Review, research studies rate previous academic performance as a significant predictor of future student performance. In addition, these 19 attributes are objective and measurable, for example, previous study, intermediate assessments and VLE accesses, or objective and system calculated such as speed of response to on-line questions and number of repeats of reviews of learning objects. In comparison, the majority (43%) of the 47 sensitive and potentially sensitive attributes are based upon questionnaire completion and student self-assessment, which although valuable in developing learning analytics models are in a number of cases subjective measures.

There is also evidence that the measurement of student attendance at lectures and tutorials provides useful predictive data of likely student outcomes (Aziz & Awlla, 2019; Fike & Fike, 2008), as is the case of student engagement with the VLE (Umer et al., 2018). A number of the dynamic attributes in Table 3.16

may be collected and analysed on a temporal basis in order to identify trends that may enhance prediction accuracy, for example VLE accesses and attendance. There is some evidence that students whose VLE activity is early in a given learning cycle (i.e. in advance of the topic being taught) are more likely to be successful (Nguyen et al., 2018). In addition, in the case of MOOCs which by definition are entirely on-line study, considerable research has been carried out on whether students likely to drop out may be identified by measuring detailed temporal on-line activity (Vitiello, 2018).

Table 6.16: Summary of Attribute Types and Associated Event (Vitiello, 2018, p7)

Type of attribute	**Associated events to be measured**
Session Related	Sessions, Requests, Active Time, Days, Timespan Clicks, Session Length, Session Requests, Day Requests
Main Page Links	About, Faqs, Home, Instructor, Progress, StudyAtCurtin
LMS	TabSelected, PreviousTabSelected, NextTabSelected, LinkClicked, OutlineSelected
Video	CaptionHidden, CaptionShown, LanguageMenuHidden, LanguageMenuShown, Loaded, Paused, Played, PositionChanged, SpeedChanged, Stopped, TranscriptHidden, TranscriptShown
Video Mobile	CaptionHiddenM, CaptionShownM, LanguageMenuHiddenM, LanguageMenuShownM, LoadedM, PausedM, PlayedM, PositionChangedM, SpeedChangedM, StoppedM, TranscriptHiddenM, TranscriptShownM
Problem	Check, CheckFail, FeedbackHintDisplayed, Graded, HintDisplayed, Rescore, RescoreFail, Reset, ResetFail, Save, SaveFail, SaveSuccess, Show, ShowAnswer
Poll & Survey	PollSubmitted, PollViewResults, SurveySubmitted, SurveyViewResults
Bookmark	Accessed, Added, Listed, Removed
Forum	CommentCreated, ResponseCreated, ResponseVoted, Searched, ThreadCreated, ThreadVoted

6.4 Chapter Summary

In this chapter I have described each of the datasets used in my research and comprehensively catalogued the very wide variety of student attributes I encountered. These datasets were selected to represent a variety of small, medium and large student cohorts and similar ranges of student attributes. In the case of the student attributes they include readily accessible and uncontroversial (from an ethical, moral and privacy perspective) features such as attendance at lectures and interim assessments as well as highly personal and sensitive features such as demographics and alcohol consumption. A proportion of these attributes are measurable and unambiguous such as attendance and age, while others are system generated such as VLE accesses and previous education results or "student provided" via questionnaire such as "weekly study time" and "current health status". I have categorised these attributes as Fixed Static (e.g. age, family size), Evolving Static (e.g. other academic work load, employment) and Dynamic (e.g. VLE accesses, attendance). I have assigned a subjective sensitivity indicator to each in order to consider the level of challenge that institutions may face in the collection and use of the attribute. As discussed in the previous section, almost 30% of all the attributes considered are not classified as sensitive or potentially sensitive and the majority of these are measurable and directly related to the student's academic background and performance. I have also identified evidence that student attendance, interim assessments and VLE activity provide useful predictive data for learning analytics. These analyses and results provide the platform for the following chapter which focuses upon my exploration of alternative AI/ML techniques for the prediction of student outcomes, using the datasets described here. In the following chapter I provide an explanation of each of the AI/ML techniques relevant to my research and describe the experiments I have conducted on the datasets described in Chapter Three and describe my contribution of identifying a novel technique for the analysis of nominal data.

CHAPTER SEVEN

Relevant AI and ML Techniques

7.1 Introduction

7.1.1 Contributions to Knowledge Relevant to this Chapter

In this chapter I explore alternative AI/ML techniques for predicting student outcomes. Section 7.3 describes my contribution to knowledge of the development of a novel technique for the analysis of nominal data. In section 7.4.3.4 I present the results of applying this technique to the analysis of the nominal attributes of the Portuguese student data set (described in section 6.2.3). In section 7.4.3.5 I compare and discuss these results with those given by applying the Chi-square test method for the analysis of nominal data:

The following sections of this chapter are supported by previously published material:

Section 7.2 Artificial Intelligence and Machine Learning Techniques (Wakelam et al., 2015)

Section 7.3 Novel technique for the analysis of nominal data (Wakelam et al., 2016)

Section 7.4.3 Portuguese secondary school student achievement (Wakelam et al., 2016)

Section 7.4.3.3 Experimental analysis (Wakelam et al., 2016)

7.1.2 Summary of Chapter Content

I describe each of the artificial intelligence, machine learning and statistical techniques relevant to my research, including my own novel technique for the analysis of nominal data. I apply a variety of techniques to freely available student datasets, both small and large and comprising both limited and multiple student attributes. I consider both numeric and nominal attributes. I describe each of my own experiments using a variety of these techniques on the datasets described in Chapter Seven (applied to freely available datasets and including a summary of an experiment conducted on a live student cohort).

7.2 Artificial Intelligence and Machine Learning Techniques

Before describing individual AI and ML techniques, I briefly summarise the methods and terms relevant to a number of these techniques in general.

As a general rule, machine learning approaches follow the following process:

- Collect the required data
- Identify and correct missing data points/anomalies as required

- Prepare the data as required by the selected machine learning model(s)
- Establish a baseline model that you aim to exceed
- Train the model on the training data (randomly selected from the dataset)
- Make predictions on the test data (randomly selected from the dataset)
- Compare predictions to the known test set targets and calculate performance metrics
- If performance is not satisfactory, adjust the model, acquire more data, or try a different modelling technique
- Interpret model and report results visually and numerically

In the case of learning analytics features are typically referred to as student attributes (e.g. attendance, course results, gender). An important component of machine learning analyses is the selection of appropriate features, discarding those which are seen to have very little or no effect on ML results accuracy and the derivation of additional features from existing ones where the accuracy of machine learning results may be improved. This process is called feature engineering and forms a significant component of data scientist's activities in the analytics process. A Forbes survey indicates that data scientists spend 80% of their time in the data preparation activity (Press, 2016). The derivation (synthesising) of additional features from existing ones is a common practice to either augment or replace parts of the existing dataset (Heaton, 2016). Heaton (2016) presents the benefits of deploying ten types of engineered features including counts, differences and logarithms on each of Deep Neural Network, Random Forest, Support Vector Machines and Gradient Boosted Machines (e.g. Decision Trees) machine learning techniques. Heaton concludes that feature engineering may not always be effective for every data set however, in some cases prediction accuracies may be improved by a statistically significant amount.

Regression and classification are supervised machine learning techniques which use known datasets (training datasets) to make predictions. Garbade (2018) provides a simple hierarchical chart illustrating this (Figure 7.1).

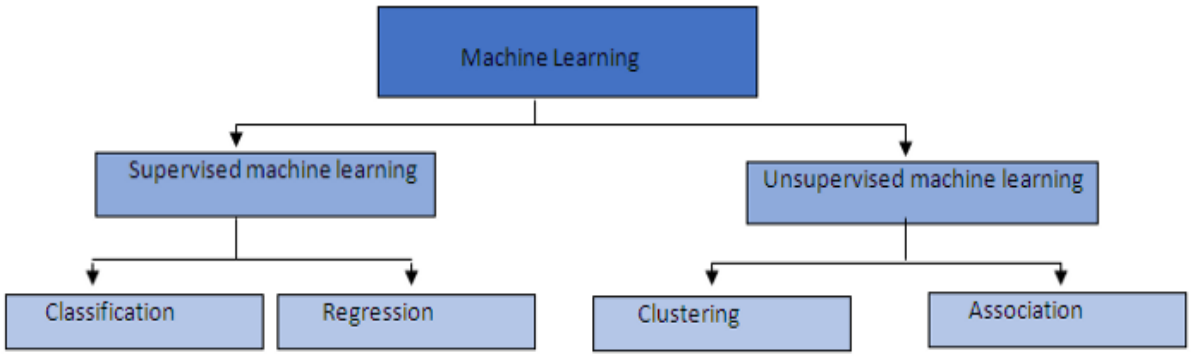


Figure 7.1: Machine Learning Branches (Garbade, 2018)

In the case of prediction by classification, the output variable (prediction) is categorical (discrete or nominal), whereas in the case of regression the output variable is numerical (continuous). This can be illustrated by the example of predicting student assessments (e.g. examination) outcomes. The application of a regression technique would provide a numerical output such as 62 marks out of 100 (i.e. 62%), whereas a classification technique would provide a categorical output such as pass or fail, or perhaps A, B, C, D, E, F. A more detailed description of the difference between regression and classification is available here (Garbade, 2018).

Some techniques are described as non-parametric, these include Support Vector Machine and K-Nearest Neighbour. This means that the technique makes no assumptions regarding the underlying data distribution of the dataset, instead determining the structure of the data model solely using the data it is presented with. This is valuable in real world problems where the data may often be very random and not in line with typical theoretical assumptions.

Each of the AI, ML and statistical techniques relevant to my research are each described in the following sections:

7.2.1 Support Vector Machine (SVM)

SVM is a supervised learning algorithm which allows us to classify data in a way in which we can then analyse new data points to confidently identify which solution space they fit within (Chang & Lin, 2011). Of particular value is that they can perform this on multi-dimensional data by mapping to a two or three dimensional space where the boundaries between the data attributes can be identified. SVM algorithms can solve linear and non-linear problems and works well for many practical problems creating a line or a hyperplane which separates the data into different classes (Figure 7.2), (Pupale, 2018). Linear problems are those where variables are of power one and their graphical representation is a straight line, for

example, x + y = 0. Note that this is also true for multi-dimensional problems, for example, ax + by + cz = 0. Non-linear problems are those which include variables with powers of 2 or more, for example, $x^2 + y^2 = 0$, or include complex multiples of variables or mathematical functions, such as xy = 0 or y = sin(x), and whose graphical representations are not straight lines.

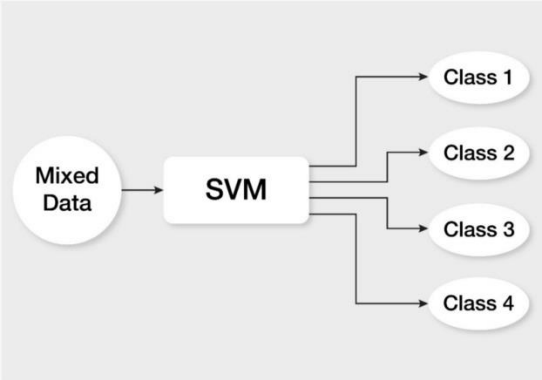


Figure 7.2: SVM Classifier (Pupale, 2018)

SVM can be employed for both classification and regression purposes. It is more commonly used in classification problems.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes (in practice, SVM can be used to classify multi-classes), as shown in Figure 7.3 (Bambrick, 2016).

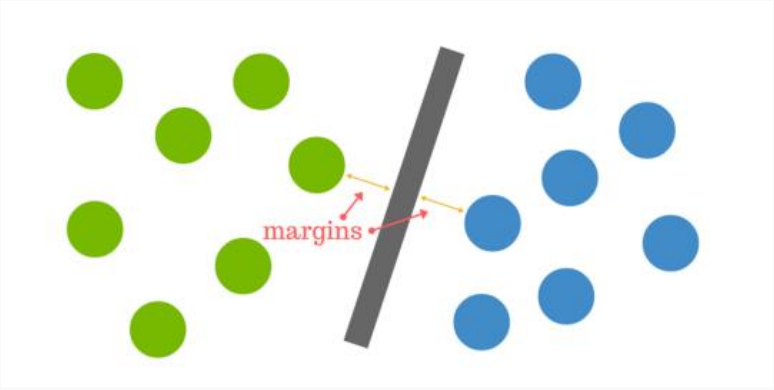


Figure 7.3 Dividing a Dataset into Two Classes (Bambrick, 2016).

SVM analysis requires the selection of a regularisation parameter, often denoted as "c", which tells the SVM optimisation the degree to which misclassifying each training example can be avoided (Patel, 2017).

Large values of c will cause optimisation to choose a smaller-margin hyperplane if doing so does a better job of getting all the training points classified correctly. Small values of c will cause optimisation to choose a larger-margin separating hyperplane, at the expense of the misclassification of more points.

The images below (Figure 7.4 and Figure 7.5) are examples of the selection of two different values of c. The selection of a low value for c led to some misclassification (Figure 7.5), whereas a higher value results like right one.

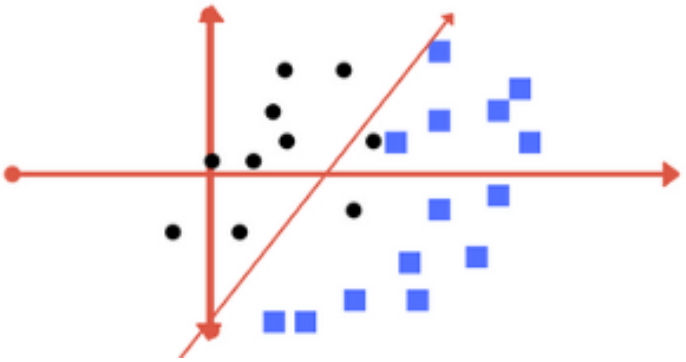


Figure 7.4: Low Value for Regularisation Parameter c (Patel, 2017)

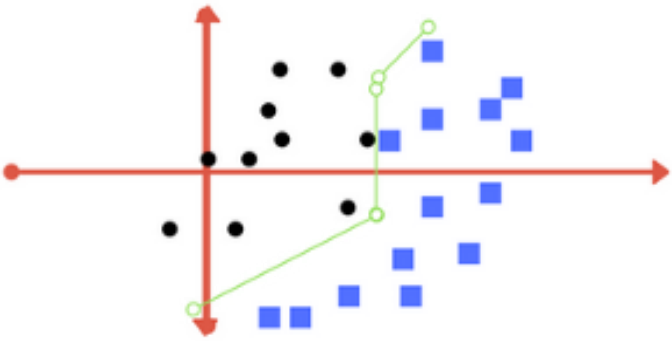


Figure 7.5: High value for Regularisation Parameter c (Patel, 2017)

In the following dataset (Figure 7.6) (Pupale, 2017) we wish to classify the red rectangles from the blue ellipses and hence find an ideal line. that separates this dataset in two classes (say red and blue).Clearly, the green line separates the data into two classes, but we are looking to identify the line which best separates the data such that when we introduce new data points we classify them accurately. In our example it is the orange line that does this.

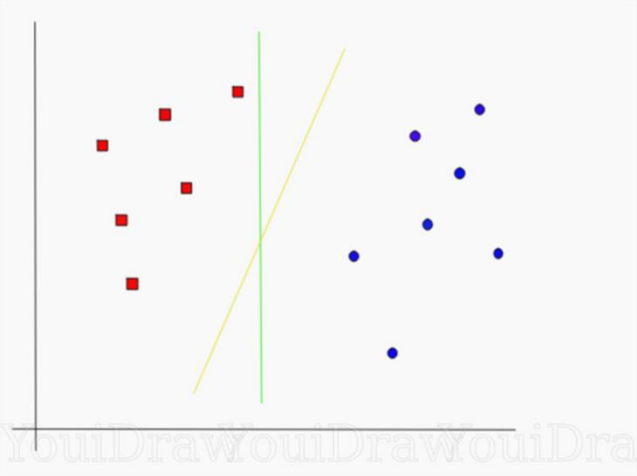


Figure 7.6: Hyperplane Classification of a Dataset (Patel, 2017)

Support vectors are the data points nearest to the hyperplane, the points of a dataset that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a dataset.

In the case of two dimensional data the line dividing the two sets of data is very simple to visualise. Where the data is three dimensional this line becomes a plane and is still visualisable in a three dimensional graph. However, for higher dimensions, we cannot visualise the dividing structure and we describe this structure as the hyperplane. For simplicity in SVM analysis we use the term "hyperplane" for all dimensions of data, including two and three dimensional, as the line that linearly separates and classifies a set of multi-dimensional data.

Clearly, the further our data points lie from the hyperplane, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

Hence, when new testing data is added, whichever side of the hyperplane a data point resides decides the class that we assign to it.

The objective of SVM is to determine where we identify the hyperplane for a given set of data allowing segregation of the classes of data.

The distance between the hyperplane and the nearest data point from either set is known as the margin (Figure 7.7). The goal is to select a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.

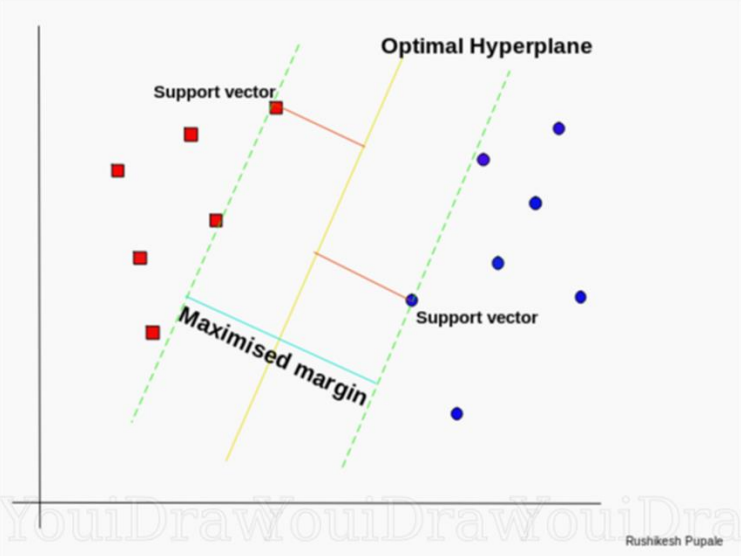


Figure 7.7: Margins and the Optimal Hyperplane (Pupale, 2018)

In the above two dimensional data example, it is simple to visualise how the two sets of data points are separated. However, in practice datasets will often appear more complex (Figures 7.8 and 7.9), (Bambrick, 2016).

Figure 7.8: Two Dimensional View of the Dataset (Bambrick, 2016)

In this case, we may move from a two dimensional to a three dimensional view of the data (Figure 7.9).

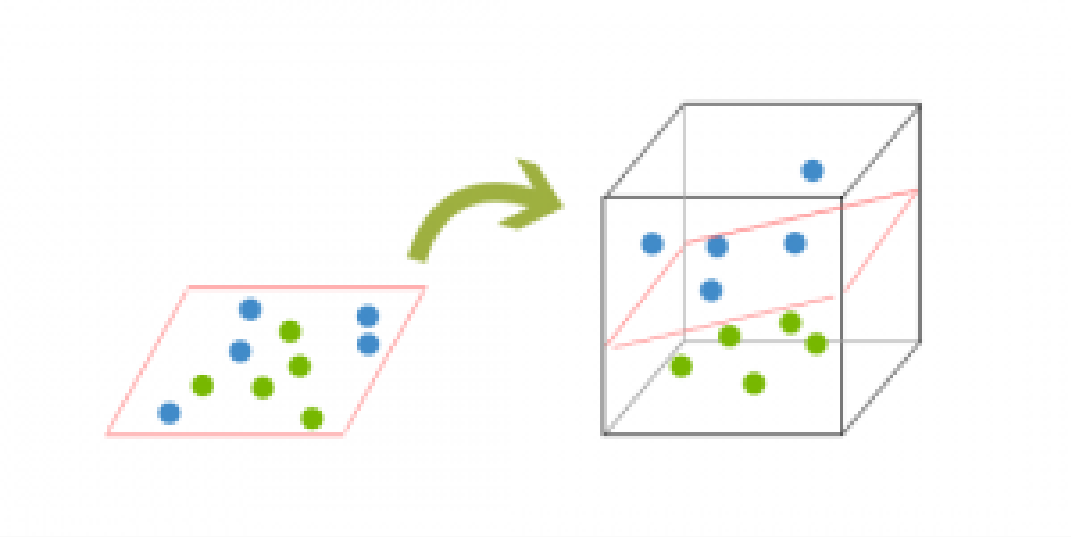


Figure 7.9: Three Dimensional View of the Dataset (Bambrick, 2016)

The hyperplane is no longer represented as a line, instead as a plane, and a clear classification of the data is visible. SVM allows data to be mapped into higher dimensions using what is described as a "kernel trick" until a hyperplane which successfully segregates the data is found. The kernel trick allows SVM to operate in the original feature space without computing the coordinates of the data in a higher dimensional space. A detailed explanation of the underlying mathematics is available here (Zhang, 2018). One of three kernel types (linear, polynomial or radial basis function (RBS)) is selected and input as a parameter before executing SVM.

The SVM algorithm calculates the position of the hyperplane by finding the data points closest to the line from both the classes. These points are called support vectors. SVM computes the distance, called the margin, between the line and the support vectors. The line (hyperplane) for which the margin is maximum is the optimal hyperplane.

SVM is suited to the analysis of numeric data. Given that SVM is based upon Euclidian distances it cannot be applied directly to categorical data. However, it is possible to allocate suitable numeric values (dummy variables) to represent the categoric data, for example "yes" is allocated the value "1" and "no" is allocated the value "2" (Peng & Li, 2019). Although this allows the application of SVM to mixed datasets, it may not exploit the strengths of the technique.

The pros of SVM are its accuracy, its good performance on smaller, cleaner datasets and that because it uses a subset of training points it can be more efficient than other techniques.

The pros and cons of SVM are as follows (Bambrick, 2016), SVM works well when the dataset is not easily understandable or unstructured, scaling well to high dimensional data, with less risk of overfitting, and the ability to select a kernel for the analysis allows the solution of complex problems. The cons are that it is less efficient on noisier datasets and in the case of large datasets the training time can be high. Also, selecting an appropriate kernel is not straightforward, long training times for large datasets and while the final model (clustering) is easily visualisable the process to arrive at it is not transparent.

7.2.2 Principal Component Analysis (PCA)

Where datasets comprise of two or even three dimensions the data may be presented graphically and provide for interpretation by the human eye and a variety of statistical techniques. However, the availability for analysis of multi-dimensional data is increasingly widespread in a variety of disciplines and in such cases direct human analysis is impossible. PCA is a widely used technique to drastically reduce dataset dimensionality, while preserving as much information from the whole dataset as possible, presenting interpretable analyses (Jolliffe & Cadima, 2016).

The PCA algorithm may be applied as follows (Jaadi, 2019).

As is the case with many machine learning techniques, we must standardise the data to ensure that each data point contributes equally to the analysis. If there are large differences (variances) between the ranges of the initial variables, those with the larger ranges will dominate over those with small ranges (for example, a variable with a range between 0 and 1000 will dominate over one that ranges between 0 and 1). This is done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

PCA seeks to identify any relationships between the variables of the input dataset and how they are varying from the mean with respect to each other. Sometimes, variables in a dataset are highly correlated with each other and therefore contain redundant information. In order to identify these correlations, we compute the covariance matrix.

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, Figure 7.10 shows the covariance matrix for a three dimensional dataset with 3 variables *x*, *y*, and *z*.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Figure 7.10: Covariance Matrix for a Three Dimensional Dataset

Note that covariance is commutative (Cov (a,b) = Cov (b,a)), therefore the covariance matrix entries are symmetric with respect to the main diagonal), hence the upper and the lower triangular portions are equal.

The value of each entry of the covariance matrix describes the magnitude of correlation of the two variables. If the covariance value is positive, then the two variables increase or decrease together (correlated). If the value is negative, then one variable increases when the other decreases (inversely correlated).

The next step in PCA is to compute the eigenvectors and eigenvalues (Smith, 2002) of the covariance matrix to identify the principal components of the data.

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables, computed in such a way that the new variables (i.e., the principal components) are uncorrelated and such that most of the information within the initial variables is squeezed or compressed into the first components.

For example, ten dimensional data gives ten principal components, however PCA aims to put maximum possible information in the first component, then maximum remaining information in the second and so on, for example Figure 7.11.

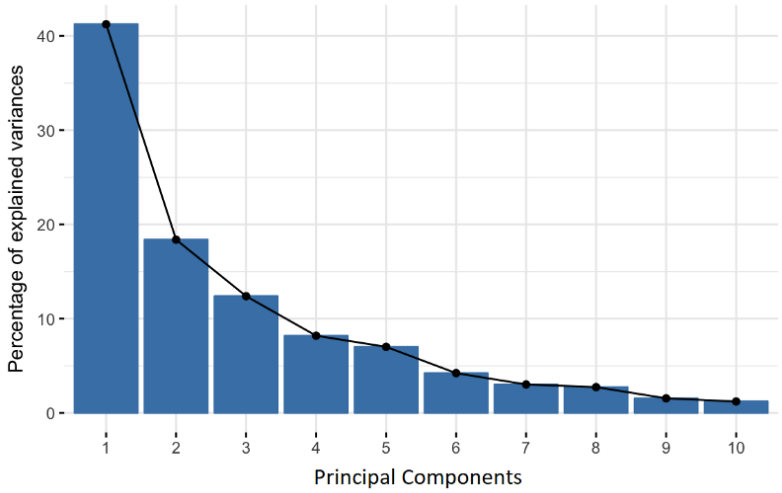


Figure 7.11: Percentage of Variance (Information) for by each Principal Component (Kassambara, 2017)

By organizing the information in our dataset into principal components in this way, dimensionality is reduced without an unacceptable loss of data, and by discarding the components with low information the remaining components may be considered as the new variables.

A simple way of thinking about principal components is to consider them as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

A description of the mathematics that the PCA algorithm applies to construct the principal components is available here (Jaadi, 2019).

PCA is suited to the analysis of numeric data, however, new developments in the application of kernel approaches to select reasonable dummy variables have now been developed to deal with categorical data (Niitsuma, H. & Okada, T., 2005).

The pros of PCA are that it removes correlated features, which in turn improves the performance of the algorithm, reduces the risk of overfitting and provides a visualisable output. The cons are that the principal components are not readable and interpretable as the original features are, the data must be normalised beforehand and categorical features must be converted to numerical.

7.2.3 Neural Networks (NN)

NN were inspired by studying how the brain works. They are composed of a large number of highly connected processing nodes which work in unison to solve specific problems (Marr, 2018). They can derive meaning from complicated or imprecise data extracting patterns or detecting trends that are too complex to be identified by other techniques.

NNs use multiple layers of mathematical processing to make sense of the information it's supplied with. An NN may have from dozens to millions of artificial neurons arranged in a series of layers (Figure 7.12). NNs are adaptive because they have the ability to change their internal structure by adjusting input weightings.

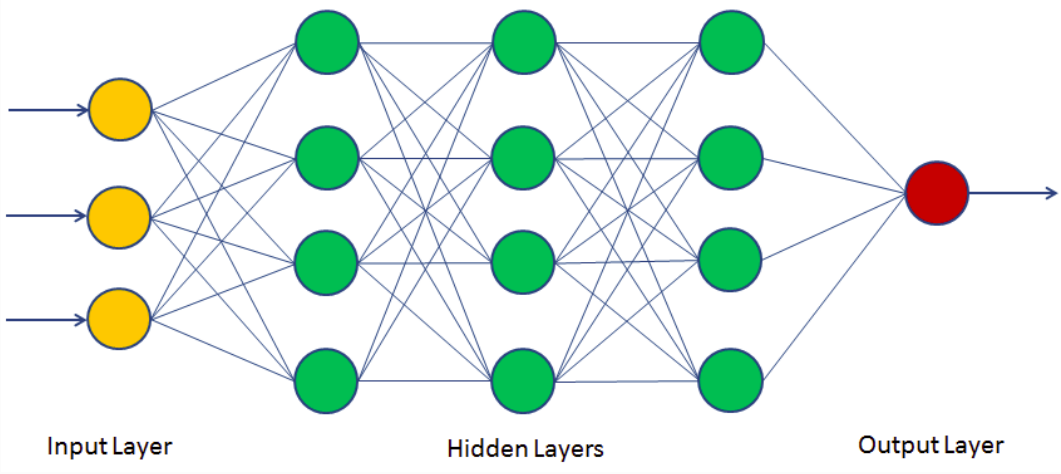


Figure 7.12: Multi-layer Neural Network (Naviani, 2019)

An NN is a set of connected input/output units in which each connection has a weight associated with it. In a learning phase, the network learns by adjusting the weights to predict the correct class label of the given inputs.

The input layer receives data to be processed from the outside world. Data progresses from the input unit through one or more hidden units, layer by layer, with the objective of transforming the data into something the output unit can use.

Most NNs are fully connected from one layer to another. Each connection is weighted, with the magnitude of the weighting defining the level of influence one unit has on another (as in the human brain). As data progresses through each unit the network more is learnt about the data.

The two main types of NNs are feedforward and feedback. In a feedforward NN the neurons in each layer are only connected to neurons in the next layer, and processing travels only in the direction of the output layer. In a feedback NN signals travel in both directions through the introduction of loops in the network.

NNs are suited to the analysis of numeric data and cannot be applied directly to categorical data. However, as with SVM, it is possible to allocate suitable numeric values (dummy variables) to represent the categoric data. Alternative methods of encoding categorical data are presented and discussed by (Brokmeier, 2019) (Potdar et al., 2017).

The pros of NNs are that they can be used for both regression and classification, they are not limited by the number of inputs and layers, and they can perform processing in parallel. In addition, they are able to deal with a non-linear dataset with large numbers of inputs, such as image recognition. The cons are that

NNs are a black box approach (difficult for humans to understand their analyses), take longer to create and require more computing power (Naviani, 2019).

In recent years, NN has been extended into a technique which uses neural networks with multiple layers to allow the computer to learn to filter inputs (patterns such as images, text or sound) through each layer in order to classify data. This extended technique is called Deep Learning (Marcus, 2018), the multiple "hidden" layers are shown in Figure 7.13 in comparison with a simple neural network (Vázquez, 2017). In essence, the difference between the neural network technique and deep learning is the depth of the model i.e. the number of layers and consequently the complexity of the number of paths that may be taken through the model. For example, traditional neural networks may only contain 2-3 hidden layers, while a deep learning network may have in excess of 100.

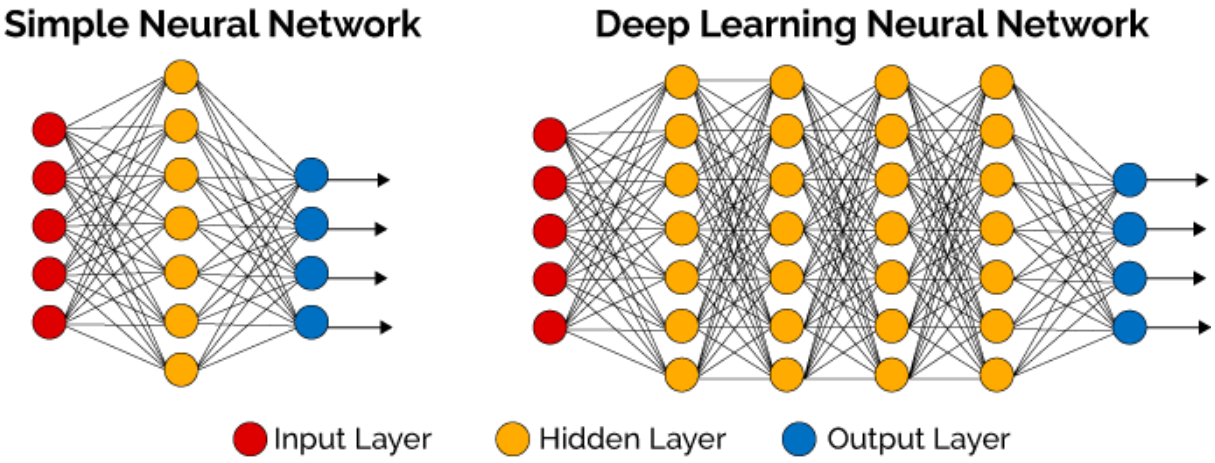


Figure 7.13: Neural Network vs Deep Learning (Vázquez, 2017)

Research into the application of deep learning in the field of educational data mining has become increasingly popular in recent years. A review of publications on this topic shows an increase of papers published from 3 in 2015 to 17 in 2018 (Hernández-Blanco et al,, 2019).

7.2.4 Growing Neural Gas (GNG)

The GNG algorithm (Fritzke et al., 1995) is an unsupervised clustering method. It iteratively grows a graph to map the data in the sample vector space. When complete, each data point may be seen as part of one of the groups allowing their classification. This mapping is a type of Self Organising Map (SOM) techniques. A SOM is a type of neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), representation of the input space of the training samples, called a map. It is therefore a method of dimensionality reduction.

GNG is a graph consisting of a set of nodes and a set of edges connecting the nodes. Each node is given a weight vector corresponding to its position in the input space and an error variable intended for identification of the parts of the network least adapted to the input signals. An edge is a line connecting a pair of nodes. Initially, GNG places two randomly generated nodes into a network and repeats (alternates) two phases until a selected stopping criterion is met. Phase 1 (the self-organising phase) is performed in a number of steps. In each step, a random input signal is generated and the neural network adapts itself to it by strengthening or creating a connection between two nodes nearest to the input signal. The nearest node and all its topological neighbours (nodes connected directly to the node by an edge) are then moved towards the input signal and the nearest node's error is increased (identifying areas where nodes are not sufficiently adapted to input signals). An aging mechanism of edges is then triggered, removing those edges that had not been strengthened for a long time from the network. The last step of the adaptation the error of each node is decreased (allowing the neural network to forget old errors allowing it to focus on the most recent ones). In phase 2 (the growing phase) a new node is created and connected into the network. This node's error is used for to identify the area where the adaptation was least successful i.e. identifying the node with the largest error and its neighbour with the largest error. A new node is created at the halfway between them. The errors of those nodes are then decreased (Fiser et al., 2013).

A detailed description of the application of GNG, demonstrated by its application to quantifying hard retinal exudates, may be found in (Csefalvay, 2019).

GNG is suited to the analysis of numeric data and cannot be applied directly to categorical data.

The pros of GNG are the technique's ability to find optimal clusters in data without prior information about the number of optimal clusters and its improved performance over other methods (Jirayusakul & Auwatanamongkol, 2007). The cons are that its computational expense is too high when dealing with a large numbers of features.

7.2.5 Decision Tree (DT)

DTs are a tool that allows the creation a tree-like picture of decisions and alternative next steps. They allow us to determine a strategy to reach a defined goal. A decision tree reaches its decision by performing a sequence of tests, with each internal node in the tree corresponding to a test of the value of one of the input attributes. The branches from each node are labelled with the possible values of the attribute, and each leaf node in the tree specifies a value to be returned by the function.

Decision Trees are a class of very powerful ML technique achieving high accuracy while being highly interpretable. The knowledge learned by a decision tree is displayed in a hierarchical structure in a way that

can be easily understood, even by non-experts. Classification And Regression Tree (CART) is an umbrella term for Decision Tree techniques which can be applied to conduct predictive modelling using classification or regression techniques (Brownlee, 2016).

The following example (Figure 7.14) shows how an individual may formulate the decision on what activity to do at the weekend, described in decision tree terms, by following a set of sequential, hierarchical decisions that lead to a final result.

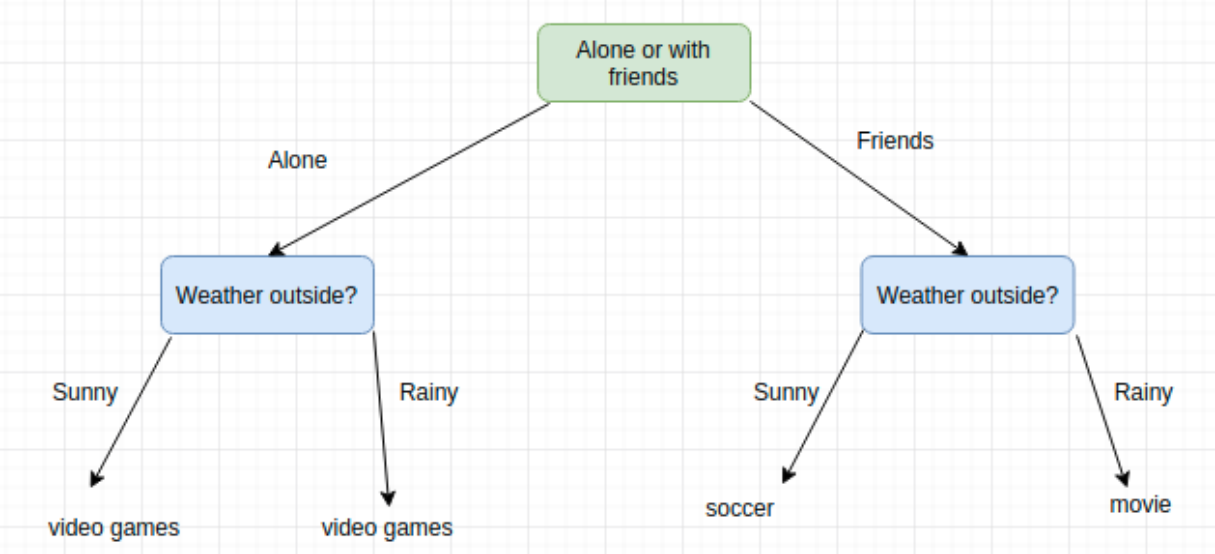


Figure 7.14: Example Decision Tree (Seif, 2018)

A decision tree model is created in two steps, induction and then pruning (Seif, 2018). Induction refers to the process of building the tree based upon the set all of the hierarchical decision boundaries based on the dataset. Because of the nature of training, decision trees can be prone to major overfitting. Pruning is the process of removing any unnecessary structure from a decision tree, reducing the complexity to combat overfitting. In addition, this reduction in complexity makes the resulting tree easier to interpret.

Induction comprises of four steps. Firstly, after extracting a training set from the dataset, the best feature to split the data on is determined ("best" is selected by considering the highest number of features to consider when looking for the best split). The data is split into subsets that contain the possible values for this best feature. This splitting basically defines a node on the tree i.e. each node is a splitting point based on a certain feature from our data. A detailed description of the mathematics that the DT algorithm applies to identify the "best" feature is available is here (Seif, 2018).

The data is then split into subsets that contain the possible values for this best feature, defining a node on the tree (each node is a splitting point based on a certain feature from the data). New tree nodes are then recursively generated, repeatedly spitting the data until a point is reached where maximum accuracy with the minimum number of splits/nodes has been optimised. A description of the mathematics for the identification of the optimum point to halt the recursion is also available in (Seif, 2018).

After the induction step is completed, the decision tree is pruned to avoid overfitting. If the decision point splitting value is too small the tree will have a large number of splits and consequently a very large and complex tree. In this case many of the splits will be redundant and have no effect on the accuracy of the model. If the values are too large then the decision tree will not perform a valuable analysis of our dataset.

Pruning is a technique that leverages this splitting redundancy to remove the unnecessary splits in our tree. It compresses part of the tree from strict and rigid decision boundaries into ones that are smoother and generalise better, reducing the tree complexity (defined as the number of splits in the tree). A description of the mathematics of various pruning methods is available here (Patel & Upadhyay, 2012).

DTs are suited to the analysis of both numerical and categoric data.

The pros of DT are its ease of understanding and interpretation, the need for very little data preparation and that the cost (effort of the algorithm compared with the accuracy of its results). The cons are that DTs are prone to overfitting, occasionally requiring dimensionality reduction (such as PCA) before application and that the overfitting can result in bias towards classes which have a majority in the dataset. This may be avoided by balancing the classes using weighting techniques.

7.2.6 Random Forest (RF)

RF consists of multiple randomly created decision trees (see Section 7.2.5). It is an "ensemble" (combination of more than one method, such that a group of weak learners can be combined to create a strong learner and hence more accurate predictions) learning method for classification and regression analyses of datasets. Each tree in the forest is built from a random sample of the original dataset and at each tree node the best split is selected randomly from a subset of features. This dual randomness removes the risk of overfitting. RF analyses are typically more accurate than Decision Trees given that they consist of multiple single trees each of which is based on a different random sample of the training data. A detailed description of the application of RF analysis is available here (Deng, 2018a).

Unlike decision trees, which require pruning to avoid overfitting, RF trees are fully grown and unpruned and therefore the feature space is split into more and smaller regions (Deng, 2018b). The pros of RF are that they are often accurate, they do not require feature scaling, categorical feature encoding, and need little

parameter tuning. In particular they do not suffer the overfitting issue which affect decision tree analyses. They can also be more interpretable than other complex models such as neural networks.

RFs are suited to the analysis of both numerical and categoric data.

7.2.7 K-Nearest Neighbour (KNN)

KNN is a technique that classifies data points based on the points that are most similar to it. It does so by using test data to make an educated guess on what an unclassified point should be classified as. It classifies data points by comparing it to its nearest points in the training set and classifies it based on which points it is closest and most similar to. The algorithm decides on "closest" by measuring the distance between these points, often using Euclidian distance measures. There are other measures which may be used, for example Cosine Similarity.

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

In the Euclidian distance measure case, KNN applies the above formula calculating the distance between each data point and the test data. The probability of these points being similar to the test data allows KNN to classify based on which points have the highest probabilities (Schott, 2019). The algorithm picks the "k" closest data points (those points with the "k" lowest distances) and using majority voting across the data points decides the final classification. The optimum value of "k" is selected by trial and error usually starting with k=1, k=2 etc.

KNN is suited to the analysis of numeric data. Given that KNN is based upon distance measures (often Euclidian) it cannot be applied directly to categorical data. However, as with SVM, it is possible to allocate suitable numeric values (dummy variables) to represent the categoric data (Peng & Li, 2019). Similarly, while this allows the application of KNN to mixed datasets, it may not exploit the strengths of the technique.

The Pros of KNN are that it is simple to use, with fast calculation times and it does not make assumptions about the data. Its cons are that the accuracy of the technique depends upon the quality of the data, it is necessary to find (trial and error usually) an optimal value for the parameter k. KNN can also be poor at successfully classifying data points where they are close to a boundary where they could be classified on one side or the other.

A detailed description of the application of KNN analysis and its underlying mathematics is available here (Soni, 2018).

7.2.8 Naïve Bayes Classification

Naïve Bayes Classification is based upon the Bayes theorem which determines the probability of an event A happening, given that an event B has occurred. This technique is called "naïve" because the assumption is made that the predictors/features are independent of each other, i.e. any one feature does not affect any other. Bayes theorem is as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A detailed description of Bayes theorem and its underlying mathematics is available here (Oppermann, 2018).

The application of the Naïve Bayes technique may best be explained by following a worked example (Chauhan, 2018). Chauhan describes having data on 1000 pieces of fruit, either bananas, oranges or some other fruit and that 3 features of each fruit are known, whether it's long or not, sweet or not and yellow or not (see Table 7.1).

Table 7.1: Fruit Dataset (Chauhan, 2018)

Fruit	**Long**	**Sweet**	**Yellow**	**Total**
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

The data shows that the proportions of each of bananas, oranges and other fruits are 50%, 30% and 20% respectively. Proportions (and consequently the probabilities) of each feature are also evident, for example 80% of bananas are long, so a probability of 0.8.

Prediction of the class of a new fruit can now be done by applying the Bayes theorem formula using the probabilities from our test dataset (Table 7.1) and the features of the new piece of fruit. Determining the probability of which type of fruit it is, as follows:

Banana:

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = \frac{P\left(\frac{Long}{Banana}\right) \times P\left(\frac{Sweet}{Banana}\right) \times P\left(\frac{Yellow}{Banana}\right) \times P(Banana)}{P(Long)\, P(Sweet)\, P(Yellow)}$$

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = \frac{(0.8) \times (0.7) \times (0.9) \times (0.5)}{0.25 \times 0.33 \times 0.41}$$

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = 0.252$$

Orange:

$$P\left(\frac{Orange}{Long, Sweet, Yellow}\right) = 0$$

Other Fruit:

$$P\left(\frac{Other}{Long, Sweet, Yellow}\right) = \frac{P\left(\frac{Long}{Other}\right) \times P\left(\frac{Sweet}{Other}\right) \times P\left(\frac{Yellow}{Other}\right) \times P(Other)}{P(Long)\, P(Sweet)\, P(Yellow)}$$

$$P\left(\frac{Other}{Long, Sweet, Yellow}\right) = \frac{(0.5) \times (0.75) \times (0.25) \times (0.2)}{0.25 \times 0.33 \times 0.41}$$

$$P\left(\frac{Other}{Long, Sweet, Yellow}\right) = 0.01875$$

In the above example, banana has the highest probability, 0.252.

The pros of the Naïve Bayes technique are its speed and ease of prediction, also performing well in multi-class situations, given the assumption of independence of features it outperforms many other models, and it performs well in the case of categoric (nominal) variables. Naïve Bayes is suited to the analysis of both numerical and categoric data. The cons are that if a variable that is not in the training set (in our example, for example, a strawberry) then the algorithm will assign a zero probability and fail to make any prediction, and that it is only a valid technique if the variables are truly independent. It should also be noted that Bayes probability outputs are only useful in the classification process and not to be regarded as accurate in their own right.

4.2.9 Knowledge Based Systems (KBS)

KBS, which are sometimes referred to as Expert Systems, use a set of rules to solve problems and support decision making based upon stored expert knowledge (Rouse, 2018). These expert rules are usually encoded by extraction from human experts. A good example is medical diagnosis, where a data base of medical conditions including symptoms and treatments is created and a doctor defines the logical steps to be followed to apply these rules in a dialogue with a patient in order to arrive at a diagnosis.

Typically a KBS comprises of two components, a knowledge base (a database of stored knowledge and rules for the analysis and application of that knowledge to determine useful outcomes in a particular and well defined field of knowledge) and an inference engine, which deduces insights from the knowledge base and rules and is capable of interrogation by human users. This engine applies selected AI/ML techniques to support and implement the defined rules. The interrogation is made possible through a variety of user interfaces.

Alternatively to those KBSs which apply expert rules, other KBSs apply what is referred to as case-based reasoning. These are a library of solutions to existing problems/situations that may be applied to a new problem.

At its simplest, a KBS may follow the equivalent of a flowchart of questions and branches leading to an outcome. In simple cases, the medical diagnosis example can illustrate this. A medical practitioner creates the flowchart (or tree) of diagnostic questions, which when followed results in a proposed diagnosis and next steps.

Pros of KBSs are that they reduce the workload on human experts and they collect and retain a record of the data and rules that they apply in diagnostic/advisory situations, information is rapidly and accurately retrievable, and they are able to provide clear explanations of how individual outcomes were arrived at (Raj, 2019).

Cons are that there creation demands substantial expert time to create the knowledge base and expert rules and then thoroughly test and prove the operation of the KBS, essential in safety critical fields such as medicine. In addition there is an on-going imperative of operating a rigorous and sustainable regular updating of the knowledge and rules in line with advancing knowledge. In both cases, by their nature, the experts are a very expensive resource.

7.2.10 Fuzzy Logic

Fuzzy logic allows us to use degrees of truth/accuracy in data analysis rather than the black or white ones and zeroes or yes and no's traditionally used in systems (Benabdellah, 2014). The application of the Fuzzy Logic technique may best be explained by following a worked example (Ghoneim, 2019). Ghoneim describes how traditional classification may classify a cup of coffee into one of two sets, hot or cold, hence a lukewarm coffee would fall into the category hot.

Fuzzy logic is an approach to computing based upon a numerical measure of the degree of truth rather than simply true or false. Therefore, each element of a set has a degree of membership to every set that it is contained in. In the lukewarm case above we might assign values of "0.7 hot" and "0.3 cold" as the respective degrees of membership to the hot and cold sets to the coffee.

Fuzzy logic rules may be used in a variety of AI/ML techniques, for example Neural Networks and Knowledge Based/Expert Systems (Priy & Rajput, 2019).

The pros of fuzzy logic its alignment with formal set theory, its ability to deal with noisy data, their construction is simple and understandable, it resembles human reasoning and requires little data and hence less memory. The cons are that a given problem can be approached in a variety of ways which may lead to ambiguity and given its application to both precise and imprecise data its accuracy may be compromised.

7.2.11 Ant Colony Optimisation

Ant colony optimisation (ACO) is an algorithm for establishing the optimal paths in data and processes that is based upon how ants leave pheromone markers to show the path to food that they have found (Sivakumar & Praveena 2015).

In ACO, artificial ants, represented by software agents search for optimum solutions to a given problem, by transforming the problem into one of finding the best path on a weighted graph. The artificial ants incrementally build solutions by moving on the graph, randomly constructing solutions determined by a set of graph nodes and edges the values of which are modified at runtime by the ants. This construction process corresponds to the pheromone model deployed by real ants. The application of the ACO technique may best be explained by following a worked example (Dorigo, 2007).

ACO associates the set of cities with a set of vertices of a graph. Given that it is possible for the salesman to move from any city to any other city, the graph is fully connected and therefore the number of vertices is equal to the number of cities. The lengths of the edges between the vertices are set in proportion to the distances between the cities and pheromone values and heuristic values are associated with the edges of

the graph. Pheromone values are modified at runtime representing the cumulated experience of the ant colony, and heuristic values are set in line with the problem itself. In the case of the traveling salesman problem the heuristic values are set to be the inverse of the lengths of the edges.

Each ant starts from a randomly selected city (vertex of graph) moving along the edges of the graph and keeping a memory of its path. In subsequent steps the ant only follows edges that do not lead to already visited vertices. The ant has constructed a solution once it has visited all graph vertices. At each step, the ant probabilistically (using the pheromone values and heuristics) chooses the edge to follow among those that lead to yet unvisited vertices. The probability that the ant will choose a particular edge is determined in line with the higher the pheromone and the heuristic value associated to that edge. When all the ants have completed their tour, the pheromone on each edge is updated, usually by a defined percentage. Each edge is then given additional pheromone proportional to the quality of the solutions to which it belongs (there is one solution per ant). This procedure is repeatedly applied until a termination criterion is satisfied. A description of the mathematics of the ACO algorithm is available here (Dorigo, 2007).

The pros of ACO are its guaranteed convergence and its adaptability to the introduction of new instances. The cons are that probability distributions can change for each iteration and their time to convergence is uncertain.

7.2.12 ANOVA

ANOVA (Analysis of Variance) investigates whether there are any statistical differences between the means of groups of independent variables. ANOVA returns the probability (p-value) of obtaining the data assuming the null hypothesis (see Section 7.2.13 for the explanation of the null and alternative hypotheses in the case of student attribute analysis). A significant p-value (conventionally $p < 0.05$) suggests that at least one group mean is significantly different from the others. A more detailed description of the application of KNN analysis and its underlying mathematics is available here (Hindle, 2016).

7.2.13 Chi-square test

A traditional statistical method which may be applied to identify potential relationships between nominal attributes is to create a contingency table of observed and expected outcomes and use this data to apply the chi-square test to establish for potential relationships (Gajawada, 2019).

The first step in this method is to declare each of the null and alternative hypotheses for the analysis. In the case of the comparison of nominal attributes there are:

Null hypothesis: There is no association between two observed nominal attributes (they appear independent of each other).

Alternative hypothesis: There appears to be an association between the two observed nominal attributes.

For each pair of nominal attributes a contingency table (sometimes referred to as a frequency table) is created. Each cell is the count of the number of times that each permutation of the two attributes occurs in the data. This table represents the "observed" data.

Using the sums of each row and column an "expected" data table is generated. The expected value for each cell is calculated by multiplying the row total by the column total, then dividing by the grand total.

For each pairing the chi-square test is applied to each pair of nominal attributes. This test returns a p-value (Lee, 2019) and a chi-square value (Gajawada, 2019) for each pair The p-value is the probability of obtaining the observed data results of a test, assuming that the null hypothesis is correct.

A value of 0.05 is conventionally used as the cut-off for significance of the p-value. If the p-value is less than 0.05, the null hypothesis is rejected and it may be concluded that there is likely to be an association between the two nominal attributes.

The chi-square value is calculated as:

$$\chi^2 = \sum (\text{Observed value}_i - \text{Expected value}_i)^2 / \text{Expected value}_i$$

Each chi-square value is then compared with the chi-square critical value defined in a look-up table tabulated by p-values and respective degrees of freedom (degree of freedom is calculated as (table rows - 1) x (table columns - 1)). If the chi-square value is greater than the critical value then the null hypothesis is rejected and it may be concluded that there is likely to be an association between the two nominal attributes.

The pros of the chi-square test are its ability to analyse categorical data, robustness with respect to distribution of the data and its relative ease of computation. The cons are that the test can be highly sensitive to sample size, it only returns a yes or no answer to the question of a likely association between the two attributes and that it only tests two variables at one time.

7.3 Novel Technique for the Analysis of Nominal Data

A large variety of techniques are available to analyse numeric data, however there are fewer techniques applicable to nominal data. In each of the appropriate AI and ML technique sections 7.2.1 to 7.2.13, I note which are applicable to numeric and/or categoric data. Four of these (SVM, PCA, NN and KNN) are able to handle categoric data by encoding the data items into dummy numeric variables, others such as Decision Trees, Random Forest, Naïve Bayes and Chi-square are able to handle both numeric and categoric data directly.

A total of 17 of the 33 student attributes in the Portuguese secondary school student achievement dataset (see Section 6.2.3) are nominal (e.g. gender) and the remainder numerical (e.g. number of school absences). While considering alternative techniques to analyse this dataset a novel method became apparent and after experimentation provided useful results.

My analyses of the Portuguese student dataset include both PCA/GNG analysis of the numeric attributes and the application of my novel method to the nominal data (Wakelam et al., 2016).

The method compares the correspondence between pairs of nominal data attributes, calculating a numerical value from all permutations of values of the possible values each attribute can take. From these numerical values a symmetry (correlation) matrix is generated allowing the inference of relative strengths of each attribute to all other attributes to be determined. In particular, this method is able to provide both the correlations between categorical co-variables and the generation of a symmetry matrix, which may be used as a correlation matrix for use in PCA. As with the analysis of numerical variables, the resulting PCA allows the development of scatter plots and exploration for potential data clusters.

To illustrate the technique, Table 7.2 presents a worked example of a dataset of 4 students, each with 2 nominal attributes.

Table 7.2: Example Dataset (Wakelam et al., 2016)

Student	**Attribute 1 (a1)**	**Attribute 2 (a2)**
s1	p	x
s2	p	y
s3	q	z
s4	p	y

After setting a counter to zero we compare every possible pairing of student attribute values in the attribute 1 column of Table 7.3 with the corresponding pair in the attribute 2 column. If the selected pair from attribute 1 have the same value and the corresponding pair from attribute 2 also have the same value then we increment the counter by 1. Similarly if they both have different values then we increment the counter by 1. Otherwise we decrement the counter by 1 (Table 7.3).

So, for example, looking at step 1 below, the values of attribute 1 are both “p” (i.e. the same), whereas the values of attribute 2 are “x” and “y” (i.e. different), so we decrement the counter by 1. However, looking

at step 2, the values of attribute 1 are “p” and “q” (different), and the values of attribute 2 are “x” and “z” (different), so we increment the counter by 1.

Table 7.3: Step by Step Process (Wakelam et al., 2016)

Step	**Student pairing**	**a1**	**a2**	**Score**	**Cumulative counter**
1	(s1 s2)	(p p)	(x y)	-1	-1
2	(s1 s3)	(p q)	(x z)	+1	0
3	(s1 s4)	(p p)	(x y)	-1	-1
4	(s2 s3)	(p q)	(y z)	+1	0
5	(s2 s4)	(p p)	(y y)	+1	1
6	(s3 s4)	(q p)	(z y)	+1	2

This process is repeated for all combinations of attribute values and the resultant counter totals are used to populate a correlation matrix. Obviously, each attribute fully correlates with itself resulting in identical values across the matrix diagonal. The resulting matrix is normalised by dividing all entries by this value to keep all correlation matrix values between -1 and +1 (Table 7.4).

Table 7.4: Normalised Correlation Matrix for Illustrative Example 1 (Wakelam et al., 2016)

	a1	**a2**
a1	1	1/3
a2	1/3	1

Positive values represent positive correlations between the respective attributes, negative values represent negative correlations and the magnitude of the value represents the strength of the correlation.

For example, where there are a high proportion of data pairs where the corresponding attributes are correspondingly the same or different this will result in a relatively higher correlation value (for example, 1/3 in Table V) between the two attributes.

For each attribute, its correlation with all other attributes is evaluated and mean value calculated over all these correlations. As a first indicator of interesting attributes, particular attention was paid to those

correlations where the magnitude of the mean value was high in comparison to the mean values of other attributes. Those correlations where the magnitude was above the mean for that attribute then provided additional correlations for consideration.

The technique was applied to each of the Mathematics and Portuguese Language datasets in turn (see Section 7.4.3). For each dataset those pairs of attributes that were most strongly correlated were identified – whether positively or negatively. This enabled the potential influences on student behaviours to be considered.

The correlations in the Mathematics data set were then compared with those in the Portuguese Language dataset.

Using the correlation matrix generated by this technique corresponding PC1 v PC2 scatter plots were produced for each of our Mathematics and Portuguese Language student datasets in order to visualize potential clusters for future analysis and comparison with any clusters identified in our numeric data. In order to visualize and more easily identify potential clusters PCA scatter plot was produced for each of the four final grade intervals (using final grades 0-5, 6-10, 11-15, 16-20 as the labels) for each student dataset.

7.4 Techniques Applied to Each Dataset

7.4.1 Small Student Dataset for Higher Education Teachers

7.4.1.1 Technique(s) Applied

Support Vector Machine (SVM) classification. This technique was selected given its accuracy and good performance on smaller, cleaner datasets and that because it uses a subset of training points it can be more efficient than other techniques.

7.4.1.2 Dataset

The work of Natek & Zwilling investigates the application of data mining techniques to small datasets see Table 6.2 above (Natek & Zwilling 2014).

7.4.1.3 Experimental Analysis and Results

For this dataset, after scaling the data to have unit standard deviation, support vector machine techniques were applied using svm-toy from the libsvm tool box (Chang & Lin, 2011), with input classes Activities Points and Exam Points to visualise the data and the decision boundary between the two classes (Figure 1). The axes are Activities points (x-axis) and Exam points (y-axis). Kernel type RBS (Radial Basis Function) was selected.

The data was split into two halves: Class 1, where Final Points ≥75 and Class 2, where Final Points <75. Students who have achieved the highest grades are represented as purple squares and those who achieved lowest as blue squares. As can be seen in Figure 7.15 svm_toy was able to make a very clear delineation between the two classes (Class 1 is shaded black and Class 2 blue). This allows the prediction of the likely Final Points class of any new student whose Activity Point and Exam Points we are presented with.

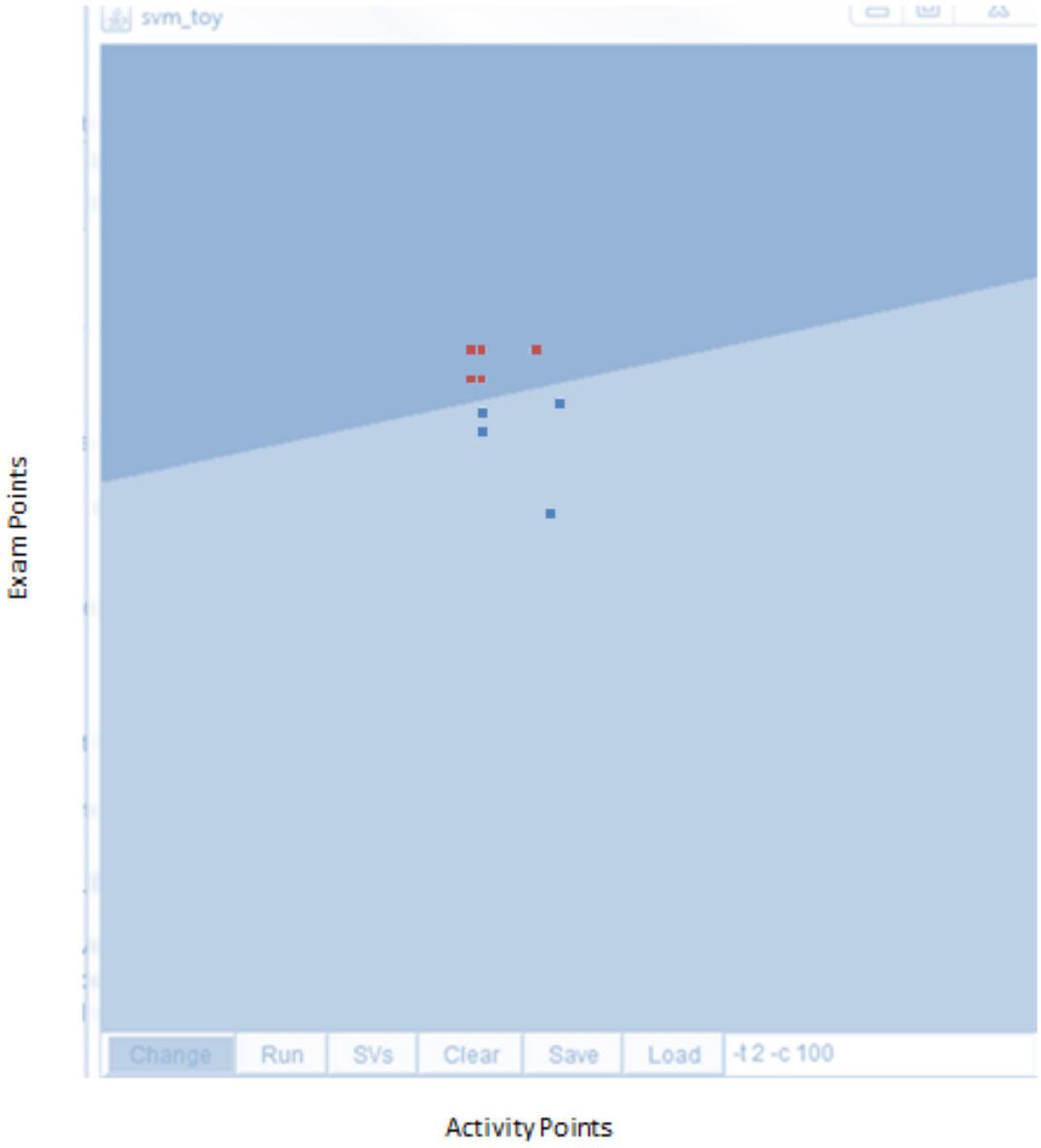


Figure 7.15: Small Student Dataset: Svm-toy: Exam Points & Activity Points for all Student Data

Please note that the 10 data points appear as 9 on the plot because two of the points are identical.

The base data was then divided into a training set (8 data points) and a test set (2 data points) and svm_train followed by svm_predict was run, using SVM classification. This obtained results of 100% accuracy.

7.4.1.4 Conclusions

In this case, of a very small dataset and only two attributes, SVM was able to clearly delineate between the two classes, allowing predictions of the likely Final Points class of any new student whose Activity Point and Exam Points it would be presented with.

7.4.2 Students' Knowledge Levels on DC Electrical Machines

7.4.2.1 Techniques(s) Applied

Data visualisation using scatter plots and Principal Component Analysis (PCA) classification. PCA was selected in this case because it removes correlated features, which in turn improves the performance of the algorithm, reduces the risk of overfitting and provides a visualisable output.

7.4.2.2 Dataset

This dataset was obtained from the research conducted into the creation of an efficient user knowledge model for adaptive learning systems (Kahraman et al., 2013) and in particular the University College Irvine (UCI) Machine Learning Repository availability of the dataset used.

The data comprises 258 students' performance in an on-line web based Electrical Engineering course. Data was measured against 5 attributes (see Table 6.3):

7.4.2.3 Experimental Analysis and Results

For this dataset, after scaling the data to have unit standard deviation Principal Component Analysis (PCA) was applied to plot the Exam performance (UNS) data classified as "Very Low", "Low", "Middle:, "High" (Figure 7.16) in order to visualise the data and look for any obvious patterns. This figure plots the first two principal components which account for 54% of the variance of the data. As can be seen, there is a great deal of overlap of the different knowledge student grades consequently making it impossible to identify boundaries between every set of grades. However, it is possible to see a clear boundary between Highest (green) and the Very Low (red) exam performance.

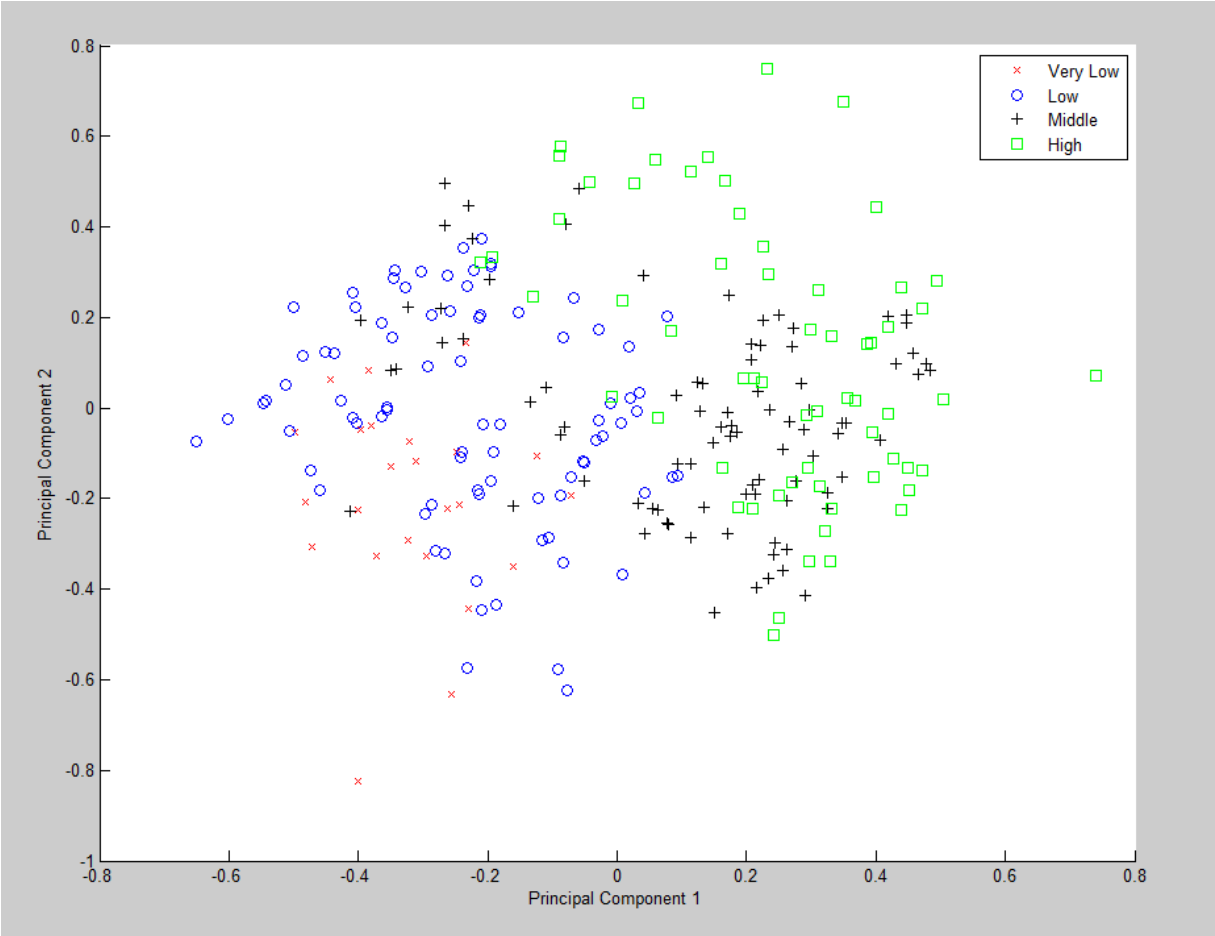


Figure 7.16: Exam Performance (UNS) Data Classified as “Very Low”, “Low”, “Middle", “High”

Pairs of attributes were then selected from the five attributes above that looked as if they might show correlations with each other and the results were plotted.

In the case of Degree of Study Time v Exam Performance., the plot (Figure 7.17) shows exactly the sort of patterns you would expect for the resulting knowledge level of students who spend more time studying. As can be seen there are a small number of outliers in the plotted results (data points that lie away from the majority), which can perhaps be explained by the cleverer students getting away with less study time.

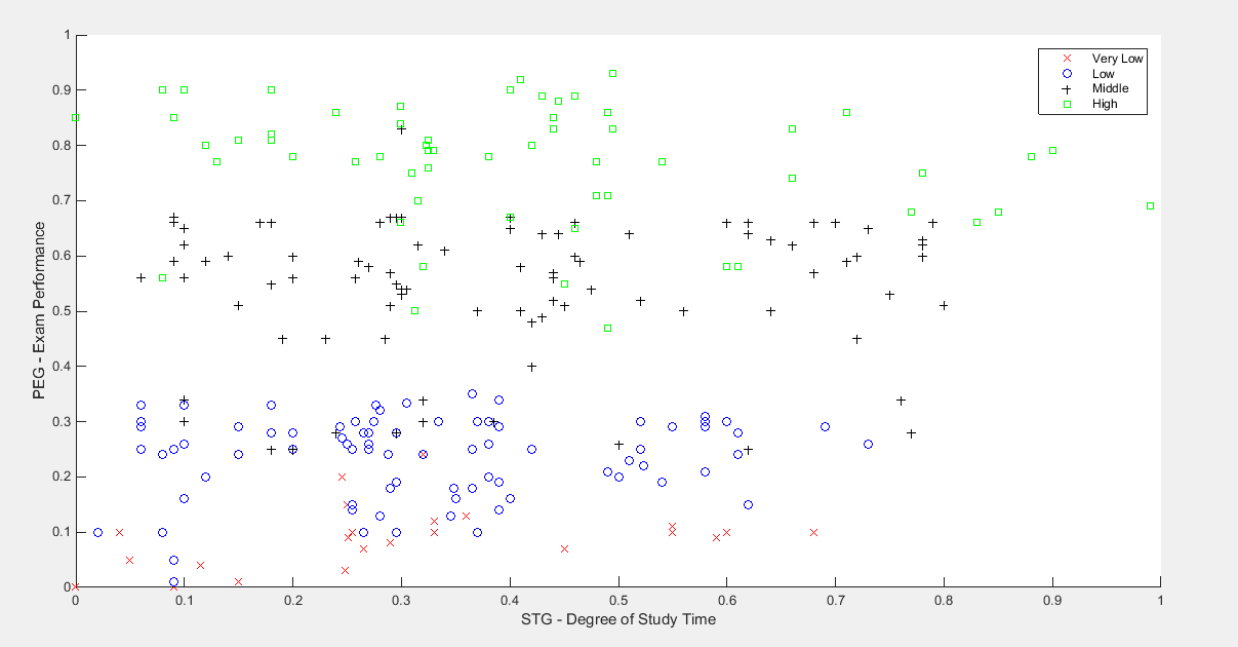


Figure 7.17: The Degree of Study Time v Exam Performance

Degree of Study Time v Exam Performance for Related Objects (related objects are the non-core, but related areas of study): Figure 7.18 shows an almost random correlation between these attributes and so no relationships between them were identified.

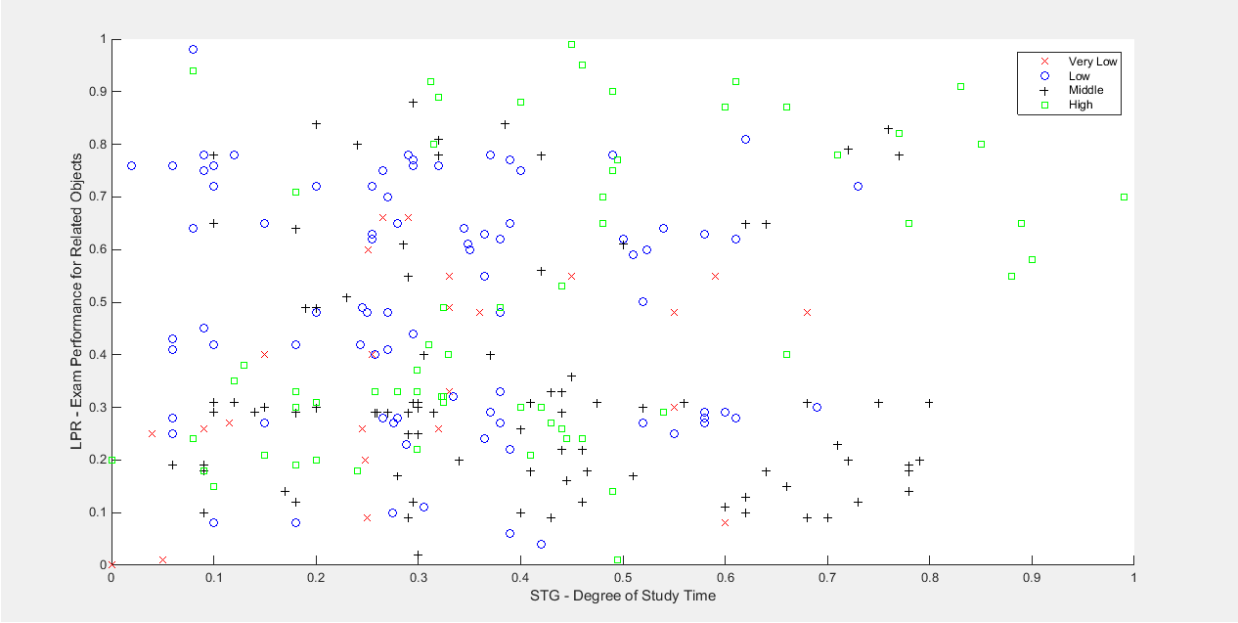


Figure 7.18: Degree of Study Time v Exam Performance for Related Objects

Exam Performance for Related Objects v Exam Performance: This plot (Figure 7.19) shows a very clear correlation between Full exam performance and Exam performance for related areas. This result can perhaps be seen to confirm that students who invest time in gaining an understanding of non-core, but related areas of study which deepen their understanding of the subject area do better in their exams.

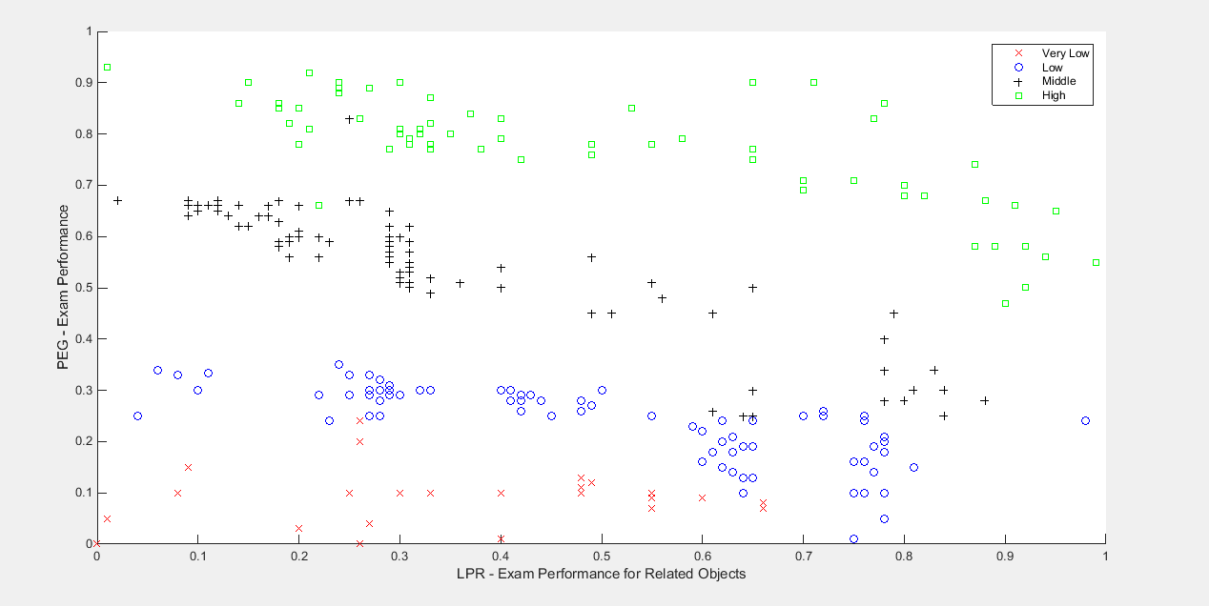


Figure 7.19: Exam Performance for Related Objects v Exam Performance

7.4.2.4 Conclusions

The application of PCA to a modestly sized dataset (258 students) with a small number of attributes (5) resulted in mixed results, in some cases correlations were evident, in others none.

7.4.3 Portuguese Secondary School Student Achievement

7.4.3.1 Technique(s) Applied

In the case of the numeric data, Principal Component Analysis (PCA) to reduce the dimensionality of the data followed by Growing Neural Gas (GNG) to identify potentially useful clusters of data. GNG was selected given the technique's ability to find optimal clusters in data without prior information about the number of optimal clusters. In the case of nominal data the novel technique was applied (see Section 7.3).

7.4.3.2 Dataset

In order to investigate the predictive accuracy of student achievement data was taken from a set of students from a Portuguese study (Cortez & Silva, 2008). This data consists of information taken from two

Portuguese secondary schools and each student has 33 attributes. The data includes three labels: first period grade, second period grade and final grade. The subjects are Mathematics (395 students) and Portuguese Language (649 students) and the data was collected during the 2005-2006 academic year. The attributes comprise 16 numeric (including the labels: first period, second period and final performance grades) and 17 nominal (Tables 7.5 and 7.6).

Table 7.5: Example of the Numeric Attributes (Cortez & Silva, 2008)

Identifier	**Description**
Age	Student's age (numeric: from 15 to 22)
Absences	Number of school absences (numeric: from 0 to 93)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

Table 7.6: Examples of the nominal attributes (Cortez & Silva, 2008)

Identifier	**Description**
Gender	Student's gender (binary: "F" - female or "M" - male)
Mjob	Mother's job (nominal: "teacher", "health" care related, civil "services" (e.g., admin or police), "at_home" or "other")
Romantic	With a romantic relationship (binary: yes or no)

For consistency the original attribute types as used in the Portuguese study were adopted, although there are a small number of the attributes defined as numeric which could be considered as ordinal.

7.4.3.3 Experimental Analysis

Analysis of nominal data

The novel method compares the correspondence between pairs of our nominal data attributes (see Section 7.3).

The technique was applied to each of the Mathematics and Portuguese Language datasets in turn. For each dataset, it was then possible to identify those pairs of attributes that were most strongly correlated – whether positively or negatively. This enabled the potential influences on student behaviours to be considered.

In addition the correlations between the Mathematics dataset and those in the Portuguese Language dataset were compared.

Using the correlation matrix generated by this technique the corresponding PC1 v PC2 scatter plots for each of the Mathematics and Portuguese Language student datasets were produced. The potential clusters for future analysis and comparison with any clusters identified in our numeric data may then be examined. In order to visualize and more easily identify potential clusters a PCA scatter plot for each of the four final grade intervals (using final grades 0-5, 6-10, 11-15, 16-20 as our labels) was produced for each student dataset.

Analysis of measurement data

After normalisation of the Mathematics and Portuguese Language student numeric datasets, respectively (by subtracting the mean and dividing by the standard deviation) a linear Principal Component Analysis (PCA) was performed, plotting each of the leading three principle components, PC1 v PC2, PC2 v PC3, PC1 v PC3. In each Figure, the amount of variance accounted for by the respective principal components is reported. For example, in Figure 1 PC1 and PC2 account for 26% of the total information in the data.

In each case a visual inspection suggested possible clusters. In order to try and identify these clusters GNG was applied, with key parameters set to 50 training runs and a maximum of 200 nodes. This technique [16] identified a small number of clusters and their respective centroids as well as allowing us to identify the actual students in each cluster.

7.4.3.4 Results

The aim was to identify interesting correlations in our student data attributes, providing the opportunity to focus on promising correlations for deeper analysis.

Nominal data

Mathematics students

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables 7.7 and 7.8 respectively.

Table 7.7: Highest mean value Mathematics Student attributes (Wakelam et al., 2016)

Attribute	Mean value
Higher Education wish	0.23
School	0.19
Parent cohabitation	0.18

Table 7.8: Lowest mean value Mathematics Student attributes (Wakelam et al., 2016)

Attribute	Mean value
Paid tutor	0.008
Gender	0.006
Extra-curricular activity	0.003

The results show potential correlations may exist between the student's wish to take Higher Education and other nominal attributes - the school attended and parent cohabitation status, followed by receipt of extra educational support, Mother's job, access to the internet, the reason for choice of school and nursery school attendance.

Mother's job also shows potential correlations with other factors, including the wish for Higher Education, parent cohabitation, school attended, educational support and choice of school.

Paid extra tuition does not correlate strongly with other factors, even parent's jobs, which might have been expected. This is also true for students receiving educational support from within the family. However, future analyses may show that such extra tuition correlates with student performance measured by their grades.

Internet access also shows potential correlations with a number of factors, including the wish for Higher Education, school attended, parent cohabitation, address, the level of educational support by the school and Mother's job.

Factors which show very low correlations with others are the level of extra-curricular activities, whether the student was male or female and paid tutoring, followed by romantic relationships, Father's job, and family size.

Portuguese Language students

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables 7.9 and 7.10 respectively.

Table 7.9: Highest Mean value Portuguese Language Student attributes (Wakelam et al., 2016)

Attribute	Mean value
Paid tutor	0.20
Higher Education wish	0.18
Parent cohabitation	0.16

Table 7.10: Lowest Mean Value Portuguese Language Student attributes (Wakelam et al., 2016)

Attribute	Mean value
Family education support	0.02
Gender	0.01
Extra-curricular activity	0.003

The results show potential correlations may exist between paid tutoring, the student's wish to take Higher Education and parent cohabitation followed by educational support and Mother's job.

Paid extra tuition shows potential correlations with a number of other factors including the level of educational support, the wish for Higher Education, parent cohabitation, and Mother's job. This is also true for extra educational support provided by the school, correlating with the use of paid tutors, parent cohabitation, and Mother's job.

Mother's job shows potential correlation with the use of paid tutoring, educational support, parent cohabitation and attendance at a nursery school.

Internet access only correlated modestly with other factors for Portuguese Language students.

Factors which show very low correlations with others are the level of extra-curricular activities, student gender and family educational support, followed by romantic interest, guardian, Father's job and school attended.

Comparisons between Mathematics and Portuguese Language analysis results

The wish to take Higher Education shows potential correlation with Mother's job, cohabitation status and receipt of extra educational support for both sets of students.

In both cases Mother's job correlates with other factors. In contrast, Father's job, along with romantic relationships and extra-curricular activities shows very low correlations with other factors in both sets.

Additional educational support provided by the school also shows potential correlation with a number of other factors in both sets.

In comparison with Portuguese Language students, paid extra tuition in the case of Mathematics students does not correlate strongly with other factors.

Interestingly, gender, often considered to be an influential factor, does not correlate well with other attributes in either set.

In the case of Mathematics students, internet access shows potential correlations with a number of factors, such as the wish to take further education, school attended, and parent cohabitation. However, in the case of Portuguese Language students, internet access shows only modest correlations.

Principal Component Analysis

As described in section 7.2.2, above, a PCA projection will allow visualization of multi-dimensional data in a two-dimensional representation. For each dataset the initial PCA plot including all final grades proved too challenging to visualize four plots were produced, one for each of the four final grade intervals. One example from each dataset is included. Principle component analysis of the Mathematics and Portuguese Language student data shows no evidence of potential clustering.

For example, a PC1 v PC2 nominal data plot of Mathematics students' achieving final grades of between 11 and 15 (Figure 7.20).

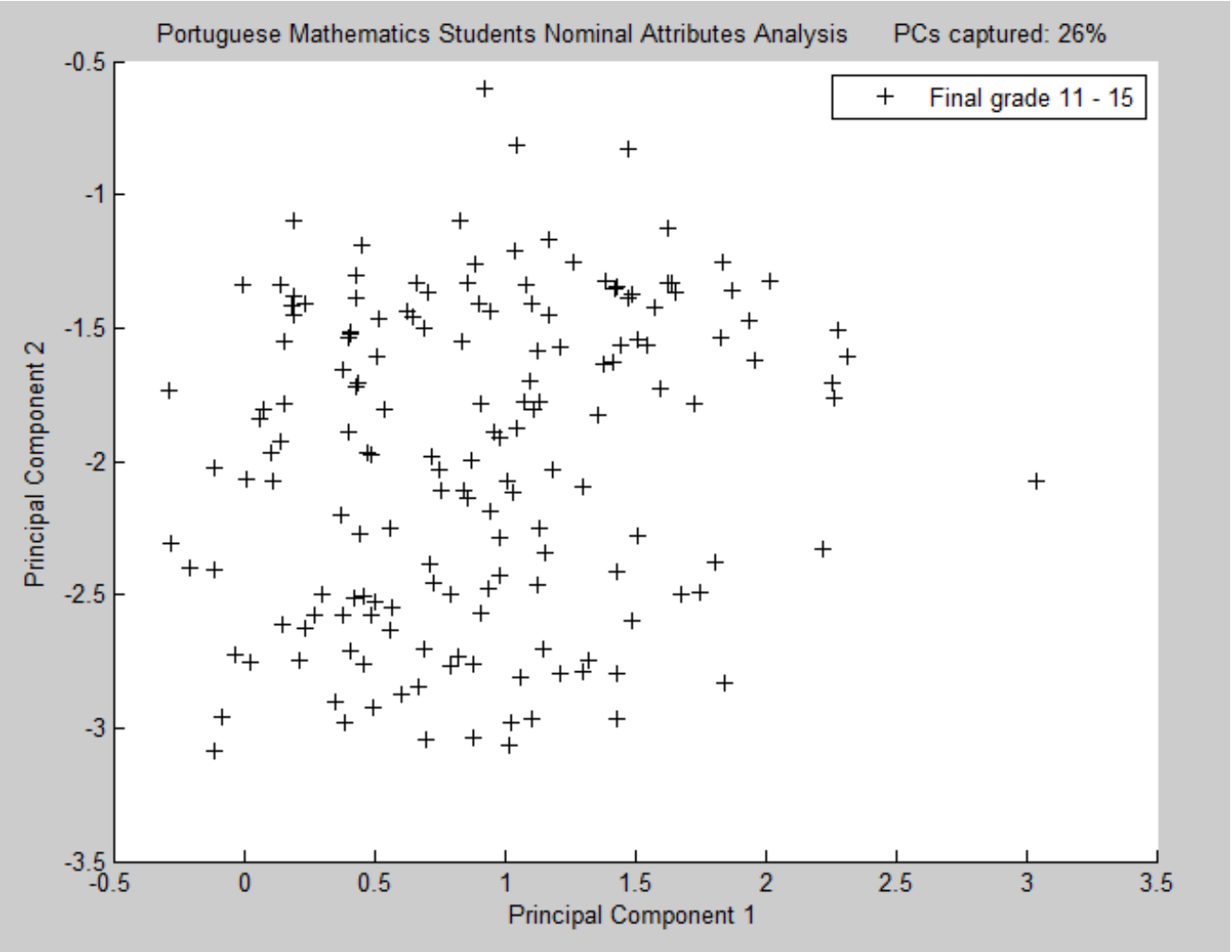


Figure 7.20: Mathematics Nominal Data PC1 v PC2 Final Grades 11-15 (Wakelam et al., 2016)

A further example shows a PC1 v PC2 nominal data plot of Portuguese Language students' achieving grades of between 11 and 15 (Figure 7.21). This data plot appears to exhibit a lower boundary delineation which is believed to be a result of a predominance of very narrow variances in the attribute values in this particular dataset.

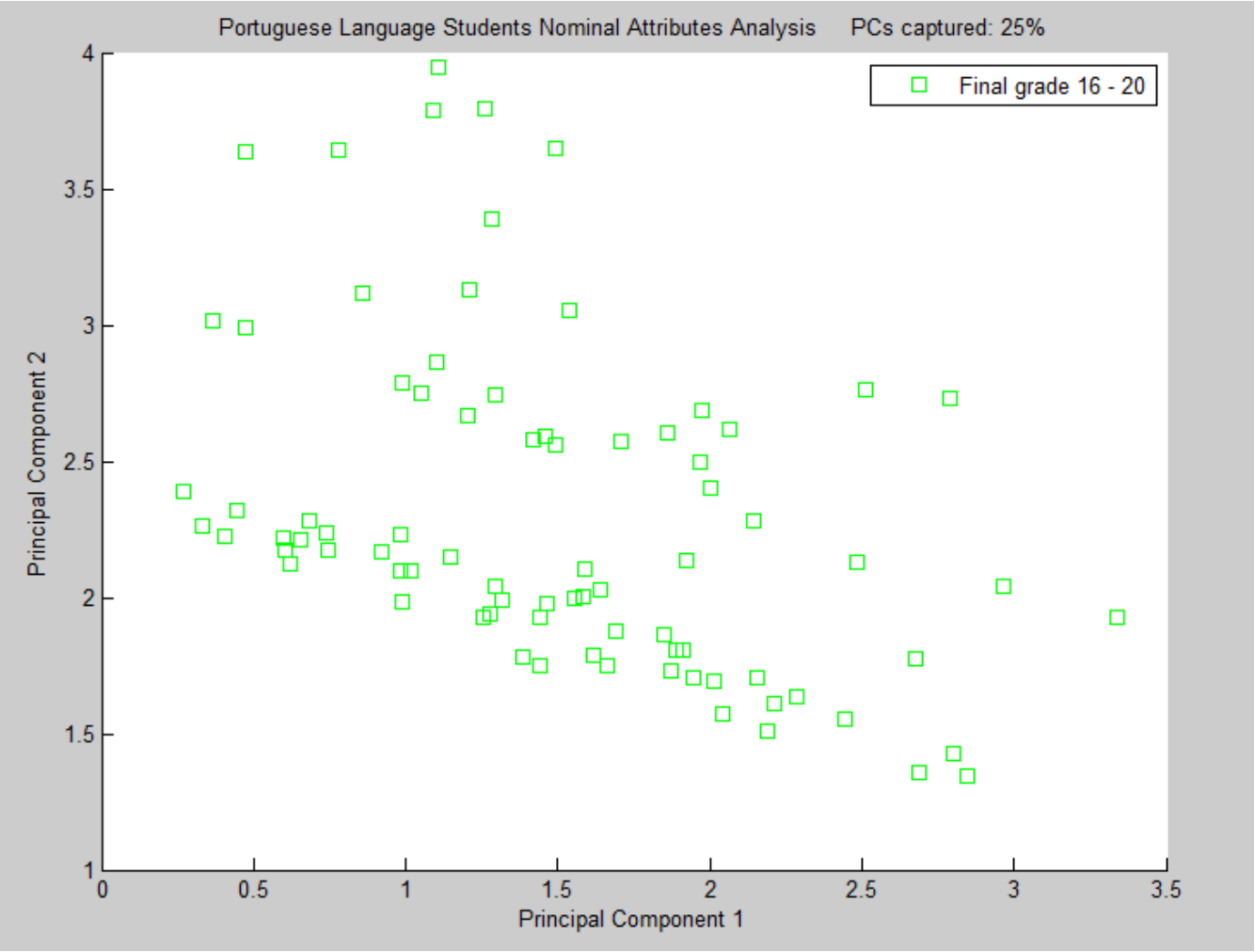


Figure 7.21: Portuguese Language Nominal Data PC1 v PC2 Final Grades 16-20 (Wakelam et al., 2016)

Measurement data

Mathematics students

GNG identified modest clustering in each of the PC1, PC2, PC3 comparisons For example, in Figure 7.22 we can see that three clusters have been identified. The centroids are shown in red and in each case the students in each cluster are identified in order to for look for potential correlations with the results of our nominal data analysis.

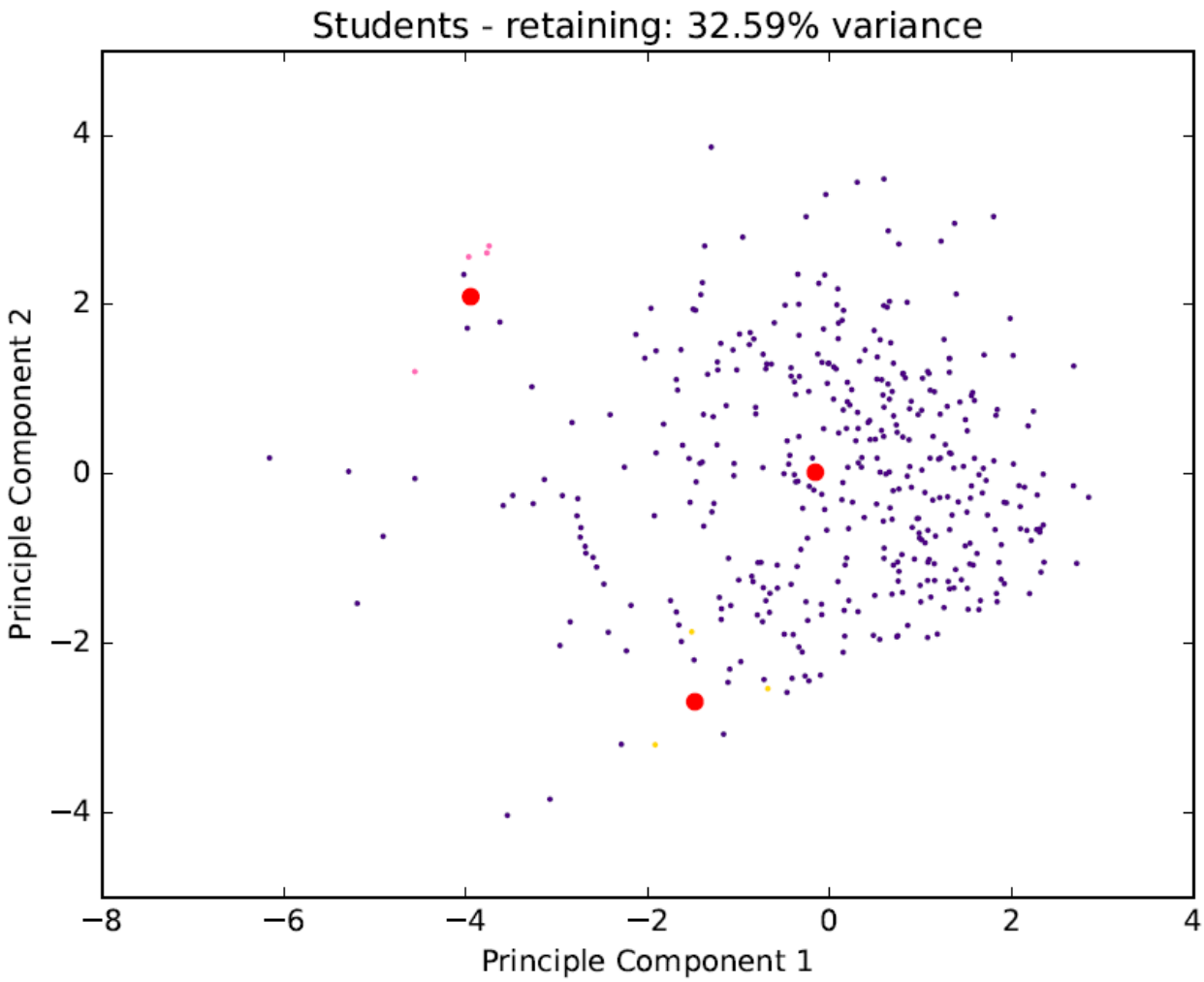


Figure 7.22: Mathematics Students' Numeric Data PC1 v PC2 Scatter Plot (Wakelam et al., 2016)

Portuguese Language students

GNG did not identify useful clustering in either of the PC1, PC2, PC3 comparisons in any of the four final grade intervals. In all cases only one cluster was identified, for example, in Figure 7.23. As above, the centroids are shown in red.

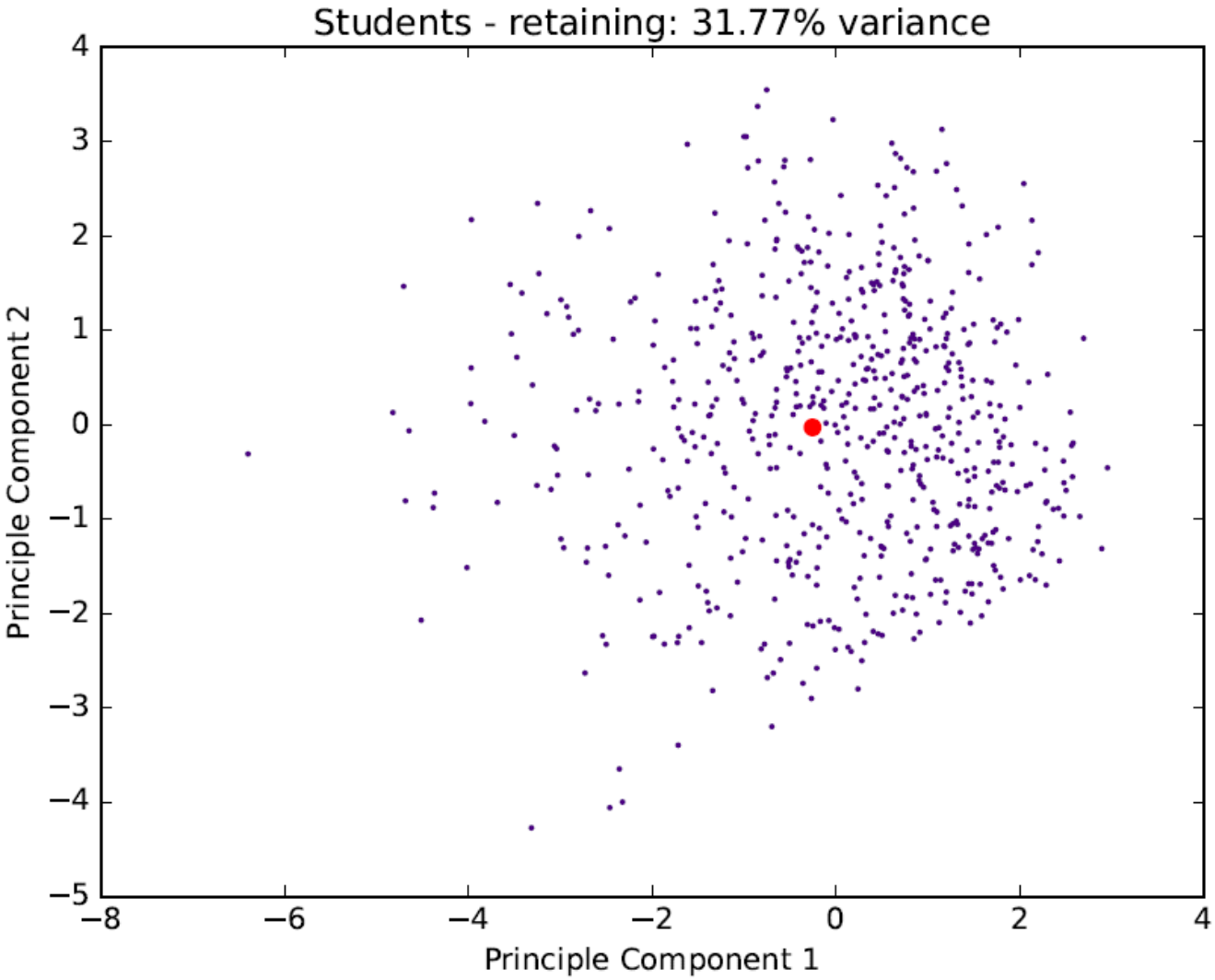


Figure 7.23: Portuguese Lang. Students' Numeric Data PC1 v PC2 Scatter Plot (Wakelam et al., 2016)

The GNG analysis was repeated, adjusting the key parameters, increasing the number of training runs from 50 to 100 and maximum nodes from 200 to 600. However, this did not result in improvement. Further work is underway to identify alternative techniques to identify potential clustering in the Portuguese Language student numeric data, such as Curvilinear Component Analysis (CCA).

The full set of PCA scatter plots generated are included in Appendix E.

7.4.3.5 Comparison of results of novel technique for the analysis of nominal data with those of a chi-square test analysis

The chi-square Test Analysis was conducted as follows:

The null and alternative hypotheses appropriate to the correlation of Portuguese student nominal attributes are:

Null hypothesis: There is no association between two observed nominal attributes (they appear independent of each other).

Alternative hypothesis: There appears to be an association between the two observed nominal attributes.

The chi-square test analysis was applied to each of the Mathematics and Portuguese Language datasets. The resulting p-values are presented in Tables 7.11 to 7.14 respectively. The resulting chi-square values tables are included in Appendix F.

Mathematics students

The results of chi-square analysis suggest that family education support, with 8 p-values < 0.05, is the attribute which has a relationship with the most other student attributes, followed by gender, mother's job and paid tutor, each with 7 p-values <0.05 (Tables 7.11 and 7.12). For example, the analysis of chi-square values for family education support show correlations with the school attended, mother's job, extra school support, family size and gender (Appendix F).

The results of chi-square analysis (Appendix F) suggest that the attributes which have a relationship with the least other attributes are parent cohabitation status and extra-curricular activities, each with one p-value < 0.05, followed by father's job (2 p-values < 0.05).

Portuguese Language Students

The results of chi-square analysis suggest that mother's job and school, each with 9 p-values < 0.05, are the attributes which have a relationship with the most other student attributes, followed by gender (8 p-values < 0.05) and Higher Education wish (7 p-values < 0.05) (Table 7.13 and 7.14). For example, the analysis of chi-square values for mother's job shows correlations with the school attended, father's job, extra-curricular activities, internet access, guardian, reason for school choice, address and gender (Appendix F).

The results of chi-square analysis (Appendix F) suggest that the attributes which have a relationship with the least other attributes are paid tutor and nursery school attendance, each with two p-values < 0.05, followed by family size, Father's job and family support each with three p-values < 0.05.

Table 7.11: Mathematics Student Attribute P-values

p-values	School	Gender	Address	Famsize	Pstatus	Mjob	Fjob	Reason	Guardian
School		0.8070989	0.00000003	0.19733506	0.36139667	0.39121104	0.07761918	0.00594783	0.31178571
Gender	0.8070989		0.57104719	0.07410444	0.64127556	0.00155644	0.31417349	0.17629237	0.34542853
Address	0.00000003	0.57104719		0.14976690	0.39749659	0.01013724	0.71166025	0.02180446	0.26251683
Famsize	0.19733506	0.07410444	0.14976690		0.00294452	0.47855952	0.49629825	0.93385862	0.83074055
Pstatus	0.36139667	0.64127556	0.39749659	0.00294452		0.50193845	0.32142913	0.91329340	0.08702398
Mjob	0.39121104	0.00155644	0.01013724	0.47855952	0.50193845		0.00000000	0.02818547	0.09284299
Fjob	0.07761918	0.31417349	0.71166025	0.49629825	0.32142913	0.00000000		0.19858757	0.02397069
Reason	0.00594783	0.17629237	0.02180446	0.93385862	0.91329340	0.02818547	0.19858757		0.60797853
Guardian	0.31178571	0.34542853	0.26251683	0.83074055	0.08702398	0.09284299	0.02397069	0.60797853	
Schoolsup	0.00546533	0.00599452	0.62332688	0.56919060	0.40121061	0.19959633	0.20029449	0.95765939	0.40163039
Famsup	0.00104303	0.00258304	0.63473800	0.02485137	0.70473021	0.04802701	0.14840630	0.08318220	0.96099677
Paid	0.73421510	0.01027837	0.29400667	0.78262295	0.35606875	0.01211452	0.47542422	0.00720650	0.40681031
Activities	0.02011224	0.04723966	0.30736695	0.99820529	0.05301016	0.12492712	0.70563527	0.07351560	0.69861101
Nursery	0.07600742	0.87049750	0.23629434	0.04246156	0.07171418	0.05395990	0.20486084	0.64441978	0.00240833
Higher	0.63124239	0.00268063	0.39437999	0.90812828	0.41817784	0.06500684	0.45736001	0.01665203	0.91950166
Internet	0.00793531	0.38063515	0.00001635	0.98857643	0.16371259	0.00000834	0.36888152	0.48522090	0.49633741
Romantic	0.22766657	0.04259422	0.91678540	0.49423824	0.42142910	0.67038577	0.69010915	0.38509486	0.04438195

Key: p-value < 0.05

Table 7.12: Mathematics Student Attribute P-values (Continued)

p-values	Schoolsup	Famsup	Paid	Activities	Nursery	Higher	Internet	Romantic
School	0.00546533	0.00104303	0.73421510	0.02011224	0.07600742	0.63124239	0.00793531	0.22766657
Gender	0.00599452	0.00258304	0.01027837	0.04723966	0.87049750	0.00268063	0.38063515	0.04259422
Address	0.62332688	0.63473800	0.29400667	0.30736695	0.23629434	0.39437999	0.00001635	0.91678540
Famsize	0.56919060	0.02485137	0.78262295	0.99820529	0.04246156	0.90812828	0.98857643	0.49423824
Pstatus	0.40121061	0.70473021	0.35606875	0.05301016	0.07171418	0.41817784	0.16371259	0.42142910
Mjob	0.19959633	0.04802701	0.01211452	0.12492712	0.05395990	0.06500684	0.00000834	0.67038577
Fjob	0.20029449	0.14840630	0.47542422	0.70563527	0.20486084	0.45736001	0.36888152	0.69010915
Reason	0.95765939	0.08318220	0.00720650	0.07351560	0.64441978	0.01665203	0.48522090	0.38509486
Guardian	0.40163039	0.96099677	0.40681031	0.69861101	0.00240833	0.91950166	0.49633741	0.04438195
Schoolsup		0.03748016	0.67999986	0.36025718	0.36093808	0.27885805	0.84738688	0.10866980
Famsup	0.03748016		0.00000001	0.97621535	0.23671024	0.04510661	0.03952930	0.80472406
Paid	0.67999986	0.00000001		0.67086123	0.04235059	0.00016954	0.00233898	0.91239112
Activities	0.36025718	0.97621535	0.67086123		0.95671697	0.05516458	0.33346999	0.69613224
Nursery	0.36093808	0.23671024	0.04235059	0.95671697		0.28047653	0.87633999	0.58475838
Higher	0.27885805	0.04510661	0.00016954	0.05516458	0.28047653		0.68553486	0.03572537
Internet	0.84738688	0.03952930	0.00233898	0.33346999	0.87633999	0.68553486		0.08336087
Romantic	0.10866980	0.80472406	0.91239112	0.69613224	0.58475838	0.03572537	0.08336087	

Key: Chi-square value > critical chi-square value for respective degrees of freedom

Table 7.13: Portuguese Language Student Attribute P-values

p-values	School	Gender	Address	Famsize	Pstatus	Mjob	Fjob	Reason	Guardian
School		0.0343680	0.00000000	0.57079370	0.47375951	0.00000012	0.00026792	0.00000000	0.23431975
Gender	0.0343680		0.51589138	0.01235582	0.09929949	0.00107948	0.35677600	0.29496891	0.60510655
Address	0.00000000	0.51589138		0.24009472	0.01591400	0.00001789	0.25055980	0.00024172	0.66850773
Famsize	0.57079370	0.01235582	0.24009472		0.00000000	0.63098248	0.29547179	0.40572510	0.87816238
Pstatus	0.47375951	0.09929949	0.01591400	0.00000000		0.62654921	0.18523659	0.36580657	0.00007664
Mjob	0.00000012	0.00107948	0.00001789	0.63098248	0.62654921		0.00000000	0.00309600	0.00238143
Fjob	0.00026792	0.35677600	0.25055980	0.29547179	0.18523659	0.00000000		0.06504276	0.00733321
Reason	0.00000000	0.29496891	0.00024172	0.40572510	0.36580657	0.00309600	0.06504276		0.46267079
Guardian	0.23431975	0.60510655	0.66850773	0.87816238	0.00007664	0.00238143	0.00733321	0.46267079	
Schoolsup	0.00167722	0.00461227	0.64736014	0.15073377	0.80963852	0.21825133	0.09919016	0.32427720	0.57219306
Famsup	0.10452559	0.00097298	0.88701826	0.31038981	0.79492938	0.09723662	0.09838749	0.09124627	0.40618622
Paid	0.84039212	0.04336247	0.43741266	0.20046395	0.68499717	0.82597208	0.92193479	0.08933767	0.18923062
Activities	0.02399373	0.00148821	0.81315592	0.70634268	0.00967687	0.02522820	0.54586311	0.00032352	0.55953903
Nursery	0.90552033	0.26665390	0.64514394	0.01031654	0.40446813	0.10847089	0.57986399	0.64554581	0.03376647
Higher	0.00052527	0.13860974	0.05068655	0.90827453	0.56261090	0.00003393	0.06492705	0.01527282	0.00000294
Internet	0.00000000	0.09313031	0.00000752	0.73364343	0.12794336	0.00000000	0.08526982	0.00246504	0.76402102
Romantic	0.06571186	0.00501671	0.43058976	0.40143697	0.17028464	0.29500307	0.88092056	0.33875641	0.00266507

Key: p-value < 0.05 

Table 7.14: Portuguese Language Student Attribute P-values (Continued)

p-values	Schoolsup	Famsup	Paid	Activities	Nursery	Higher	Internet	Romantic
School	0.00167722	0.10452559	0.84039212	0.02399373	0.90552033	0.00052527	0.00000000	0.06571186
Gender	0.00461227	0.00097298	0.04336247	0.00148821	0.26665390	0.13860974	0.09313031	0.00501671
Address	0.64736014	0.88701826	0.43741266	0.81315592	0.64514394	0.05068655	0.00000752	0.43058976
Famsize	0.15073377	0.31038981	0.20046395	0.70634268	0.01031654	0.90827453	0.73364343	0.40143697
Pstatus	0.80963852	0.79492938	0.68499717	0.00967687	0.40446813	0.56261090	0.12794336	0.17028464
Mjob	0.21825133	0.09723662	0.82597208	0.02522820	0.10847089	0.00003393	0.00000000	0.29500307
Fjob	0.09919016	0.09838749	0.92193479	0.54586311	0.57986399	0.06492705	0.08526982	0.88092056
Reason	0.32427720	0.09124627	0.08933767	0.00032352	0.64554581	0.01527282	0.00246504	0.33875641
Guardian	0.57219306	0.40618622	0.18923062	0.55953903	0.03376647	0.00000294	0.76402102	0.00266507
Schoolsup		0.05474428	0.30204484	0.44098938	0.64938045	0.02966997	0.50868588	0.01627958
Famsup	0.05474428		0.01629382	0.84981677	0.47882240	0.02969844	0.06703121	0.55112223
Paid	0.30204484	0.01629382		0.09377563	0.48251130	0.53903672	0.41753061	0.64090771
Activities	0.44098938	0.84981677	0.09377563		0.31160236	0.25260308	0.03585763	0.14284897
Nursery	0.64938045	0.47882240	0.48251130	0.31160236		0.27775539	0.85527592	0.55818761
Higher	0.02966997	0.02969844	0.53903672	0.25260308	0.27775539		0.07312144	0.01134162
Internet	0.50868588	0.06703121	0.41753061	0.03585763	0.85527592	0.07312144		0.37488548
Romantic	0.01627958	0.55112223	0.64090771	0.14284897	0.55818761	0.01134162	0.37488548	

Key: Chi-square value > critical chi-square value for respective degrees of freedom

Comparison between Mathematics and Portuguese Language analysis results

In both Mathematics and Portuguese Language chi-square analyses mother's job and gender are related to a large number of other attributes. Father's job has the least relationships with other attributes in both student data sets. Each of paid tutor and family education support provided contradictory results figuring highly in the case of Mathematics students and low in Portuguese Language students.

The results of the comparison between the chi-square analysis and Novel method are as follows:

The results generated by the novel method showed modest correspondence with those generated by the chi-square analysis. In the case of the Portuguese Language students, Higher Education wish was identified as an important attribute in terms of its relationship to other attributes by both methods (see Table 4.9). For Portuguese Language students, family education support was identified as an attribute with the fewest relationships with other attributes by both methods (see Table 4.10), as was extra-curricular activities for mathematics students. There was also some cross correspondence between the two methods across the different student populations, for example the novel method also identified Higher Education wish as an important attribute for Mathematics students as identified by the chi-square method for Portuguese Language students. Some differences in method performance may be related to the difference in populations sizes between the two student data sets (Portuguese language student dataset of 649 almost two thirds (64%) larger than that of the mathematics students). In the case of the chi-square test, its limitation of sensitivity to sample size may be relevant and this may also be true for the novel method. Future work is recommended to explore this with varying dataset sizes, see Chapter Nine: Conclusions and Future Work (see Section 9.4.7).

7.4.3.6 Conclusions

A novel approach to the analysis of the nominal data has been applied, comparing the correspondence between pairs of nominal attributes.

An investigation of whether the analysis would identify interesting information in the dataset shows that to some extent it did. Our PCA plot of the Mathematics nominal data showed no evidence of clustering.

Numeric data analysis techniques were then applied to identify clustering and potential correlations in our numeric attributes identifying some potentially interesting patterns.

In the case of the Mathematics student data using PCA followed by the GNG technique some clustering of the data was identified, however the corresponding analysis of the Portuguese Language student data did not identify useful clusters.

A comparison of the results between the application of the novel technique for the analysis of nominal data and each of contingency table and chi-square test showed only very modest correspondences.

7.4.4 Open University Student Dataset

7.4.4.1 Technique(s) Applied

Naïve Bayesian classification, Classification and Regression Tree (CART), K-Nearest Neighbour (KNN).

7.4.4.2 Dataset

The OU is an excellent example of the use of very large datasets comprising in excess of 32,000 students across 22 courses and 28, mixed numeric and nominal, attributes per student (see Section 6.2.4). The potential for multi-year aggregation of module and student data to improve prediction accuracy is a powerful benefit in their approach.

7.4.4.3 Review

The diligence of the OU analytics team, working alongside their institutional privacy and ethics teams and students themselves, has resulted in their successfully overcoming the institutional barriers which limit progress for many HE organisations. In particular, the inclusion of student demographic data provides the ML techniques with valuable additional data. A detailed discussion of the leading contributions to Learning Analytics by the OU is given in Chapter Two, Literature Review.

7.4.5 University of Hertfordshire Strategic IT Management module

A detailed description of my experiment conducted on a live final year university module student cohort of 23, where individual student data is limited to lecture/tutorial attendance, virtual learning environment accesses and intermediate assessments is given in Chapter Eight. Techniques applied were Decision Tree, Random Forest and K-Nearest Neighbour regression.

7.5 Chapter Summary

In this chapter I have provided an explanation of each of the AI/ML techniques relevant to my research. I have then described each of the experiments and analyses I carried out on the datasets identified in the previous chapter, including a brief introduction to an experiment conducted on a live student cohort (detailed in Chapter Nine). In each case I have presented my results indicating likely useful attribute correlations to student performance which may prove useful in learning analytics and student outcome prediction. I have also described my contribution of identifying a novel technique for the analysis of nominal data and compared the results of its application to the Portuguese student dataset with those achieved by contingency table and chi-square test techniques. In the following chapter I describe a live

experiment to identify students potentially at risk, conducted on a small student cohort of 23, with minimal available student attributes totalling 3.

CHAPTER EIGHT

Experiment to establish the potential for student performance prediction in small cohorts with minimal available attributes using learning analytics techniques

8.1 Introduction

8.1.1 Contributions to Knowledge Relevant to this Chapter

This experiment directly supports my contribution demonstrating that it is possible and useful to predict student performance on courses comprising relatively small student cohorts, where a very limited set of student attributes are readily available for analysis. In addition, the results of this experiment directly support my contribution of demonstrating how the analysis of these limited attributes: attendance, VLE accesses and intermediate assessments, may provide potentially useful intervention guidance to academic leadership.

All sections of this chapter have been published previously (Wakelam et al., 2020) with the exception of section 8.7.1.3.

8.1.2 Summary of Chapter Content

In this chapter I describe an experiment conducted on a final year university module student cohort of 23, where individual student data is limited to lecture/tutorial attendance, virtual learning environment accesses and intermediate assessments. I found potential for predicting individual student interim and final assessment marks in small student cohorts with very limited attributes and that these predictions could be useful to support module leaders in identifying students potentially "at risk". This chapter addresses the following research questions:

Section 1.2.1, "*How accurately can we predict student performance on courses comprising relatively small student cohorts, where a very limited set of student attributes are readily available for analysis?"*

Section 1.2.2, "*How useful would these analyses be in order to provide course leadership with the opportunity to make timely supportive interventions at appropriate points during a module?*"

Section 1.2.3, "*Which data mining techniques are suitable for predicting student performance?*"

8.2 Motivation for Experiment

As discussed in Chapter Two, Literature Review, Section 2.2.1 Learning Analytics, academics have traditionally used their interactions with students through class activities and interim assessments to identify those who may be at risk of failure or withdrawal. Reduced lecture/tutorial attendance, mitigated by on-line availability of course material, generally considered to be a factor in lower lecture/tutorial

attendances, makes such direct identification of students at risk more challenging for academic staff. The introduction of data mining and machine learning techniques providing increasingly accurate predictions of student examination assessment marks have focussed upon the analysis of so called “big data” of large student populations and wide ranges of data attributes per student. Many university modules comprise relatively small student cohorts, with institutional protocols limiting the student attributes available for analysis. It appears that very little research attention has been devoted to this area of analysis and prediction and its potential value to academic staff to support timely interventions.

In this experiment I am interested in the application of learning analytics for the prediction of intermediate and final student assessment marks, where the student cohort is small and with very limited attributes for each student. In order to provide appropriate benchmarks for comparison, the comparative prediction accuracies across a variety of techniques, applied to large student cohorts with multiple student attributes, are discussed. This analysis is supported by corresponding traditional statistical analyses of potential correlations and their significance between the predictive attributes used, and evaluation of how much of the variance in the final assessment marks can be attributed to each of the available student attributes.

Ethical approval limited the student attributes available to my experiment to interim and final course assessments, VLE accesses and student attendance at lectures and tutorials. We may consider each of these as “low sensitivity” attributes (see Section 3.3.1 Potentially useful student attributes) and therefore unlikely to present ethical and privacy obstacles in the majority of academic institutions. In the case of each of these three attributes there is evidence that they are useful predictors of student success. This is discussed in Chapter Two Literature Review, Section 2.2.1 Learning Analytics, with supporting reference citations: Assessments (Sclater et al., 2016); VLE accesses (Doijode & Singh, 2017), Attendance (Aziz & Awlla, 2019;Fike & Fike, 2008).

8.3 Experiment Design

Three machine learning techniques, Decision Tree (DT), K-Nearest Neighbours (KNN) and Random Forest (RF) analyses to analyse and predict student performance, were applied and compared at appropriate points during module delivery. These points were selected to coincide with intermediate assessments. DT, KNN and RF methods were selected given their ability to perform well when some values are missing (Quinlan, 2014) and their widespread core use in learning analytics research (Ashraf et al., 2018). Given that the experiment is designed to analyse student performance breakdown, missing values may be expected. In the case of this experiment, missing values occur where a student chooses not to take part in an interim assessment. For example, only the highest two of the three multiple choice

assessments (see Table 8.1) count towards the student's final mark and in some cases students who scored highly in the first two of these assessments chose to not sit the third. After module completion, RF analysis was applied retrospectively at each intermediate assessment point to make overall module score predictions and evaluate their accuracy.

In order to investigate statistical significance between the means of the prediction results of each of the machine learning techniques applied (Decision Tree, K-Nearest Neighbours and Random Forest) against actual and predicted student overall module results the ANOVA (Analysis of Variance) technique was applied and p-values discussed.

This technique was applied in order to determine whether the associations between student attributes and the final assessment results are statistically significant and the resulting p-values are discussed.

The results of the relative importance of each of the student attributes, generated by the application of the Random Forest technique, are also presented and discussed.

In addition, graphical (histogram) analyses of the correlations between each of the major attributes (Attendance vs Final marks, VLE accesses vs Final marks, Interim assessments vs Final marks) are presented and discussed.

8.4 Module Description

The selected course instance is a Level 6 (Final Year undergraduate) Computer Science module, duration 15 weeks (including a 3 week vacation period and 2 weeks allocated for submission and review of each of the two final assessments) comprising 5 intermediate summative assessments and no final examination. Each week students are expected to attend a two hour lecture and a one hour tutorial. During the course of the module there are 10 lectures and 9 tutorials. Three EVS (Electronic Voting System) in-class tests are included, with the best two results counting towards the final overall module assessment (see Table 8.1). The module has a profile of early "low stakes" assessments with "higher stakes" assessments later in the module. The module VLE comprises of 8 sections, including the course guide for example, however student focus was overwhelmingly on the News and Teaching sections.

Table 8.1: Module Assessments

Week No.	Name	Description	Number of Weeks to Complete Assessment	Submit on Week No.	Result Publication Week No.	Percentage Contribution to Final Result
1	EVS1	Multiple choice	Immediate	4	4	5%
2	EVS2	Multiple choice	Immediate	6	6	5%
3	EVS3	Multiple choice	Immediate	10	10	5%
4	Group Presentation	Group work and presentation	6	11	12	40%
5	Individual Report	Technical Report	8	15	18	50%

Note that only the highest two scores of the three EVS results contribute to the final result.

8.5 Dataset Description

The student cohort is 23. For each student the attributes collected comprise attendance at lectures/tutorials, VLE accesses and intermediate assessment results spread throughout the module (Table 8.2). These attributes were supplemented by the addition of synthesised attributes: Delta increase in attendance from prior period; Cumulative VLE News section accesses; Cumulative VLE Teaching section accesses and Cumulative VLE accesses (see Section 4.2 Artificial Intelligence and Machine Learning Techniques, discussion on the potential benefits of feature engineering). Ethics approval limited analysis to dynamic data collected during course execution. Static attributes such as gender, age, prior academic results were not included.

Table 8.2: Student Attributes

Attribute	Data Range
Lecture/tutorial attendance	1-19
Delta increase in attendance from prior period	1% - 100%
Cumulative VLE News section accesses	0 - unlimited
Cumulative VLE Teaching section accesses	0 - unlimited
Cumulative VLE accesses	0 - unlimited
EVS1 Result	0% - 100%
EVS2 Result	0% - 100%
EVS3 Result	0% - 100%
Group Presentation Result	0% - 100%
Individual Report Result	0% - 100%

8.6 Methodology

Three machine learning techniques were applied, Decision Tree (regression), K-Nearest Neighbours and Random Forest to predict student assessment marks, using only their attendance, VLE accesses, and intermediate summative assessments results. The aim of these techniques is to create a model that takes these input values to predict the value of a target variable, in this case the students' assessment marks.

8.6.1 Summary of Machine Learning Techniques

A description of each of these machine learning techniques is given in Chapter Four, Relevant AI and ML Techniques, as follows: Decision Tree (see Section 7.2.5), K-Nearest Neighbours (see Section 7.2.7) and Random Forest (see Section 7.2.6).

8.6.2 Design of Experiments to meet Research Questions

Commencing at module registration, each student's attendance at lectures and tutorials was recorded, both as a simple count and as a percentage of overall module tutorials/lectures to date. As well as cumulative attendance, the delta increases between the measurement points were recorded, which were each selected to coincide with intermediate assessments. A continuous count of individual student "accesses" on items

in the VLE was maintained. Of the 8 sections of the VLE, 99% of student accesses were in only 2 sections, News and Teaching. The News section included all module announcements and weekly reminders of tasks to complete. The Teaching section included all course material. For the purposes of the experiment each of these two section accesses in the analyses were included. Intermediate and final assessment results were recorded for each student. This resulted in the dataset shown in Table 8.2. For each analysis point, each of Decision Tree, K-Nearest Neighbours and Random Forest analyses were carried out and the resultant predictions compared with actual student results and the level of accuracy measured. These analyses included the overall module result at module completion. Regression methods were selected to enable the prediction of an actual assessment mark, as opposed to classification methods which would simply predict a pass or fail. This data mining method is often used in the construction of predictive models (Daniel, 2015). The measurement methods used were percentage relative error/accuracy, Mean Squared Error (MSE) and Correlation Coefficient (CC). Prediction accuracies between the analysis methods were compared. I then repeated the analyses combining the two VLE section accesses (see Table 6) into one total in order to determine sensitivity. The progressive prediction results at each assessment point were shared with the module leader for consideration of potential interventions during module delivery. To provide module leadership with data which could potentially support their choice of intervention approach, tabular and graphical comparative analyses of attendance, VLE accesses and intermediate assessment results were also provided. Additionally, the prediction analyses at each assessment point were repeated, based upon the assessment results data alone, excluding attendance and VLE "accesses" in order to compare results. Upon availability of the overall module result after module completion, the collected data at each assessment point was revisited and overall module result prediction analyses performed at each point. I selected Random Forest for these analyses given that it delivered the most accurate predictions in earlier analyses. Upon module completion, the correlation between all assessments, including overall module results was investigated.

Statistical significance between the means of the prediction results of each of the machine learning techniques and between student attributes and the final assessment results was investigated using the ANOVA technique and p-values discussed. The results of the relative importance of each of the student attributes, generated by the application of the Random Forest technique, are also presented and discussed.

8.6.3 Performance Measurement

Percentage relative accuracy is measured as the percentage accuracy of the prediction compared to the actual student result. This permitted a direct comparison with the measurement method used by Ashraf et al. 2018 which compared the results of various data mining techniques, as described in section 2.2 above.

Mean Squared Error measures how close a prediction (regression) line is to the set of actual data points, by calculating the distances from the points to the prediction line (distances are the "errors"), squaring them and calculating their average (mean). The squaring removes any negative signs as well as giving more weight to the larger differences. Correlation Coefficient (CC) measures how strongly variables are related to each other by dividing their covariance by the product of their standard deviations. A CC of +1 indicates a perfect positive correlation, which means that as variable X increases, variable Y increases and while variable X decreases, variable Y decreases. A CC of -1 indicates a perfect negative correlation. For the purposes of identifying the strongest overall correlations for each analysis technique the average using absolute CC values were calculated.

8.7 Experimental Results

8.7.1 Research Question 1 and Research Question 3

Q1: How accurately can we predict student performance on courses comprising relatively small student cohorts, where a very limited set of student attributes are readily available for analysis?

Q3: Which data mining techniques are suitable for predicting student performance?

The value and usefulness of prediction based upon small student cohorts (in this case 23) and where organisational barriers limit the availability of student data. The results under each of machine learning analyses and traditional statistical methods are summarised. The full machine learning results summary is included in Appendix G.

8.7.1.1 Machine Learning Analyses

For each of three prediction accuracy measures, Relative % Accuracy, Mean Squared Error and Correlation Coefficient, the results of each of Decision Tree, K-Nearest Neighbours and Random Forest analyses, carried out at each assessment point (Tables 8.3, 8.4 and 8.5), are presented. In each case, this includes both the analyses results where VLE News and Teaching accesses are included as separate attributes and where they are combined as one attribute. Prediction accuracy is calculated as 100% - Absolute value of (Actual assessment result – predicted result)/100%. The results of each technique are then discussed.

Table 8.3: Prediction Accuracy Measured by Relative %Accuracy

Relative % Accuracy	EVS1	EVS2	EVS3	Group Pres'n	Indiv. Rep	Module Result	Ave % Accuracy	Ave % Accuracy (Excl. Module result)
Decision Tree Regression	72%	33%	57%	74%	64%	90%	65%	60%
Decision Tree Regression (Combined VLE Clicks)	77%	32%	57%	96%	69%	88%	70%	66%
K Nearest Neighbour, K=1	71%	54%	52%	90%	70%	88%	71%	67%
K Nearest Neighbour, K=1 (Combined VLE Clicks)	73%	46%	52%	89%	57%	81%	66%	63%
K Nearest Neighbour, K=2	74%	49%	66%	86%	74%	89%	73%	70%
K Nearest Neighbour, K=2 (Combined VLE Clicks)	74%	58%	63%	89%	69%	81%	72%	71%
K Nearest Neighbour, K=3	74%	55%	74%	74%	72%	89%	73%	70%
K Nearest Neighbour, K=3 (Combined VLE Clicks)	76%	60%	68%	88%	73%	82%	75%	73%
Random Forest	80%	56%	70%	81%	71%	91%	75%	72%
Random Forest (Combined VLE Clicks)	80%	50%	65%	90%	71%	86%	74%	71%

The overall module result is an arithmetic combination of the intermediate assessment results (see Table 8.1) and therefore we would expect all the prediction methods at the module result assessment point to deliver the most accurate results. This is clearly the case with accuracy between 81% and 91%, averaging 86%. The less than 100% accuracy in each case may be explainable by a combination of inaccuracies in

the prediction techniques used and the influence of attendance and VLE access data. Random Forest and K-Nearest Neighbours (k=3) with VLE accesses combined delivered the highest average prediction each with accuracies of 75%. Importantly for potential intervention opportunities, predictions at each of the intermediate assessment points using these analysis techniques, although mixed (between 56% and 88%) were promising in several cases, with accuracies at 70% or above at 9 of the 12 points. The least accurate results were delivered by Decision Tree Regression and K-Nearest Neighbours (k=1) with VLE accesses combined, averaging 65% and 66% respectively.

Table 8.4: Prediction Accuracy Measured by Mean Squared Error

Mean Squared Error	EVS1	EVS2	EVS3	Group Pres'n	Indiv. Rep	Module Result	Ave MSE	Ave MSE (Excl. Module result)
Decision Tree Regression	0.0767	0.1489	0.1051	0.0411	0.0603	0.0137	0.0743	0.0743
Decision Tree Regression (Combined VLE Clicks)	0.0459	0.1435	0.1019	0.0127	0.0603	0.0158	0.0634	0.0634
K Nearest Neighbour, K=1	0.0806	0.0969	0.1464	0.0216	0.0611	0.0213	0.0713	0.0713
K Nearest Neighbour, K=1 (Combined VLE Clicks)	0.0736	0.1101	0.1426	0.0247	0.0838	0.0315	0.0777	0.0777
K Nearest Neighbour, K=2	0.0527	0.0982	0.0781	0.0261	0.046	0.0217	0.0538	0.0538
K Nearest Neighbour, K=2 (Combined VLE Clicks)	0.0634	0.0755	0.0841	0.0229	0.0586	0.032	0.0561	0.0561
K Nearest Neighbour, K=3	0.0527	0.0842	0.0591	0.0334	0.0532	0.0181	0.0501	0.0501
K Nearest Neighbour, K=3 (Combined VLE Clicks)	0.0613	0.0669	0.0692	0.0028	0.0526	0.0289	0.0470	0.0470
Random Forest	0.0341	0.0657	0.0756	0.0359	0.0461	0.0191	0.0461	0.0461
Random Forest (Combined VLE Clicks)	0.0465	0.0922	0.0726	0.0189	0.0542	0.0196	0.0507	0.0507

As with the Relative % Error measure, the most accurate prediction results (in the case of MSE these are the closest results to zero) are as expected at the overall module result assessment point. At this point, MSE values are between 0.01 and 0.03. Similarly to Relative % Error measure, RF and KNN (K=3) with VLE accesses combined delivered the most accurate prediction results, excluding the overall module result predictions, with average MSE values of 0.046 and 0.047 respectively. The least accurate results

were delivered by KNN (K=1) with VLE accesses combined, DT and KNN (K=1) with average MSE values of 0.08, 0.07 and 0.07 respectively.

Table 8.5: Prediction Accuracy Measured by Correlation Coefficient

Correlation Coefficient	EVS1	EVS2	EVS3	Gp Pres'n	Indiv. Rep	Module Result	Ave CC (Absolute Values)	Ave CC (Excl. Module result)
Decision Tree Regression	-0.0912	-0.4518	0.0706	-0.0224	0.1732	0.7386	0.2580	0.1618
Decision Tree Regression (Combined VLE Clicks)	0.2754	-0.5090	-0.0426	0.7853	0.1732	0.6942	0.4133	0.3571
K Nearest Neighbour, K=1	0.042	0.0843	0.2329	0.558	0.0262	0.5363	0.2466	0.1887
K Nearest Neighbour, K=1 (Combined VLE Clicks)	-0.0295	-0.0083	-0.1433	0.4638	0.2651	0.1394	0.1749	0.1820
K Nearest Neighbour, K=2	-0.1536	-0.34	0.0701	0.3899	0.1541	0.5424	0.2750	0.2215
K Nearest Neighbour, K=2 (Combined VLE Clicks)	-0.0295	0.1093	0.0683	0.5106	-0.0404	0.0455	0.1339	0.1516
K Nearest Neighbour, K=3	-0.0876	-0.3019	0.2973	0.146	-0.1536	0.7391	0.2876	0.1973
K Nearest Neighbour, K=3 (Combined VLE Clicks)	-0.3137	0.0535	0.1928	0.5069	-0.0402	0.2075	0.2191	0.2214
Random Forest	0.4165	0.1443	0.0648	0.1352	0.1711	0.5985	0.2551	0.1864
Random Forest (Combined VLE Clicks)	0.0438	-0.2986	0.1289	0.62	0.0732	0.579	0.2906	0.2329

As with average % accuracy and MSE, CC prediction results are strongest at the overall module result assessment point, with CC values between 0.05 and 0.74. However, in the case of CC, it is DT with VLE accesses combined that delivers the strongest prediction results with an average CC of 0.4, followed by RF with VLE accesses combined and KNN, K=3 each with an average CC of 0.29. The least accurate results were delivered by KNN, K=1 and K=2, with VLE accesses combined giving CC values of 0.13 and 0.17 respectively. The remaining analysis techniques delivered promising prediction results with CC values between 0.22 and 0.28. In order to investigate the corresponding effect of attendance and VLE access data, the analyses were repeated using only the assessments and excluding all other data. The results were mixed with only very small variations leading me to believe that inaccuracies in the

prediction techniques themselves are the major contributor. An illustrative subset of the results is presented (Table 8.6).

Table 8.6: Comparison of Analyses Including all Attributes against those using Assessment Results Only.

Analysis Technique	Prediction Accuracy Measure	EVS3	Gp Pres'n	Indiv Rep
K Nearest Neighbour, K=3 (Combined VLE Clicks)	**Relative % Accuracy**	74% / 67%	74% / 82%	72% / 73%
	Mean Squared Error	0.0591 / 0.0553	0.0334 / 0.0427	0.0532 / 0.0498
	Correlation Coefficient	0.2973 / 0.4164	0.146 / -0.2093	-0.1536 / 0.1254

Results including all attributes are shown first and results using the assessment results only (i.e. excluding attendance and VLE accesses) are shown second. We can see that the comparative results are mixed. Recommendations for further work include investigating the predictive effect of cumulative multi-year analyses on the inclusion of attendance and VLE accesses data. After module completion, an overall module result prediction analysis at each assessment point, using Random Forest analysis was performed (Table 8.7).

Table 8.7: Module Result Prediction at each Assessment Point

Analysis Technique	Prediction Accuracy Measure	EVS1	EVS2	EVS3	Gp Pres'n	Indiv Rep	Ave Accuracy
Random Forest	**Relative % Accuracy**	82%	82%	86%	83%	85%	84%
	Mean Squared Error	0.0325	0.0334	0.0253	0.0323	0.0206	
	Correlation Coefficient	-0.0207	0.1114	0.3763	0.1866	0.5483	

Average student final result prediction accuracies of between 82% and 86% were obtained using Random Forest analyses. However, the variance between individual student predictions and their actual final result at each assessment point was high, with accuracies ranging from 11% to 99% (Table 8.8). MSE and CC accuracies performed in line with relative % accuracy analyses.

Table 8.8: Range of Individual Student Final Result Percentage Prediction Accuracies at Assessment Points

Prediction Accuracy	EVS1	EVS2	EVS3	Grp Pres'n	Indiv Report
Lowest	38%	11%	52%	28%	35%
Highest	98%	98%	100%	99%	99%

8.7.1.2 Correlations between assessments

An analysis of the cross-correlation between each of the interim assessments and the overall module result (Table 8.9) shows moderate, high and very high correlations with the overall module result. Of

these 5 interim assessments, high and very high correlations were found between the two major interim assessments (Group Presentation and Individual Report) and the overall module result. The initial three interim assessments were all moderately correlated with the overall module result.

Table 8.9: Assessments Correlation Matrix

	EVS1	EVS2	EVS3	Group Pres'n	Indiv. Report	Overall Module Result
EVS1	1.00	0.53	0.63	0.47	0.44	0.55
EVS2		1.00	0.59	0.49	0.60	0.66
EVS3			1.00	0.42	0.42	0.51
Grp Pres'n				1.00	0.73	0.90
Indiv. Rep					1.00	0.95
Overall Module Result						1.00

Key:		
Very Highly Correlated	**(0.9 to 1.0)**	
Highly Correlated	**(0.7 to 0.89)**	
Moderately Correlated	**(0.5 and 0.69)**	
Low Correlation	**(0.3 to 0.49)**	

8.7.1.3 Statistical Analysis of the Associations and Statistical Significance of Attributes and Final Assessment Results

Evaluation of the Significance of each Attribute to the Final Assessment result

In order to evaluate the significance of each attribute to the final assessment result I have used the Multiple Linear Regression ANOVA technique (see Section 7.2.12). The results are shown in Table 8.10.

Table 8.10: Multiple Linear Regression ANOVA Analysis of Attributes vs Final Assessment Result

	Coefficients	***Standard Error***	***t Stat***	***p-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Overall Module Result (Intercept)	0.6010	0.0016	368.4634	1.1E-25	0.5975	0.6046	0.5975	0.6046
Module Att.	0.0000	0.0000	65535	#NUM!	0.0000	0.0000	0.0000	0.0000
Δ Att. EVS1 to EVS2	0.0006	0.0031	0.1969	#NUM!	-0.0061	0.0073	-0.0061	0.0073
Δ Att. EVS2 to EVS3	0.0007	0.0026	0.2466	0.8094	-0.0051	0.0064	-0.0051	0.0064
Δ Att. EVS3 to Gp Pres'n	0.0007	0.0023	0.3218	0.7532	-0.0043	0.0058	-0.0043	0.0058
StudyNet News clicks	0.0065	0.0032	2.0034	0.0682	-0.0006	0.0135	-0.0006	0.0135
StudyNet Teaching clicks	-0.0064	0.0033	-1.9092	0.0804	-0.0136	0.0009	-0.0136	0.0009
EVS1 Result	0.0104	0.0027	3.8687	0.0022	0.0045	0.0162	0.0045	0.0162
EVS2 Result	0.0105	0.0025	4.1540	0.0013	0.0050	0.0160	0.0050	0.0160
EVS3 Result	0.0054	0.0026	2.1051	0.0570	-0.0002	0.0110	-0.0002	0.0110
Gp Pres'n Result	0.0710	0.0033	21.4461	0.0000	0.0638	0.0782	0.0638	0.0782
Indiv Rep Result	0.0997	0.0032	31.3066	0.0000	0.0927	0.1066	0.0927	0.1066

In the case of this experiment the null hypothesis is that the attribute measured has no effect upon the actual module result. In the majority of analyses, a measure of 0.05 is used as the p-value cut off for

significance (McDonald, 2009). If the p-value is ≤ 0.05, we reject the null hypothesis that there is no difference between the means and conclude that a significant difference does exist. A value of $p > 0.05$ is the probability that the null hypothesis is true. A statistically significant test result ($p \leq 0.05$) means that the test hypothesis is false or should be rejected. In the case of my experiment, a p-value > 0.05 means that the respective attribute has no effect upon the actual module result.

I have set 5% (95% hypothesis testing confidence level) as the measure of the significance of any attribute to the actual module result. Therefore, attributes with a p-value ≤ 0.05 are statistically significant to the actual module result.

The results (Table 8.10) suggest that each of the EVS1, EVS2, Group Presentation Assessment and Individual Report Assessment attributes are statistically significant to the actual module result. It is worth noting that the p-values of StudyNet News Clicks (0.068), StudyNet Teaching Clicks (0.080) and EVS3 Assessment Results (0.057) are relatively close to our p-value cut off measure of 0.05. It may be that the inclusion of data accumulated from one or more previous occurrences of the module could improve the significance of any of these attributes to the actual module result (see Chapter Nine: Conclusions and Future Work, section 9.4.3).

In respect of the attributes Module Attendance and Delta increase in Attendance between EVS and EVS Assessments, p-values generated returned indeterminate (#NUM!) results. Such results can arise if one of the attribute's results is linearly dependent upon the others or is predictable from the other attributes (Energy, 2011). In such cases, the removal of the respective attribute results from the analysis may resolve the problem. In order to explore this possibility, I removed the Module Attendance attribute and applied Multiple Linear Regression ANOVA analysis to the resulting data. The results are shown in Table 8.11.

Table 8.11: Multiple Linear Regression ANOVA Analysis of Attributes vs Final Assessment Result Excluding Overall Attendance

	Coefficients	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Overall Module Result (Intercept)	0.6010	0.0016	368.4634	1.08E-25	0.5975	0.6046	0.5975	0.6046
Δ Att. EVS1 to EVS2	0.0006	0.0031	0.1969	0.8472	-0.0061	0.0073	-0.0061	0.0073
Δ Att. EVS2 to EVS3	0.0007	0.0026	0.2466	0.8094	-0.0051	0.0064	-0.0051	0.0064
Δ Att. EVS3 to Gp Pres'n	0.0007	0.0023	0.3218	0.7532	-0.0043	0.0058	-0.0043	0.0058
StudyNet News clicks	0.0065	0.0032	2.0034	0.0682	-0.0006	0.0135	-0.0006	0.0135
StudyNet Teaching clicks	-0.0064	0.0033	-1.9092	0.0804	-0.0136	0.0009	-0.0136	0.0009
EVS1 Result	0.0104	0.0027	3.8687	0.0022	0.0045	0.0162	0.0045	0.0162
EVS2 Result	0.0105	0.0025	4.1540	0.0013	0.0050	0.0160	0.0050	0.0160
EVS3 Result	0.0054	0.0026	2.1051	0.0570	-0.0002	0.0110	-0.0002	0.0110
Gp Pres'n Result	0.0710	0.0033	21.4461	6.16E-11	0.0638	0.0782	0.0638	0.0782
Indiv Rep Result	0.0997	0.0032	31.3066	7.1E-13	0.0927	0.1066	0.0927	0.1066

Removal of student attendance attributes from the analysis was successful in eliminating indeterminate (#NUM!) results.

As above, 5% is set as the measure of the significance of any attribute to the actual module result. The results (Table 8.11) delivered almost identical results in terms of p-values for the attributes included compared to those with Attendance included (Table 8.10), also suggesting that each of the EVS1, EVS2, Group Presentation Assessment and Individual Report Assessment attributes are statistically significant to the actual module result. It is also the case that the p-values of StudyNet News Clicks (0.068), StudyNet Teaching Clicks (0.080) and EVS3 Assessment Results (0.057) are relatively close to our p-value cut off measure of 0.05.

Statistical Significance Tests for Comparing each of the Machine Learning Techniques

The results of the ANOVA analysis comparing the overall module prediction results of each of DT, KNN (K=3) and RF machine learning techniques (Table 8.12) are shown in Table 8.13.

Table 8.12: Student Module Result Predictions for each Machine Learning Technique

`	**Predicted Module Result**			**Actual Module Result**
Student	**DT**	**KNN (K=3)**	**RF**	
1	0.61	0.69	0.57	0.62
2	0.56	0.61	0.58	0.64
3	0.27	0.62	0.61	0.48
4	0.62	0.56	0.62	0.61
5	0.55	0.59	0.63	0.62
6	0.77	0.67	0.56	0.71
7	0.76	0.66	0.61	0.75
8	0.74	0.69	0.74	0.84
9	0.74	0.73	0.73	0.87
10	0.43	0.53	0.50	0.00
11	0.77	0.64	0.44	0.70
12	0.61	0.63	0.63	0.64
13	0.58	0.58	0.56	0.53
14	0.77	0.68	0.71	0.67
15	0.62	0.55	0.63	0.61
16	0.62	0.52	0.60	0.54
17	0.62	0.62	0.57	0.57
18	0.76	0.71	0.73	0.77
19	0.49	0.54	0.53	0.42
20	0.49	0.53	0.37	0.39
21	0.61	0.59	0.58	0.65
22	0.62	0.63	0.62	0.59
23	0.62	0.61	0.64	0.60

Table 8.13: ANOVA analysis of Comparison of Machine Learning Technique Predictions

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
DT	23	14.22078	0.618295	0.015336
KNN (K=3)	23	14.20671	0.617683	0.003727
RF	23	13.76106	0.598307	0.007772

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	0.005944	2	0.002972	0.332268	0.718488	3.135918
Within Groups	0.590347	66	0.008945			
Total	0.596291	68				

The p-value of 0.718488 is > 0.05 which means that our predictions from each of the techniques are not statistically significantly different.

The results of the ANOVA analysis comparing the overall module prediction vs actual results of each of DT, KNN (K=3) and RF machine learning techniques (Table 8.14) are shown in Table 8.15.

Table 8.14: Student Module Result Predictions vs Actual Results for each Machine Learning Technique

	Differences between Predicted and Actual Module Results		
Student	DT	KNN (K=3)	RF
1	0.01	0.11	0.09
2	0.14	0.05	0.10
3	0.43	0.30	0.28
4	0.00	0.08	0.01
5	0.12	0.05	0.02
6	0.08	0.06	0.21
7	0.02	0.12	0.18
8	0.12	0.18	0.12
9	0.15	0.15	0.15
10	0.00	0.00	0.00
11	0.10	0.08	0.37
12	0.04	0.01	0.02
13	0.10	0.10	0.07
14	0.15	0.01	0.06
15	0.01	0.09	0.03
16	0.14	0.04	0.11
17	0.09	0.10	0.01
18	0.02	0.08	0.06
19	0.16	0.29	0.26
20	0.25	0.35	0.07
21	0.07	0.09	0.12
22	0.05	0.07	0.05
23	0.03	0.02	0.07

Table 8.15: ANOVA analysis of Comparison of Machine Learning Technique Predictions vs Actual results

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
DT	23	2.28	0.09913	0.009336
KNN (K=3)	23	2.43	0.105652	0.008744
RF	23	2.46	0.106957	0.009277

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	0.000809	2	0.000404	0.044343	0.956655	3.135918
Within Groups	0.601835	66	0.009119			
Total	0.602643	68				

The p-value of 0.956655 is > 0.05 which means that the differences between our predictions and actual module results from each of the techniques are not significantly different.

The relative importance of each of the student attributes

The Random Forest prediction analysis method (see Section 4.2.6) also provides a measure of the importance of each attribute. This measure is the increase in prediction error if the values of that attribute are permuted across the out-of-bag observations. The measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble.

In the case of my experiment, Table 8.16 presents the outputs from the RF analysis measures of the importance of each attribute to the actual module result in descending order from the most important.

Table 8.16:

Attribute	Importance Measure	Order of Importance
Module Attendance	0	6=
Δ Attendance EVS1 to EVS2	0.6398	5
Δ Attendance EVS2 to EVS3	0	6=
Δ Attendance EVS3 to Gp Pres'n	0	6=
Δ Attendance Gp Pres'n to Indiv. Rep	0	6=
Δ Attendance Indiv. Rep to Module End	0	6=
StudyNet News clicks	0	6=
StudyNet Teaching clicks	0.7071	2=
EVS1 Results	0	6=
EVS2 Results	0.7071	2=
EVS3 Results	-0.7071	13
Gp Pres'n Results	0.7071	2=
Indiv. Report	0.9171	1

The analysis shows Individual Report Assessment attributes as the most important relative to the actual module result, followed by the Group Presentation Assessment attribute, EVS2 Assessment result and

StudyNet Teaching clicks attributes in equal second place. These results are similar to those found in the machine learning analyses above (see Section 8.7.1.1) and relatively unsurprising given the significant contributions of 40% and 50% of these assessments in the calculation of the overall module result. The high placing in order of importance of the EVS2 Results and in particular StudyNet Teaching Clicks attributes has not been identified in other analyses. This may be explained by, anecdotally, students placing some importance in their performance on the second of three EVS tests given that only the best of two of the three contributed to their overall module result and their reluctance to relinquish their best score to the final EVS assessment. It may be the case that StudyNet Teaching Clicks was relatively highly placed because of the value of on-line material including the opportunity to try out exemplar material.

Graphical analyses of Student Attributes vs Overall Assessment Marks

The following graphical analyses compare each of overall student attendance (Figure 8.1), VLE accesses (Figure 8.2) and each of the 5 interim assessments (Figures 8.3, 8.4, 8.5, 8.6 and 8.7) respectively against overall module assessment marks. Each figure is discussed in turn.

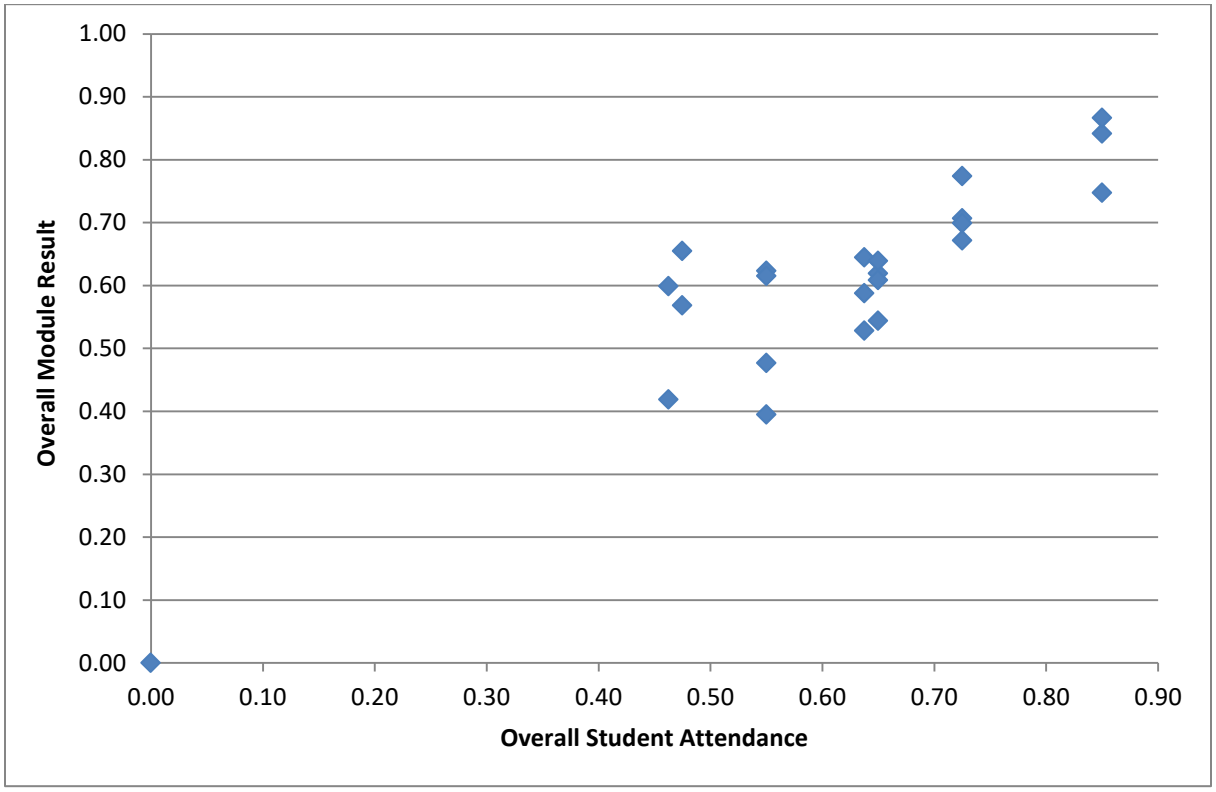


Figure 8.1: Overall Student Attendance v Overall Module Result

There appears to be some correlation between student attendance and their overall module result. This is not unexpected given published research findings (Aziz & Awlla, 2019;Fike & Fike, 2008) which show positive correlations between students' attendance and performance.

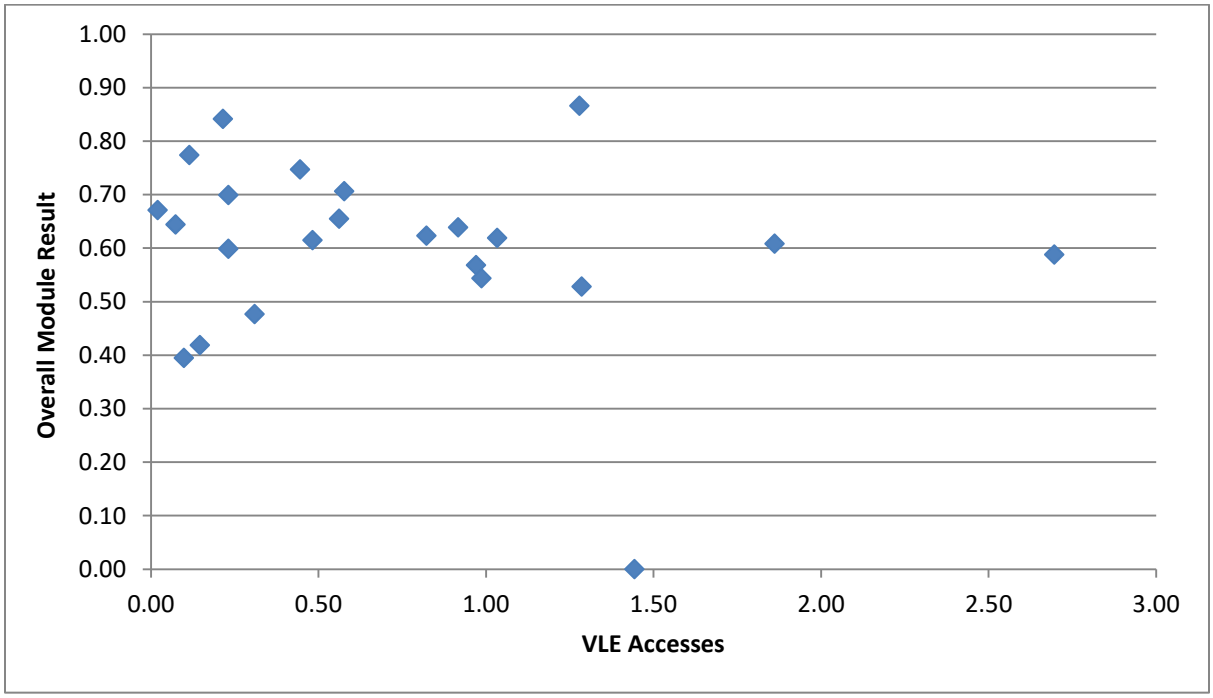


Figure 8.2: VLE Accesses v Overall Module Result

There does not appear to be any clear relationship between a student's VLE accesses and their overall module result in the experiment. (This was also the case when VLE Teaching and VLE News accesses were plotted separately). This is contrary to the recent study (Heuer & Breiter, 2018) analysing student VLE activity across 22 courses and 32,593 OU students which found student VLE accesses to be an important indicator of student performance. However, given that nearly all OU learning takes place online, the frequency of VLE access will be very high and therefore have a significant impact on student performance.

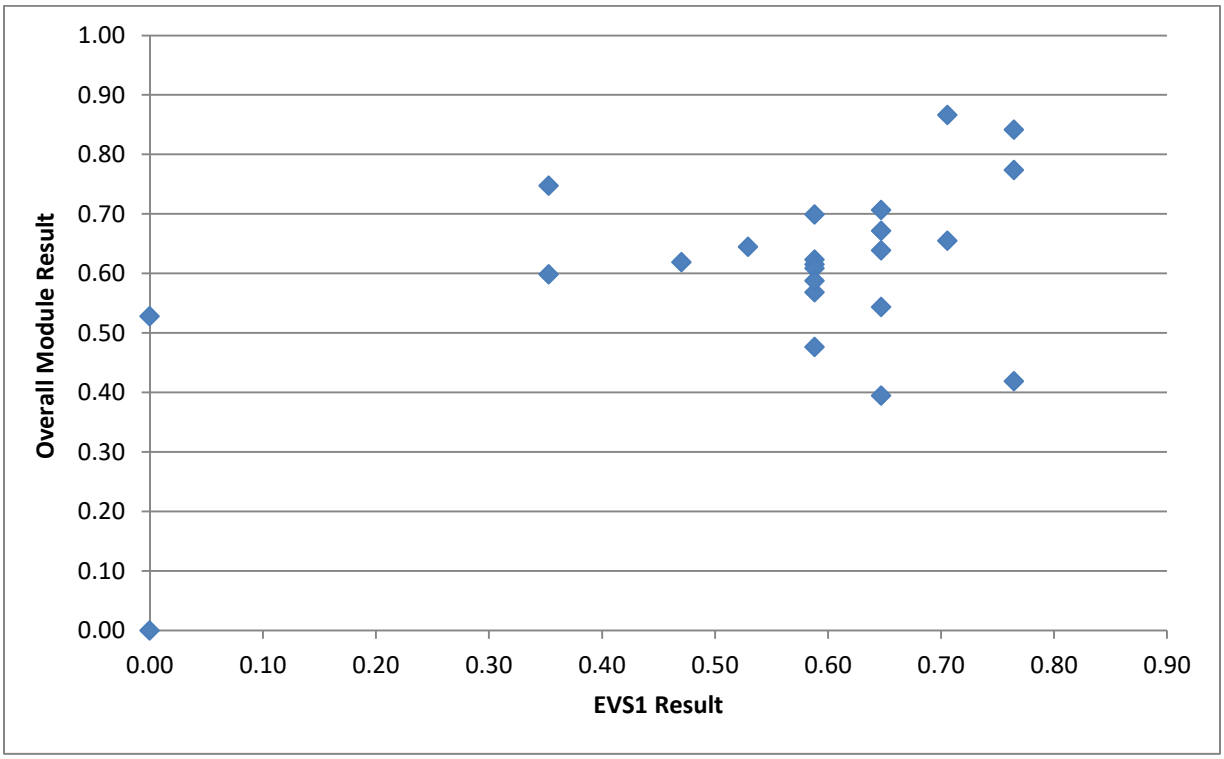


Figure 8.3: EVS1 Result v Overall Module Result

There appears to be some modest correlation between the results of the student's EVS1 assessment and their overall module result.

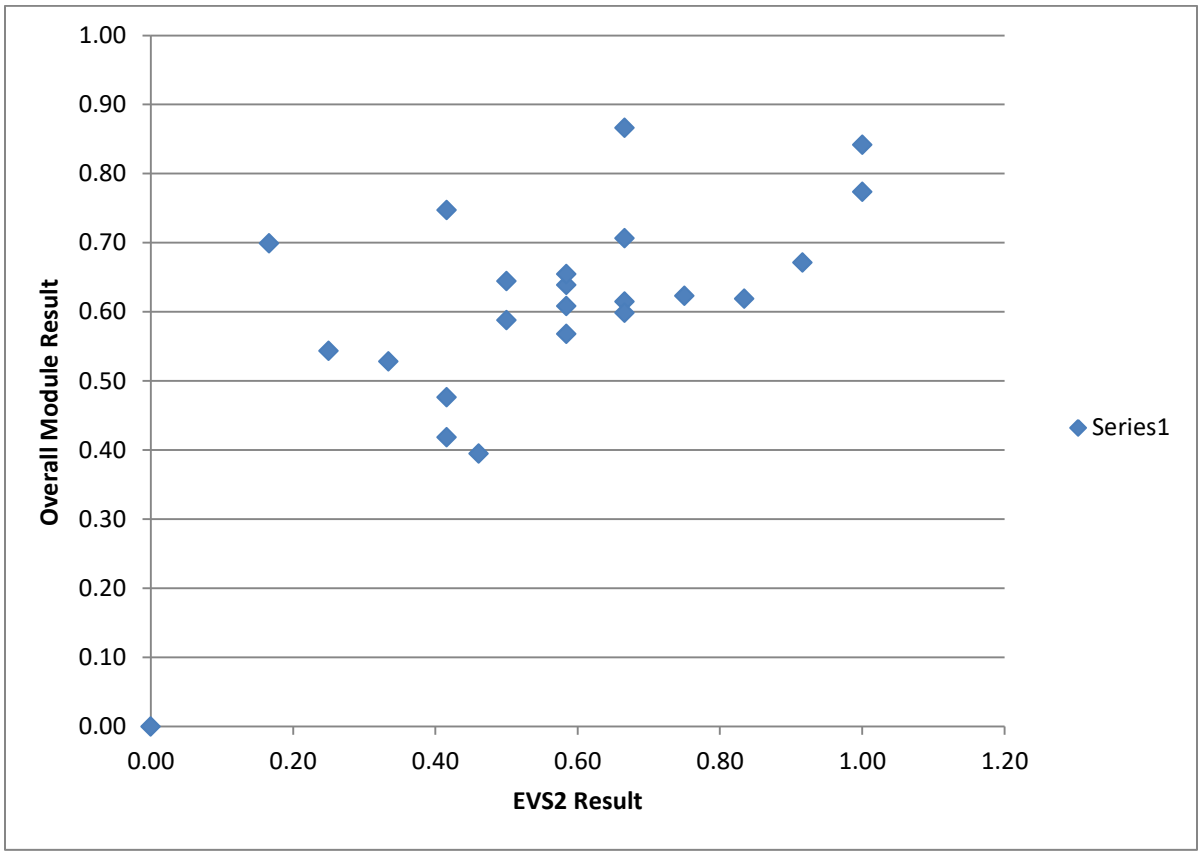


Figure 8.4: EVS2 Result v Overall Module Result

As is the case with EVS1 assessment results, there appears to be some modest correlation between the results of the student's EV2 assessment and their overall module result.

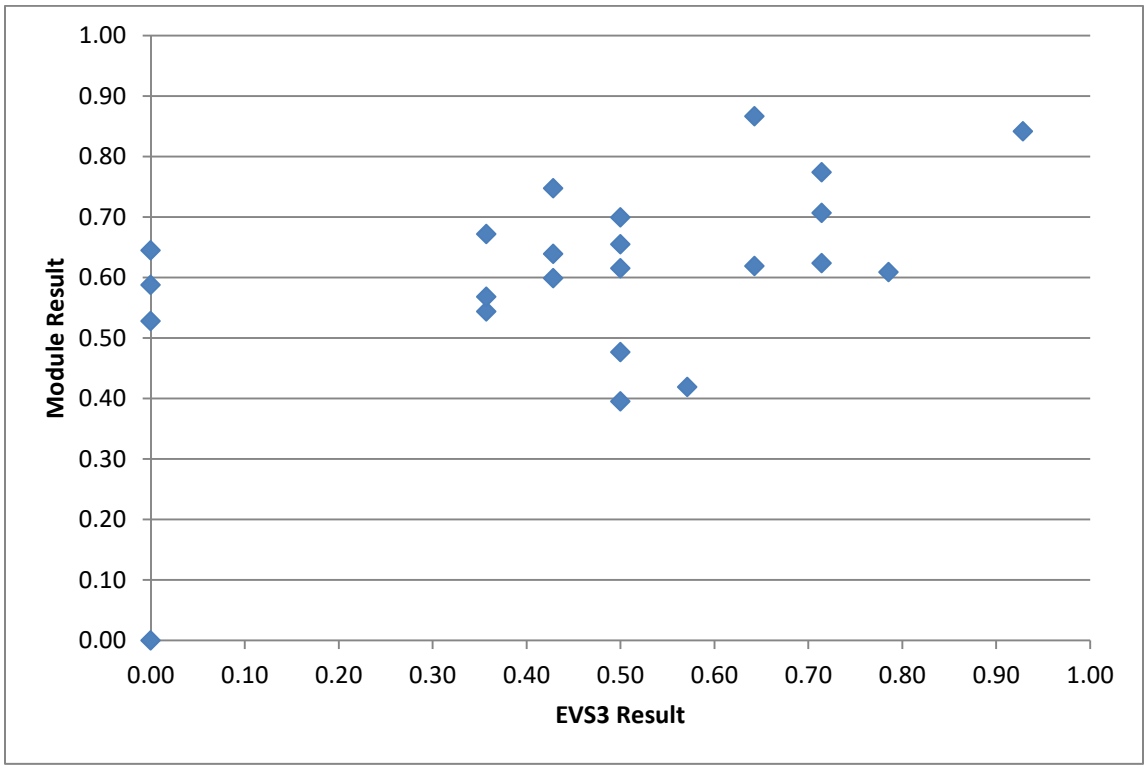


Figure 8.5: EVS3 Result v Overall Module Result

As is the case with EVS1 and EVS2 assessment results, there appears to be some modest correlation between the results of the student's EV3 assessment and their overall module result.

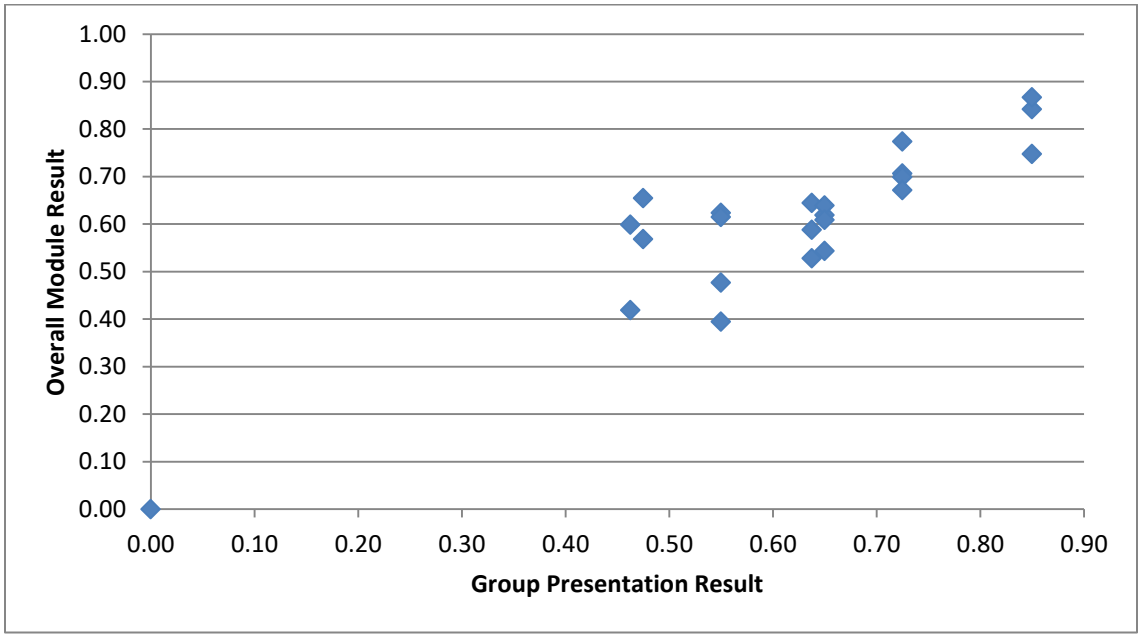


Figure 8.6: Group Presentation Result v Overall Module Result

There appears to be strong correlation between student Group Presentation assessment result and their overall module result. This corresponds to published research findings showing some evidence that interim assessment as part of the overall course assessment is a strong predictor of student success (Sclater et al., 2016). In the case of this experiment, where the Group Presentation assessment represents 40% of the overall module mark, this is not unexpected.

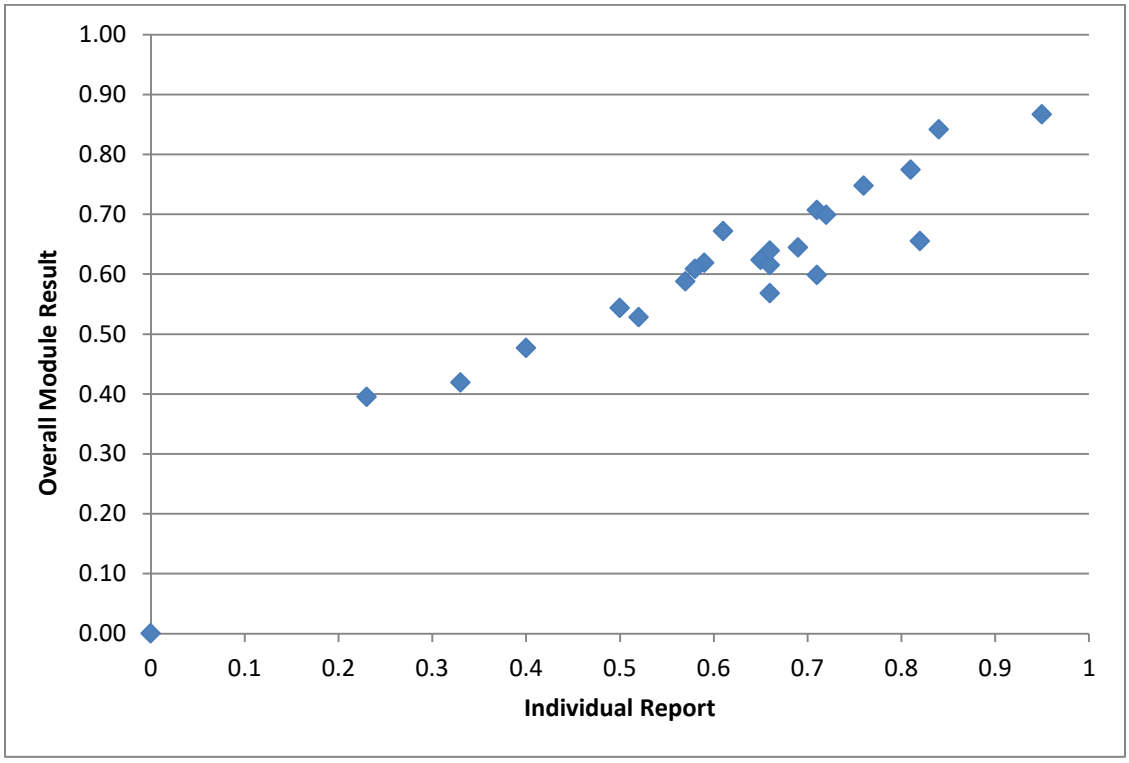


Figure 8.7: Individual Report Result v Overall Module Result

There appears to be strong correlation between student Individual Report assessment result and the overall module result. This corresponds to published research findings showing some evidence that interim assessment as part of the overall course assessment is a strong predictor of student success (Sclater et al., 2016). In the case of this experiment, where the Individual Report assessment represents 50% of the overall module mark, this is not unexpected.

8.7.1.4 Graphical Analyses to Support Potential Interventions

Example graphical analyses performed at EVS3 and individual report assessment points are discussed and shown below (Figures 8.8 to 8.13). In each figure, the student identification number (1 to 23) is labelled on the x axis. Note that student 10 withdrew from the module prior to assessment commencement.

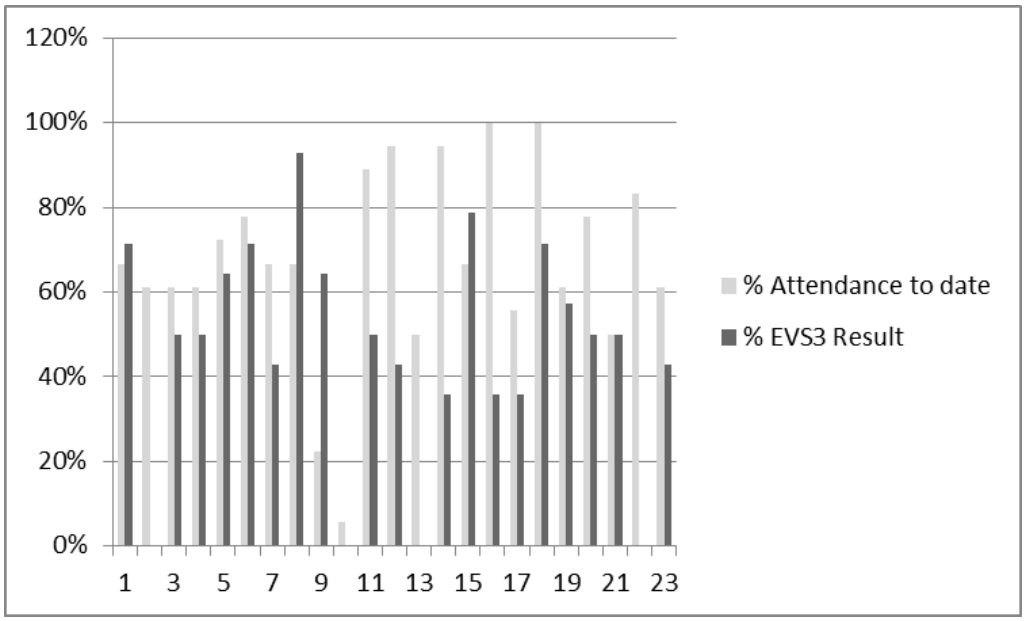


Figure 8.8: Attendance to Date v EVS3 result

Machine learning predictions for students 12 and 14 highlighted 62% and 97% negative disparities with their actual and expected progress raising concerns with module leadership. We can see from this table that in both cases their attendance records are very high and therefore not a cause for leadership concern. Student 22 had scored well in EVS1 and EVS2 assessments and given that the best two of the three assessments only are included chose not to take EVS3.

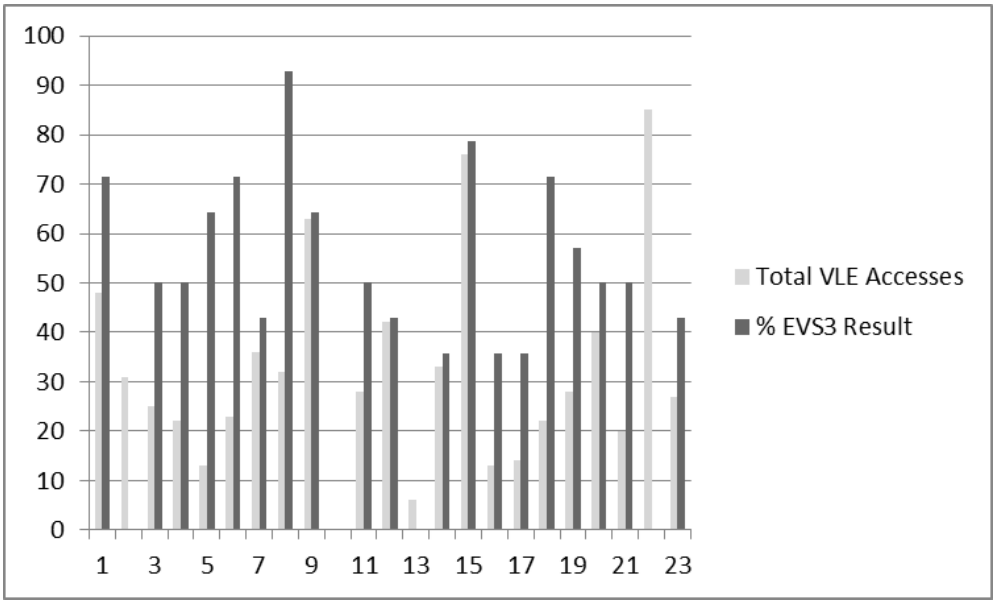


Figure 8.9: Total VLE accesses v EVS3 result

A glance at this chart shows that both student 12 and student 14 are registering average VLE accesses and this could be an area for concern and potential intervention.

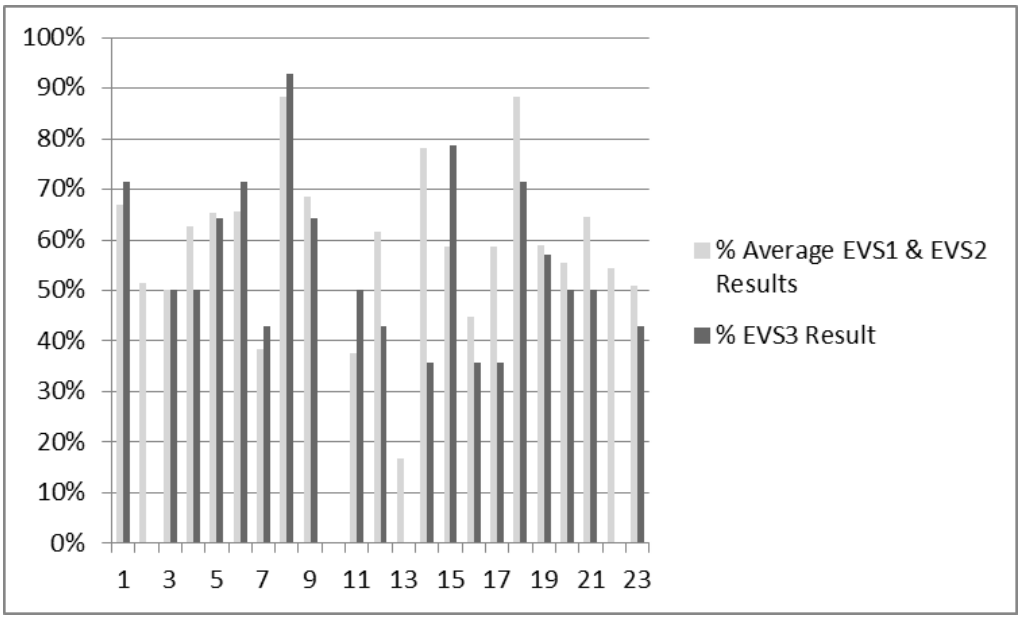


Figure 8.10: Average of EVS1 and EVS2 results v EVS3 result

As above, using students 12 and 14 as examples, we can see that their high average EVS1 and EVS2 results indicate why machine learning prediction disparities were evident.

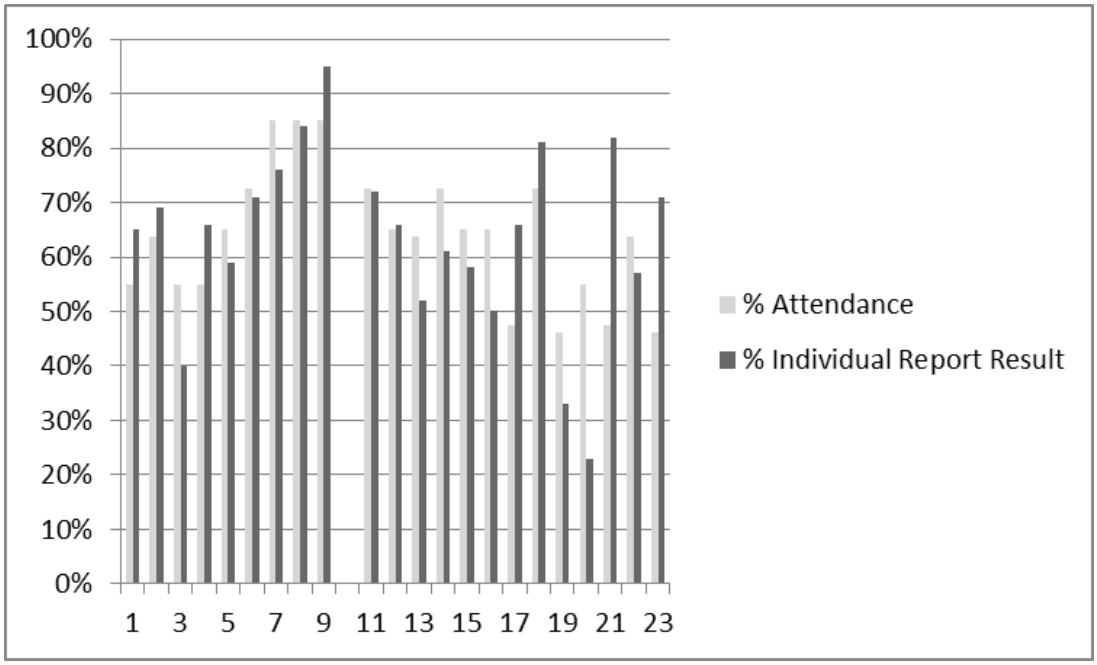


Figure 8.11: Attendance to Date v Individual Report Result

Machine learning predictions for students 19 and 20 highlighted 159% and 179% negative disparities with their actual and expected progress raising concerns with module leadership. We can see from this table that both students are maintaining average attendance.

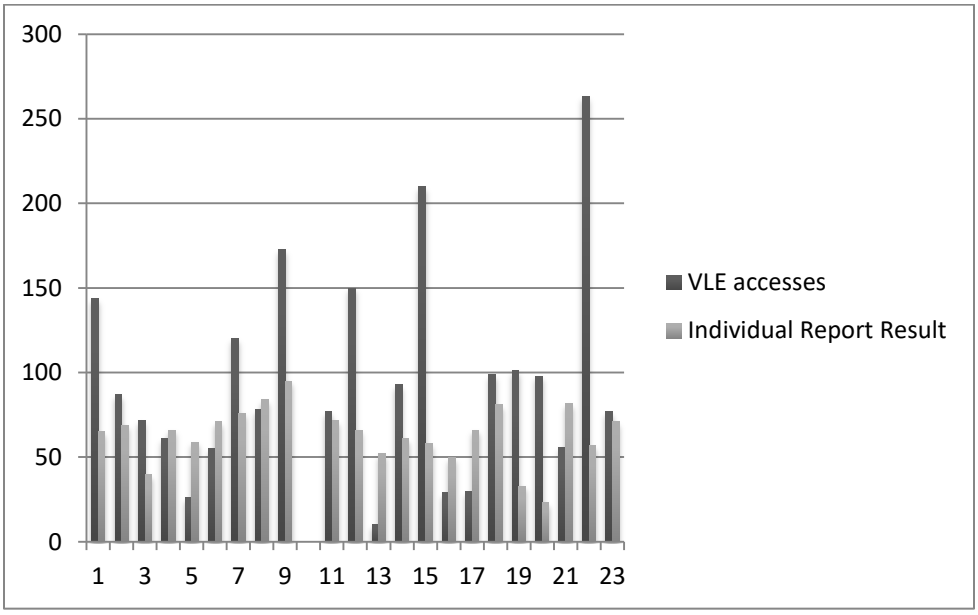


Figure 8.12: Total VLE Accesses v Individual Report Result

A consideration of this chart shows that both student 19 and student 20 are registering above average VLE accesses but may still be an area for potential intervention.

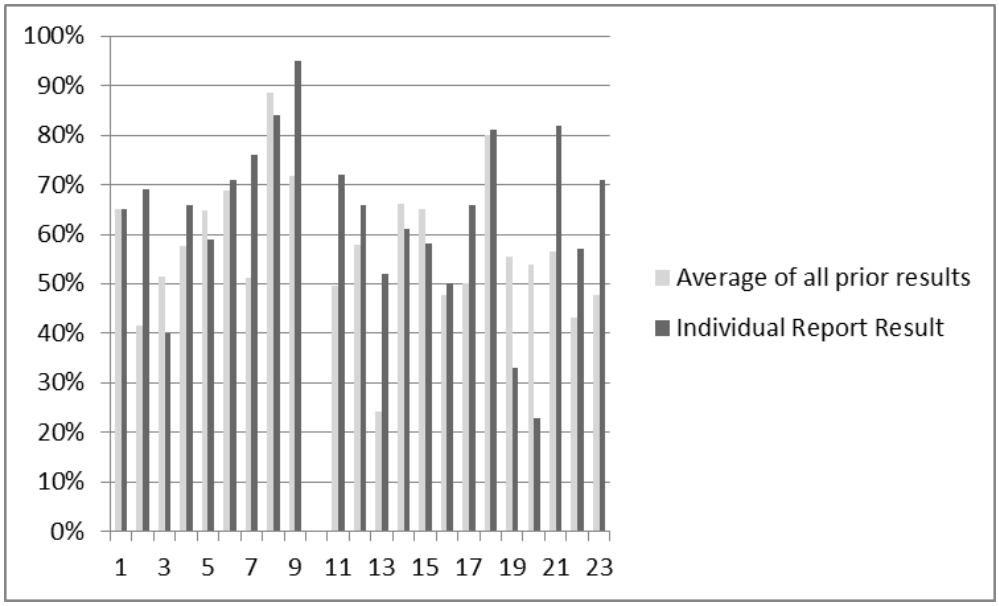


Figure 8.13: Average of EVS1, EVS2, EVS3 and Group Presentation Results v Individual Report Result

As above, using students 19 and 20 as examples, we can see that both have good average assessment results to date. In this case, module leadership considered intervention unnecessary.

8.7.2 Research Question 2

How useful would these analyses be in order to provide course leadership with the opportunity to make timely supportive interventions at appropriate points during a module?

The value and usefulness of prediction analyses for intervention opportunities may now be considered. For these analyses to be of value for interventions they must be available to module leadership while sufficient time is left for successful interventions to be made and any consequent positive effects to be achieved by the student. The early and mid-timed assessments in the selected module provided this opportunity. The progressive prediction analyses conducted may also provide module leadership with useful data in respect of module and assessment design. For example, if the predictions on individual assessments were consistently accurate it may be that these assessments are adding little value in their current form and require revision. Adaptive learning systems dynamically adjust the number of questions upwards and downwards and dynamically adjust student learning paths depending upon student performance (Wakelam et al., 2015). The graphical analyses (see Section 8.7.1.4) proved useful for module leadership to perform "at a glance" assessments of student activities. For example, where a student prediction suggests a performance risk, module leadership were able to quickly view their attendance and VLE usage in support of personal experience of the student. This in itself may suggest intervention methods, ranging from encouraging improved attendance or more usage of VLE material. In the case of this module, students where machine learning predictions identified potential poor outcomes could be reviewed, supported by "at a glance" comparisons of their attendance, VLE accesses and prior assessment marks. This information coupled by module leadership knowledge of each student through face to face lectures and tutorials supported direct interventions, including coaching and the provision of additional teaching material. These interventions may be grouped under the heading of providing additional scaffolding to students. Research conducted by Stubbs et al. at Manchester Metropolitan University discusses how a meta-framework for assisting the design of learning frameworks to educational designers to support improved learning outcomes (Stubbs et al., 2006).

8.8 Discussion and Conclusions

8.8.1 Research Question 1

Is it possible and useful to predict student performance on courses comprising relatively small student cohorts, where a very limited set of student data is readily available for analysis?

Experimental results show some potential for analysing and predicting student assessment marks on courses comprising relatively small student cohorts, and where only a very limited set of student data is readily available for analysis. The average prediction accuracy across all machine learning techniques used was 67%, with K-Nearest Neighbours and Random Forest prediction accuracy between 66% and 75%. This compares favourably with student prediction accuracy levels achieved across a variety of machine learning techniques applied to large student cohorts with significantly more student attributes (Ashraf, 2018). The results in Ashraf and colleagues' study ranged from 50% to 97% (Tables 2.5 and 2.6). Importantly for potential intervention opportunities, some promising results were obtained at the point of the third assessment, approximately two thirds of the way through the module, with prediction accuracies of 74% and 70% for K-Nearest Neighbours and Random Forest Analyses respectively. Reducing the attributes used in the analyses gave mixed results. Combining VLE News and Teaching accesses into one total had very little effect upon prediction accuracy, in some cases giving a 1% improvement and in others the reverse. Reducing the attributes to only the intermediate assessment results gave us mixed results in comparison with prediction accuracy using all available attributes, hence it was not possible to reliably consider student interventions. Similarly, this provided little opportunity to determine the effect of including attendance and VLE accesses on prediction accuracy. I believe that the inclusion of all available attributes may be considered as at least benign to the analyses. There is some evidence (Heuer & Breiter, 2018) that the analysis of VLE accesses alone can be a useful predictor of student performance. Future work accumulating year on year module data to investigate the effects on prediction accuracy of multi-year data may provide further insight. As may be expected, the final assessment, the student's Individual Report which is submitted in week 15 of 18, contributing 50% to their overall mark, correlated very highly (correlation coefficient 0.95) with their overall module result. Additionally, the penultimate assessment, the Group Presentation, submitted in week 11 of 15, correlated highly (correlation coefficient 0.9) with the overall module result. Usefully, for the potential of earlier intervention opportunities, given their early assessment points of weeks 4, 6, 10 of 15, moderate correlations were found (correlation coefficients of 0.55, 0.66 and 0.51 respectively) between EVS1, EVS2 and EVS3 and the overall module result. In particular, student usage of VLE material and correlations between attendance and VLE usage on assessment marks provided valuable insights.

8.8.2 Research Question 2

How useful would these analyses be in order to provide course leaders with the opportunity to make timely supportive interventions at appropriate points during the module?

The analyses demonstrated three opportunities for module leadership to identify potentially "at risk" students and to consider appropriate timely interventions. These were machine learning analyses at intermediate assessment points, and the identification, post module completion, of which intermediate assessments provided the likeliest indicators of overall module success. Student performance in their third assessment, week 10 of 15, appears to be a useful measure of individual progress. In this experiment, module leadership were then able to review attendance and VLE access patterns for students whose performance was of concern. Alongside personal experience of the student in question an intervention decision could then be made. In the case of the module, the analyses led to module leadership identifying two specific opportunities for direct interventions, both following the third assessment, EVS3. In each case a student's predicted performance showed a likelihood of failing their next assessment. In case 1, further analysis showed a reduction in tutorial attendance. In case 2, analysis showed a combination of reduced lecture/tutorial attendance coupled with minimal activity in the VLE. This enabled leadership to engage in positive discussions with each student and provide specific guidance on their future studies. A variety of possible interventions are described in section 4.2, but could be as simple as evidence based discussions drawing a student's attention to their attendance, arranging additional individual or group lectures/tutorials or the availability of further and focussed supporting material on the VLE. Graphical analyses allowing the visualisation of relationships between attributes provides module leadership with further opportunities to identify any interesting correlations which could support positive interventions. These graphical presentations compared different combinations of attendance, VLE usage and assessment results providing easily referenceable "at a glance" supporting material to machine learning results for module leadership. In the case of the module in this experiment, module leadership found these representations supported intervention decisions. Given their significant mark contribution to the overall module result this was to be expected. Additionally, promising results at the earlier third assessment point gave module leadership the opportunity to consider interventions in time for their effects to be useful.

8.8.3 Research Question 3

Which data mining techniques are suitable for predicting student performance?"

Each of the three selected data mining techniques, Decision Tree, K-Nearest Neighbours and Random Forest, delivered some measure of success in my experiment in predicting student interim and final

assessment scores. Measured by relative % error, the average success rates across all six prediction points ranged between 65% and 75%, with K Nearest Neighbour, K=3 (combined VLE accesses) and Random Forest techniques achieving the highest at 75%. Decision Tree (regression) achieved the lowest average success rate at 65%. These results compare favourably with the results summarised in Ashraf and colleagues' study of relevant analytics between 2011 and 2017 which included implementations with student numbers in excess of 10,000 and 77 attributes in some cases (Ashraf et al., 2018) which ranged from 50% to 97%.

8.8.4 Implications to Practice and/or Policy

University expectations are currently that the application of learning analytics necessitates the availability of so-called "big data" in particular for modules with large student cohorts. Our results show that university practice can now usefully consider smaller scale deployments of learning analytics. Where student attributes for analysis are limited to readily available data such as student attendance, VLE accesses and intermediate assessment results, with no inclusion of demographic/personal data, either none, or very limited modifications are necessary to university policies. It is good practice to provide students with a clear explanation of what data is being collected and how the analysis is being done, allowing them to individually opt in or opt out of learning analytics implementations. In addition, alternative intervention methods should be documented and where possible students given the opportunity to express their preferences, for example, from dashboard presentation of predictions, system generated emails, offers of face to face supportive meeting with course tutors.

8.9 Chapter Summary

In this chapter I describe a live experiment to identify students potentially at risk, conducted on a small student cohort of 23, with minimal available student attributes of attendance, VLE accesses and prior assessment marks. I apply each of Decision Tree, K-Nearest Neighbour and Random Forest machine learning techniques, achieving assessment prediction accuracies averaging 67% across the three methods, with K-Nearest Neighbours and Random Forest prediction accuracies between 66% and 75%. These prediction accuracies compare favourably with published research across a variety of machine learning techniques applied to large student cohorts with significantly more student attributes, which achieved accuracies between 50% and 97%. The predictions demonstrated opportunities to identify potentially at-risk students and to consider appropriate timely interventions. In the following chapter I present the conclusions of my research and make recommendations for future work.

CHAPTER NINE

Conclusions and Future work

9.1 Introduction

9.1.1 Contributions to Knowledge Relevant to this Chapter

Each of my contributions to knowledge are included in the respective section of my conclusions below.

9.1.2 Summary of Chapter Content

In this chapter I present the conclusions drawn from each of my research questions (see Section 1.2). I then present recommendations for future work.

9.2 Conclusions

I have structured my conclusions to align with each of my research questions, in each case noting how they support my contributions to knowledge.

9.2.1 Research Question 1: Small Student Cohorts and Limited Student Attributes

How accurately can we predict student performance on courses comprising relatively small student cohorts, where a very limited set of student attributes are readily available for analysis?

While there is evidence to show that predictions based upon large cohorts with multiple student attributes can provide educators with useful support in identifying students at risk (Heuer & Breiter, 2018), there is little evidence of the value that can be derived where cohorts are small and very limited attributes are available for analysis. What are the relative predictive accuracies that may be achieved in the analysis of student outcomes when the student cohort is small (23 in the case of my experiment) and student attributes are limited to lecture/tutorial attendance, Virtual Learning Environment (VLE) accesses and five formal interim assessments?

I conducted a live experiment to identify students potentially at risk, conducted on a small student cohort of 23, with minimal available student attributes of: attendance, VLE accesses and prior assessments, applying each of Decision Tree, K-Nearest Neighbour and Random Forest machine learning techniques. I achieved assessment prediction accuracies averaging 67% across the three methods, with K-Nearest Neighbours and Random Forest prediction accuracies between 66% and 75% respectively. These prediction accuracies compare favourably with published research across a variety of machine learning techniques applied to large student cohorts with significantly greater numbers of student attributes (which achieved accuracies between 50% and 97%).

This supports my contribution to knowledge of:

- Establishing the potential for predicting individual student interim and final assessment marks in small student cohorts with very limited attributes and showing that these predictions could be useful to support module leaders in identifying students potentially at risk during the course of their studies (Wakelam et al., 2020). Demonstrating through the analysis of these limited attributes: attendance, VLE accesses and intermediate assessments, how useful intervention guidance may be provided to academic leadership.

9.2.2 Research Question 2: The Opportunity to Make Interventions

How useful would these analyses be in order to provide course leadership with the opportunity to make timely supportive interventions at appropriate points during a module?

The value of the implementation of learning analytics is directly related to their success in consequent application to support students and institutions through appropriate timely interventions. What are the methods and timeliness of such interventions which are critical to their success, and which methods are preferred by students and therefore most likely to be successful? What student ethical, moral and privacy issues must be taken into consideration?

The predictions achieved in my live experiment with a small cohort and minimal attributes successfully demonstrated opportunities to identify potentially at-risk students and to consider appropriate timely interventions. These intervention opportunities included both those which are academic staff actionable, such as one on one coaching, and the capability of automated alerts to students, such as low attendance at lectures and tutorials. Both were actionable from an early enough stage in the module execution to allow time for the student, with support, to respond positively and improve the chances of a successful outcome. In addition, the intervention opportunities included those which could lead to addressing issues with multiple students in the cohort during module execution, or module re-design for future occurrences.

I have described the significant impacts upon students of a failure to progress and the very significant financial and reputational impacts upon institutions. I have presented a comprehensive list of the factors which potentially affect student performance, including how they may be identified, noting that only 4 of the 27 potential factors identified are detectable by current AI/ML techniques and that almost none are concerned with the student's intellectual capability to complete the course of study. I have catalogued consequential non-computer facilitated and computer facilitated methods of student interventions, discussing their usefulness in achieving positive learning outcomes and research into how students prefer to receive interventions. In respect of preferences it is clear from the surveys that for interventions to be

successful, strong student preferences must be identified and taken into account before intervention protocols are put in place. Similarly, addressing the legal, ethical and moral considerations of learning analytics and consequent interventions is an essential prerequisite.

This additionally supports my contribution to knowledge in section 9.2.1 of:

- Establishing and publishing the potential for predicting individual student interim and final assessment marks in small student cohorts with very limited attributes and showing that these predictions could be useful to support module leaders in identifying students potentially at risk during the course of their studies. Demonstrating through the analysis of these limited attributes: attendance, VLE accesses and intermediate assessments, how useful intervention guidance may be provided to academic leadership.

9.2.3 Research Question 3: Data Mining Techniques

Which data mining techniques are suitable for predicting student performance?

Which data mining techniques are available for the prediction of student performance and how do their respective predictive accuracies compare when applied to differing student cohort sizes and differing varieties of student attributes? Which of these techniques are applicable to each of numeric and nominal data? What are the student attributes which may be available to learning analytics and how might students and institutions view their respective sensitivity to privacy issues and therefore present potential restrictions of their use in a learning analytics context?

I have identified and described the variety of AI/ML techniques available for the analysis of student data for outcome prediction, highlighting their advantages and disadvantages. This includes the available techniques for the analysis of both measurement (quantitative) and categorical data types. While a large variety of techniques are available to analyse measurement (quantitative) data, there are fewer techniques applicable to nominal data. I summarise the results of what I believe to be a novel technique to analyse nominal data by making a systematic comparison of data pairs, comparing the results with those of the chi-square test statistical method.

My experiments upon freely available student datasets, representing a variety of small, medium and large student cohorts and similar ranges of student attributes, using appropriately selected methods led to the selection of three methods to apply to my live experiment: Decision Tree, K-Nearest Neighbour and Random Forest. These methods delivered promising results, see RQ1 above. My research and analysis identified a large variety of potentially useful static and dynamic student attributes, ranging from the uncontroversial, such as attendance and intermediate assessment results to very sensitive demographic

data. In this respect, almost 30% of all of the attributes considered are not classified as sensitive or potentially sensitive and the majority of these measurable and directly related to the student's academic background and performance.

In support of the attributes used in my live experiment, I have identified previously published evidence that student attendance, interim assessments and VLE activity provide useful predictive data for learning analytics.

This supports my contribution to knowledge of:

- Established and publishing a novel technique for the analysis of nominal data, an important subset of student attribute data alongside numeric attributes.

9.2.4 Research Question 4: Current Intelligent Educational Technologies

What progress has been made in the development and deployment of intelligent learning/training systems and prototypes and what are the institutional barriers to the adoption of learning analytics, alongside corresponding approaches to their resolution?

What intelligent learning/training systems and prototypes, including adaptive learning and intelligent tutoring systems, are currently available in the education and commercial sectors? What are the institutional barriers which must be overcome in order to successfully implement learning analytic and intervention systems, the corresponding critical success criteria and alternative approaches to their resolution?

I have presented a survey of existing intelligent learning/training systems in each of the education and commercial sectors, comparing the results with the equivalent survey conducted in 2015 (Wakelam et al., 2015) in order to examine progress. Most notable is the increased percentage of system implementations or prototypes in the commercial sector, an increase of 10 percentage points to 32%. This trend of more investment in this sector may prove beneficial in the case of educational learning analytics in its likely cross fertilisation of ideas and techniques. These systems track student progress in real-time, applying learning analytic techniques to measure students' progress and personalise their teaching through reinforcement learning, modification of learning paths and tutor/trainer alert. The techniques and measurement of student attributes mirror and are directly relevant to research into learning analytics.

As is the case in any major computer system design and implementation, the deployment of learning analytics in educational institutions must overcome a variety of challenges and barriers to success, ranging from organisational and political obstacles to academic staff and students' concerns and needs. Using available research and my own experience in the software industry I have catalogued these

challenges and documented critical success criteria, including a mapping between the two. The successful deployment of any learning analytics and intervention system is critically dependent upon executive management, design and implementation management acknowledgement and implementation of these principles.

9.3 Significance of this Research and Relevance to Teaching Practice

University expectations may be that the application of learning analytics necessitates the availability of so-called "big data", in particular, modules with large student cohorts. Based on these expectations, the implementation of an application of learning analytics based upon the large variety and volumes of student data is a very significant step for universities, both in terms of implementation and operational cost and in terms of the supporting infrastructure which must be put in place. This may be a daunting prospect for many institutions already pursuing other critical objectives which must also ensure that the sometimes difficult to evidence benefits of LA to them justify its implementation.

My results show that university practice may now usefully consider smaller scale deployments of learning analytics. Such a deployment may serve as a pilot or proof of concept to the institution, allowing modest and more manageable first steps into the exploitation of learning analytics. This is a valuable first step since in addition to a requirement for relatively modest funding, it enables the university to explore academic staff reaction and feedback as well as the student ethical, privacy and legal considerations, e.g. where student attributes for analysis are limited to readily available data such as student attendance, VLE accesses and intermediate assessment results, with no inclusion of demographic/personal data, with the result that either none, or very limited modifications are necessary to university policies.

In addition, given that many university modules average class sizes of approximately 20 students (Huxley et al., 2017) this research supports the potential value of applying learning analytics under these circumstances.

9.4 Recommendations for Future Work

Seven main areas for future work present themselves:

9.4.1 My live experiment performed Decision Tree, K-Nearest Neighbours and Random Forest machine learning analyses using regression techniques. Regression techniques generate a numerical (continuous) output variable, which in the case of my experiment was a prediction of each student's assessment mark, for example 62 marks out of 100 (i.e. 62%). An alternative approach would be to perform the same machine learning analyses, but using classification techniques. The output variable from classification techniques is categorical (nominal or ordinal) separating the data into multiple classes, which in the case

of my experiment may be pass or fail, or perhaps A, B, C, D, E, F. A comparison of the resulting prediction accuracies may then be made.

9.4.2 My live experiment was conducted on a module where both the student cohort was small and only limited student attributes were available for analysis. It would be interesting to conduct further experiments where the student cohort is small, but where a wider selection of student attributes is available, for example, prior student module marks and examination results from previously attended institutions. Similarly, to conduct an experiment where the student cohort is much larger, but with the same student attributes as with this experiment, in each case comparing achieved prediction accuracies against those achieved in my experiment.

9.4.3 The data available to my live experiment was restricted to the one occurrence of the selected module. It may be the case that including data accumulated from one or more previous occurrences of the module may improve the prediction accuracies of the chosen machine learning methods.

9.4.4 The module in my experiment comprised a relatively even spread of formal assessments, with two at an early stage. The effects upon prediction accuracy of applying the same experimental analyses to a module where there are either fewer intermediate assessments or where they are conducted later in the module may be of value. In particular, would a different and more back-ended spread of intermediate assessment allow for timely and successful interventions?

9.4.5 A logical extension to the experiment conducted on the live student cohort would be to design and conduct an experiment which tracks and measures resulting changes in individual student attendance, VLE accesses and assessment scores resulting from applied academic staff interventions. This may provide useful guidance to academic staff on which intervention methods are the most successful.

9.4.6 Further experimentation of nominal data analysis using the novel method in comparison with the chi-square method would be interesting, particularly in their application to larger datasets.

9.4.7 The results of the learning analytics applied during my live experiment provided module leadership with useful data on student performance and assessment predictions. Given that the primary objective of these analytics and predictions is to support academic staff in identifying potential intervention opportunities, further work to establish the most useful and efficient methods to do so would be of value. For example, the development and evaluation of methods of providing the data to academic staff in the most easily assimilated and actionable ways possible. Also, the prototyping of appropriately non-technical dashboards and exploration and analysis of the most timely intervention method approaches would be of value. For example, an experiment to examine the appropriate balance between

automatically generated alerts (via email or SMS) or provision of additional learning material versus the requirement for personal intervention by staff.

I feel privileged by the interest in my study taken by the University of Hertfordshire and the opportunity for my personal involvement in the proposed learning analytics strategic study planned for 2020.

REFERENCES

Agresti, A. and Kateri, M., 2011. *Categorical data analysis* (pp. 206-208). Springer Berlin Heidelberg.

Alexander, B., Ashford-Rowe, K., Barajas-Murph, N., Dobbin, G., Knott, J., McCormack, M., Pomerantz, J., Seilhamer, R. and Weber, N., 2019. *EDUCAUSE Horizon Report 2019 Higher Education Edition* (pp. 3-41). EDU19.

Allinson, C.W. and Hayes, J., 1996. *The cognitive style index: A measure of intuition-analysis for organizational research.* Journal of Management studies, 33(1), pp. 119-135.

Alva, P., 2016. *Using Machine Learning and Computer Simulations to Analyse Neuronal Activity in the Cerebellar Nuclei During Absence Epilepsy*. PhD thesis, Hatfield, University of Hertfordshire.

Andonie, R., 2010. *Extreme data mining: Inference from small datasets*. International Journal of Computers Communications & Control, 5(3), pp. 280-291.

Arnold, K.E. and Pistilli, M.D., 2012, April. *Course signals at Purdue: Using learning analytics to increase student success*. In Proceedings of the 2nd international conference on learning analytics and knowledge LAK '12 (pp. 267-270). ACM.

Ashraf, A., Anwer, S. and Khan, M.G., 2018. *A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques*. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 44(1), pp. 122-136.

Ashrafi, P., 2016. *Predicting the absorption rate of chemicals through mammalian skin using Machine Learning algorithms.* PhD thesis, Hatfield, University of Hertfordshire.

Ashby*, A., 2004. *Monitoring student retention in the Open University: definition, measurement, interpretation and action.* Open Learning: The Journal of Open, Distance and e-learning, 19(1), pp. 65-77.

Atif, A., Bilgin, A. and Richards, D., 2015. *Student Preferences and Attitudes to the use of Early Alerts.* In Proceedings of the 21st Americas Conference on Information Systems, pp. 3410-3424.

Australian Government Department of Education and Training, 2016. *Attrition, Success and Retention*. Higher Education statistics, Appendix 4.

Aziz, S.M. and Awlla, A.H., 2019. *Performance Analysis and Prediction Student Performance to Build Effective Student Using Data Mining Techniques*. UHD Journal of Science and Technology, 3(2), pp. 10-15.

Bambrick, N 2016, Support *Vector Machines: A Simple Explanation*, KDnuggets, viewed 28 August 2019, https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Banihashem, S.K., Aliabadi, K., Ardakani, S.P., Delaver, A. and Ahmadabadi, M.N., 2018. *Learning analytics: A critical literature review*. Interdisciplinary Journal of Virtual Learning in Medical Sciences (In press), 9.

Barnett, R. 2014. *Conditions of Flexibility Securing a more responsive Higher Education system*. Higher Education Academy, York, UK.

BBC 2019, Artificial Intelligence, viewed 28 August 2019, http://www.bbc.com/future/tags/artificialintelligence

British Computer Society 2019, *Home Page*, viewed 28 August 2019, https://www.bcs.org/

Benabdellah, N.C., Gharbi, M. and Bellafkih, M., 2014, May. *Ant colony algorithm and new pheromone to adapt units sequence to learners' profiles*. In 2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14) (pp. 1-8). IEEE.

Bennett, E., 2019. *Students' learning responses to receiving dashboard data (Junhong Xiao trans.)*. Society for Research into Higher Education. Journal of Distance Education in China.

Bentler, P.M. and Bonett, D.G., 1980. *Significance tests and goodness of fit in the analysis of covariance structures*. Psychological bulletin, 88(3), p.588.

Bhalchandra, P., Muley, A., Joshi, M., Khamitkar, S., Darkunde, N., Lokhande, S. and Wasnik, P., 2016. *Prognostication of student's performance: An hierarchical clustering strategy for educational datase*t. In Computational Intelligence in Data Mining—Volume *1*, pp. 149-157. Springer, New Delhi.

Bhatia, A., Kaur, L., 2014. *Global Training & Development trends & Practices: An Overview.* International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-8), 3(8), pp. 77-78.

Brokmeier, P, 2019. *An Overview of Categorical Input Handling for Neural Networks.* Towards Data Science, viewed 1 November 2019, https://towardsdatascience.com/an-overview-of-categorical-input-handling-for-neural-networks-c172ba552dee

Brownlee, J 2016, *Classification And Regression Trees for Machine Learning*, Machine Learning Mastery, viewed 28 August 2019, https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

Buckingham Shum, S., Ferguson, R. and Martinez-Maldonado, R., 2019. Human-Centred Learning Analytics. *Journal of Learning Analytics*, *6*(2), pp. 1-9.

Cardiac Tutor 2019, *Home page*. Viewed 10 October 2019, https://artiteacher.wordpress.com/2018/05/16/cardiac-tutor/

Castles, J., 2004. *Persistence and the adult learner: Factors affecting persistence in Open University students*. Active learning in Higher Education, 5(2), pp. 166-179.

Chang, C.C. and Lin, C.J., 2011. *LIBSVM: a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), p.27.

Chang, V., 2016. *Review and discussion: E-learning for academia and industry*. International Journal of Information Management, 36(3), pp. 476-485.

Chauhan, G 2018, *All about Naive Bayes*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

Chipman, S.E.F., 2010. *Applications in Education and Training: A Force Behind the Development of Cognitive Science*. Topics in Cognitive Science, 2, pp. 386–397.

Choi, S.P., Lam, S.S., Li, K.C. and Wong, B.T., 2018. *Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions*. Journal of Educational Technology & Society, 21(2), pp. 273-290.

Clement, B., Roy, D., Oudeyer, P., Lopes. M., 2014. *Online Optimization of Teaching Sequences with Multi-Armed Bandits*. 7th International Conference on Educational Data Mining.

Cleveland-Innes, M. and Campbell, P., 2012. *Emotional presence, learning, and the online learning environment*. The International Review of Research in Open and Distributed Learning, 13(4), pp. 269-292.

Cleveland-Innes, M., Stenbom, S. and Hrastinski, S., 2014. *The influence of emotion on cognitive presence in a case of online math coaching*. In the 8th EDEN Research Workshop, 27-28 October, Oxford UK, pp. 87-94.

Clow, D., 2012. *The learning analytics cycle: closing the loop effectively*. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge LAK '12 (pp. 134-138). ACM.

Clow, D., 2013. *An overview of learning analytics*. Teaching in Higher Education, 18(6), pp. 683-695.

Coe, R., Aloisi, C., Higgins, S. and Major, L.E., 2014. *What makes great teaching? Review of the*

underpinning research. Project Report. Sutton Trust, London.

Coffield, F., Moseley, D., Hall, E. and Ecclestone, K., 2004. *Learning styles and pedagogy in post-16 learning: A systematic and critical review*. Learning & Skills Research Centre, London, UK.

Constantinidou, F. and Baker, S., 2002. *Stimulus modality and verbal learning performance in normal aging*. Brain and language, 82(3), pp. 296-311.

Corrin, L., Kennedy, G., French, S., Shum, S.B., Kitto, K., Pardo, A., West, D., Mirriahi, N. and Colvin, C., 2019. *The ethics of learning analytics in Australian Higher Education*. University of Melbourne, Centre for the Study of Higher Education.

Cortez, P. & Silva, A., 2008. *Using Data Mining to Predict Secondary School Student Performance*. In the Proceedings of 5th Annual Future Business Technology Conference, pp. 5–12.

Cross, S., Galley, R., Brasher, A. and Weller, M., 2012. *OULDI-JISC Project Evaluation Report: the impact of new curriulum design tools and approaches on institutional process and design cultures*. Open University UK.

Csefalvay, C 2018, *Quantifying hard retinal exudates using Growing Neural Gas algorithms*, Medium, viewed 28 August 2019 https://medium.com/starschema-blog/growing-neural-gas-models-theory-and-practice-b63e5bbe058d

Daily Mirror 2019, *Artificial Intelligence*, viewed 28 August 2019, https://www.mirror.co.uk/all-about/artificial-intelligence

Dalipi, F., Imran, A.S. and Kastrati, Z., 2018, *MOOC dropout prediction using machine learning techniques: Review and research challenges*. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 1007-1014). IEEE.

Daniel, B., 2015. *Big data and analytics in Higher Education: Opportunities and challenges*. British journal of educational technology, 46(5), pp. 904-920.

Dearden, L., 2015. *Driverless cars trialled on UK roads for first time in four towns and cities.* The Independent UK, 11 February.

DeFreitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., Dunwell, I. and Arnab, S., 2015. *Foundations of dynamic learning analytics: Using university student data to increase retention*. British Journal of Educational Technology, 46(6), pp. 1175-1188.

Delone, W.H. and McLean, E.R., 2003. *The DeLone and McLean model of information systems success: a ten-year update*. Journal of management information systems, 19(4), pp. 9-30.

Deng, H 2018a, *An Introduction to Random Forest*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/random-forest-3a55c3aca46d

Deng, H 2018b, *Why random forests outperform decision trees*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5

Digest of Education Statistics 2017, *Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected years, 2006 to 2015*, viewed 28 August 2019, https://nces.ed.gov/programs/digest/d16/tables/dt16_326.30.asp.

Doijode, V. and Singh, 2017, N., *Predicting student success based on interaction with virtual learning environment*. Oklahoma State University.

Dolmans, D.H., Loyens, S.M., Marcq, H. and Gijbels, D., 2016. *Deep and surface learning in problem-based learning: a review of the literature*. Advances in health sciences education, 21(5), pp. 1087-1112.

Dorigo, M 2007, *Ant colony optimization*, Scholarpedia, viewed 28 August 2019, http://www.scholarpedia.org/article/Ant_colony_optimization

Drachsler, H. and Greller, W., 2016, April. *Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics*. In Proceedings of the sixth international conference on learning analytics & knowledge LAK '16, pp. 89-98. ACM.

Energy, M 2011, *#NUM! in the Results of a Regression Analysis*, Microsoft Community, viewed 12 December 2019, https://answers.microsoft.com/en-us/msoffice/forum/msoffice_excel-mso_other/num-in-the-results-of-a-regression-analysis/3578e346-fcf0-40c1-a855-df0e2c0f8329

European Union 1995, *Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*, viewed 12 December 2019 , https://www.refworld.org/docid/3ddcc1c74.html

Everything About Data Science, 2015, *Types of Statistical Data: Numerical, Categorical, and Ordinal*, viewed 28 August 2019, http://scaryscientist.blogspot.com/2015/02/classification-of-data-types.html

Fatima, D, Fatima, S, 2015. *A Survey on Research work in Educational Data Mining*. IOSR Journal of Computer Engineering (IOSR-JCE), 17 (2), pp. 43-49.

Fatima, D., Fatima, S. and Prasad, A.K., 2015. *A survey on research work in educational data mining.* IOSR Journal of Computer Engineering (IOSR-JCE), 17(2), pp. 43-49.

Fazey, D.M. and Fazey, J.A., 2001. *The potential for autonomy in learning: Perceptions of competence, motivation and locus of control in first-year undergraduate students*. Studies in Higher Education, 26(3), pp. 345-361.

Felder, R.M. and Silverman, L.K., 1988. *Learning and teaching styles in engineering education.* Engineering education, 78(7), pp. 674-681.

Ferguson, R., 2012. *Learning analytics: drivers, developments and challenges*. International Journal of Technology Enhanced Learning, 4(5/6), pp. 304-317.

Ferguson, R., Clow, D., Macfadyen, L., Essa, A., Dawson, S. and Alexander, S., 2014, March. *Setting learning analytics in context: Overcoming the barriers to large-scale adoption*. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge LAK '14 (pp. 251-253). ACM.

Ferguson, R., Cooper, A., Drachsler, H., Kismihók, G., Boyer, A., Tammets, K. and Monés, A.M., 2015, March. *Learning analytics: European perspectives*. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge LAK '15 (pp. 69-72). ACM.

Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., Rienties, B., Ullmann, T. and Vuorikari, R., 2016. *Research evidence on the use of learning analytics: Implications for education policy*. Joint Research Centre, Seville, Spain.

Ferguson, R. and Clow, D., 2017, March. *Where is the evidence?: a call to action for learning analytics.* In Proceedings of the seventh international learning analytics & knowledge conference (pp. 56-65). ACM.

Ferguson, R., Coughlan, T., Egelandsdal, K., Gaved, M., Herodotou, C., Hillaire, G., Jones, D., Jowers, I., Kukulska-Hulme, A., McAndrew, P. and Misiejuk, K., 2019. *Innovating Pedagogy 2019: Open University Innovation Report 7*. The Open University, Milton Keynes.

Feynman, R.P., 1974. *Cargo cult science*. Engineering and Science, 37(7), pp. 10-13.

Fike, D.S. and Fike, R., 2008. *Predictors of first-year student retention in the community college*. Community college review, 36(2), pp. 68-88.

Fišer, D., Faigl, J. and Kulich, M., 2013. *Growing neural gas efficiently*. Neurocomputing, 104, pp. 72-82.

Fleming, N.D. and Mills, C., 1992. *Not another inventory, rather a catalyst for reflection*. To improve the academy, 11(1), pp. 137-155. García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating Bayesian networks' precision for detecting students' learning styles. *Computers and Education*, *49*(3), 794–808. http://doi.org/10.1016/j.compedu.2005.11.017

Fritzke, B., 1995. *A growing neural gas network learns topologies*. In Advances in neural information processing systems (pp. 625-632).

Gajawada, S 2019, *Chi-Square Test for Feature Selection in Machine learning*, Towards Data Science, viewed 12 December 2019, https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

Garbade J 2018, *Regression Versus Classification Machine Learning: What's the Difference?*, Medium, viewed 28 August 2019, https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7

Garrison, D.R. and Arbaugh, J.B., 2007. *Researching the community of inquiry framework: Review, issues, and future directions*. The Internet and Higher Education, 10(3), pp. 157-172.

Geake, J., 2008. *Neuromythologies in education*. Educational research, 50(2), pp. 123-133.

Ghoneim, S 2018, *Fuzzy logic and how it is curing cancer*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/fuzzy-logic-and-how-it-is-curing-cancer-dc6bcc961ded

Gibbons, S., Neumayer, E. and Perkins, R., 2015. *Student satisfaction, league tables and university applications: Evidence from Britain*. Economics of Education Review, 48, pp. 148-164.

Graf, S., 2007. Adaptivity in learning management systems focussing on learning styles. PhD thesis, Athabasca University, Canada.

Graf, S., Kinshuk and Liu, T.C., 2009. *Supporting teachers in identifying students' learning styles in learning management systems: An automatic student modelling approach*. Journal of Educational Technology & Society, 12(4), pp. 3-14.

Hassanzadeh, A., Kanaani, F. and Elahi, S., 2012. *A model for measuring e-learning systems success in universities*. Expert Systems with Applications, 39(12), pp. 10959-10966.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J., 2005. *The elements of statistical learning: data*

mining, inference and prediction. The Mathematical Intelligencer, 27(2), pp. 83-85.

Heaton, J.T., 2017. *Automated Feature Engineering for Deep Neural Networks with Genetic Programming.* PhD thesis, Nova Southeastern University, Florida.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D. and Navarro-Colorado, B., 2019. *A Systematic Review of Deep Learning Approaches to Educational Data Mining*. Complexity, 2019. vol. 2019, Article ID 1306039.

Herodotou, C., Hlosta, M., Boroowa, A., Rienties, B., Zdrahal, Z. and Mangafa, C., 2019. *Empowering online teachers through predictive learning analytics*. British Journal of Educational Technology. 50(6), pp. 3064-3079.

HESA 2018a, *Non-continuation summary: UK Performance Indicators 2016/17*, viewed 28 August 2019, https://www.hesa.ac.uk/news/08-03-2018/non-continuation-summary.

HESA 2018b, *HE student enrolments by HE provider: Academic Year 2017/18*. Viewed 28 August 2019, https://www.hesa.ac.uk/data-and-analysis/students/whos-in-he.

Heuer, H. and Breiter, A., 2018. *Student Success Prediction and the Trade-Off between Big Data and Data Minimization*. DeLFI 2018-Die 16. E-learning Fachtagung Informatik.

Hindle, A 2016, *Research Point: ANOVA Explained*, Edanz, viewed 10 October 2019, https://en-author-services.edanzgroup.com/blogs/statistics-anova-explained

Hocking, A., Geach, J., Sun, Y., Davey, N. and Hine, N., 2015. *Unsupervised image analysis & galaxy categorisation in multi-wavelength Hubble space telescope images*. Proceedings of the ECMLPKDD, pp. 105-114.

Hoic-Bozic, N., Dlab, M.H. and Mornar, V., 2015. *Recommender system and web 2.0 tools to enhance a blended learning model*. IEEE Transactions on education, 59(1), pp. 39-44.

Honicke, T. and Broadbent, J., 2016. *The influence of academic self-efficacy on academic performance: A systematic review*. Educational Research Review, 17, pp. 63-84.

Horning, N., 2013. *Introduction to decision trees and random forests*. American Museum of Natural History's Center for Biodiversity and Conservation, 2, pp. 1-27.

Höver, K.M. & Steiner, C.M., 2009. *Adaptive Learning Environments: A Requirements Analysis in Business Settings*. International Journal of Advanced Corporate Learning, 2(3), pp. 27–33.

Howard-Jones, P.A., 2014. *Neuroscience and education: myths and messages.* Nature Reviews

Neuroscience, 15(12), p.817.

Husmann, P.R. and O'Loughlin, V.D., 2019. *Another nail in the coffin for learning styles? Disparities among undergraduate anatomy students' study strategies, class performance, and reported VARK learning styles*. Anatomical sciences education, 12(1), pp. 6-19.

Huxley, G., Mayo, J., Peacey, M.W. and Richardson, M., 2018. *Class size at university*. Fiscal Studies, 39(2), pp. 241-264.

Jaadi, Z 2019, *A step by step explanation of Principal Component Analysis*, Dmitry.AI, viewed 28 August 2019, https://dmitry.ai/t/topic/231

Jefferies, A. & Hyde, R., 2010. *Building the future students' blended learning experiences from current research findings*. Electronic Journal of E-learning, 8, pp. 133 – 140.

Jenkins, M., Walker, R., Voce, J., 2014. *Achieving flexibility? The rhetoric and reality of the role of learning technologies in UK Higher Education*. In Proceedings ascilite (Vol. 2014, pp. 544-548).

Jirayusakul, A. and Auwatanamongkol, S., 2007. *A supervised growing neural gas algorithm for cluster analysis*. International Journal of Hybrid Intelligent Systems, 4(4), pp. 217-229.

Jisc 2019, *Home Page*, viewed 10 October 2019, https://www.jisc.ac.uk/#

Johnson, L., Becker, S.A., Estrada, V. and Freeman, A., 2014. *NMC horizon report: 2014* (pp. 1-52). The New Media Consortium.

Johnson, L., Becker, S.A., Cummins, M., Estrada, V., Freeman, A. and Hall, C., 2016. *NMC horizon report: 2016 Higher Education edition* (pp. 1-50). The New Media Consortium.

Jolliffe, I.T. and Cadima, J., 2016. *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

Jordan, K., 2014. *Initial trends in enrolment and completion of massive open online courses*. The International Review of Research in Open and Distributed Learning, 15(1).

Kahraman, H.T., Sagiroglu, S. & Colak, I., 2013. *The development of intuitive knowledge classifier and the modeling of domain dependent data*. Knowledge-Based Systems, 37, pp. 283–295.

Kassambara, A., 2017. *Principal Component Analysis in R: prcomp vs princomp*, viewed 28 August 2019, http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

Kaur, H. and Chawla, S., 2014. *Web Data Mining: Exploring Hidden Patterns, its Types and Web Content Mining Techniques and Tools.* International Journal of Innovative Science and Modern Engineering, 3(1).

Knewton 2019, *Knewton Adaptive Learning*, viewed 28 August 2019, https://www.knewton.com/

Kolb, D.A., 1981. *Learning styles and disciplinary differences*. The modern American college, 1, pp. 232-255.

Kubat, M., 2017. *An introduction to machine learning* (Vol. 2). Cham, Switzerland: Springer International Publishing.

Kurzweil, R., 2014. *2029: the year when robots will have the power to outsmart their makers*. The Guardian UK, 22 February.

Kuzilek, J., Hlosta, M. and Zdrahal, Z., 2017. *Open university learning analytics dataset*. Scientific data, 4, p.170171.

LACE - Learning Analytics Community Exchange 2019, *Home Page*, viewed 28 August 2019, http://www.laceproject.eu/

Lang, C., Siemens, G., Wise, A. and Gasevic, D. eds., 2017. *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Lee, A 2019, *P-values Explained By Data Scientist For Data Scientists*, viewed 12 December 2019, https://towardsdatascience.com/p-values-explained-by-data-scientist-f40a746cfc8

Lesgold A., Lajole S., Bunzo M., Eggan G., 1988. *Sherlock: A Coached Practice Environment for an Electronics Troubleshooting Job*. Viewed 10 October 2019, https://apps.dtic.mil/dtic/tr/fulltext/u2/a201748.pdf

Lewthwaite, S. and Sloan, D., 2016, April. *Exploring pedagogical culture for accessibility education in computing science.* In Proceedings of the 13th Web for All Conference (p.3). ACM.

Lin, T.C., Yu, W.W.C. and Chen, Y.C., 2012. *Determinants and probability prediction of college student retention: new evidence from the Probit model*. International Journal of Education Economics and Development, *3*(3), pp. 217-236.

Liu, G. and Wang, X., 2008. *An integrated intrusion detection system by using multiple neural networks*. In 2008 IEEE Conference on Cybernetics and Intelligent Systems (pp. 22-27). IEEE.

Mampadi, F., Chen, S.Y., Ghinea, G. and Chen, M.P., 2011. *Design of adaptive hypermedia learning*

systems: A cognitive style approach. Computers & Education, 56, pp. 1003–1011.

2011. Design of adaptive hypermedia learning systems: A cognitive style approach. Computers & Education, 56(4), pp.1003-1011.

Marburger, D.R., 2001. *Absenteeism and undergraduate exam performance.* The Journal of Economic Education, 32(2), pp. 99-109.

Marcus, G., 2018. *Deep learning: A critical appraisal.* arXiv preprint arXiv:1801.00631.

Marengo, A., Pagano, A., Monopoli,G., 2015 *Adaptive System Prototype: Automated and Customised Learning Experience*, INTED2015 Proceedings, pp. 4536-4544.

Marist College 2019, *Home page*, viewed 28 August 2019, https://www.marist.edu/

Marković, M.G., Jakupović, A. and Kovačić, B., 2014. *A prevalence trend of characteristics of intelligent and adaptive hypermedia e-learning systems*. WSEAS Transactions on Advances in Engineering Education, 11, pp. 80-101.

Marr , B 2018, *What Are Artificial Neural Networks - A Simple Explanation For Absolutely Anyone*, Forbes, viewed 28 August 2019, https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/#3693c04c1245

Massa, L.J. and Mayer, R.E., 2006. *Testing the ATI hypothesis: Should multimedia instruction accommodate verbalizer-visualizer cognitive style?*. Learning and Individual Differences, 16(4), pp. 321-335.

McDonald, J.H., 2009. *Handbook of biological statistics*. Baltimore, MD: sparky house publishing, pp. 6-59.

McKenzie, K. and Schweitzer, R., 2001. *Who succeeds at university? Factors predicting academic performance in first year Australian university students.* Higher Education research & development, 20(1), pp. 21-33.

Mearman, A., Pacheco, G., Webber, D., Ivlevs, A. and Rahman, T., 2014. *Understanding student attendance in business schools: An exploratory study*. International Review of Economics Education, 17, pp. 120-136.

Mo, S., Zeng, J. and Tan, Y., 2010, September. *Particle swarm optimization based on self-organizing topology driven by fitness.* In 2010 International Conference on Computational Aspects of Social Networks (pp. 23-26). IEEE.

Natek, S. & Zwilling, M., 2014. *Student data mining solution-knowledge management system related to Higher Education institutions.* Expert Systems with Applications, 41(14), pp. 6400–6407.

National Association of Secondary School Principals (USA) and Keefe, J.W., 1979. *Student learning styles: Diagnosing and prescribing programs*. NASSP.

Naviani, A 2019, *Neural Network Models in R*, DataCamp, viewed 28 August 2019, https://www.datacamp.com/community/tutorials/neural-network-models-r

Niitsuma, H. and Okada, T., 2005, *Covariance and PCA for categorical variables*. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, (pp. 523-528)

Nguyen, Q., Huptych, M. and Rienties, B., 2018. *Using Temporal Analytics to Detect Inconsistencies between Learning Design and Student Behaviours.* Journal of Learning Analytics, 5(3), pp. 120-135.

Nkambou, R., Mizoguchi, R. and Bourdeau, J. eds., 2010. *Advances in intelligent tutoring systems* (Vol. 308). Springer Science & Business Media.

O'Donnell, E. and O'Donnell, L., 2015. *Technology-Enhanced Learning: Towards providing supports for PhD students and researchers in Higher Education*. In Handbook of research on scholarly publishing and research methods (pp. 231-251). IGI Global.

Open University 2019. *Distance Learning Courses and Adult Education - The Open University*, viewed 28 August 2019, http://www.open.ac.uk/

Open University 2017, *Open University Learning Analytics dataset*, viewed 28 August 2019, https://analyse.kmi.open.ac.uk/open_dataset

Open University 2019, *OU Analyse*, viewed 28 August 2019, https://analyse.kmi.open.ac.uk/

Open University 2014, *Policy on Ethical use of Student Data for Learning Analytics, viewed 28 August 2019*, https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf

Oppermann, A 2018, *Bayes' Theorem: The Holy Grail of Data Science*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/bayes-theorem-the-holy-grail-of-data-science-55d93315defb

Oxman, S., Wong, W. and Innovations, D.V.X., 2014. *White paper: Adaptive learning systems.* Integrated Education Solutions, pp. 6-7.

Papatheodorou, T. and Potts, D., 2016. *Pedagogy in Practice*. The Early Years Foundation Stage: Theory

and Practice, p.111.

Pardo, A. and Siemens, G., 2014. *Ethical and privacy principles for learning analytics*. British Journal of Educational Technology, *45*(3), pp. 438-450.

Parmar, B.H. and Khalpada, K., 2015. *Analysis & Survey of Different Data Mining Techniques for Predicting Student's Performance*. *ETCEE–2015*, p.64.

Parnell, A., Jones, D., Wesaw, A. and Brooks, D.C., 2018. *Institutions' Use of Data and Analytics for Student Success*. Educause Center for Analysis and Research (ECAR), 11.

Pashler, H., McDaniel, M., Rohrer, D. and Bjork, R., 2008. *Learning styles: Concepts and evidence*. Psychological science in the public interest, 9(3), pp. 105-119.

Patel, S 2017, *Chapter 2 : SVM (Support Vector Machine) - Theory*, Medium, viewed 28 August 2019, https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

Patel, N. and Upadhyay, S., 2012. *Study of various decision tree pruning methods with their empirical comparison in WEKA*. International journal of computer applications, 60(12) pp. 20–25.

Peng, J. and Li, S., 2019. *Preprocessing of categorical predictors in SVM, KNN and KDC*, LibreTexts, viewed 12 November 2019, https://stats.libretexts.org/Bookshelves/Advanced_Statistics_Computing/RTG%3A_Classification_Methods/4%3A_Numerical_Experiments_and_Real_Data_Analysis/Preprocessing_of_categorical_predictors_in_SVM%2C_KNN_and_KDC_(contributed_by_Xi_Cheng)

Perrotta, C. and Williamson, B., 2018. *The social life of Learning Analytics: cluster analysis and the 'performance'of algorithmic education*. Learning, Media and Technology, 43(1), pp. 3-16.

Potdar, K., Pardawala, T.S. and Pai, C.D., 2017. *A comparative study of categorical variable encoding techniques for neural network classifiers.* International Journal of Computer Applications, 175(4), pp.7-9.

Press, G 2016, *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*, Forbes, viewed 13 October 2019, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#b04af0a6f637

Prinsloo, P. and Slade, S., 2018. *Student Consent in Learning Analytics: The Devil in the Details?* In Learning Analytics in Higher Education (pp. 118-139). Routledge.

Priy, S, Rajput, A 2019, *Fuzzy Logic | Introduction*, GeeksforGeeks, viewed 28 August 2019, https://www.geeksforgeeks.org/fuzzy-logic-introduction/

Pupale, R 2018, *Support Vector Machines (SVM) — An Overview*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Purdue University 2019, *Home Page*, viewed 28 August 2019, https://www.purdue.edu/

Quinlan, J.R., 2014. *C4. 5: programs for machine learning*. Elsevier.

Raley, N., 2012, *Intelligent Tutoring Systems: A Literature Synthesis.*

Raj 2019, *Creating Smart — Knowledge Base Systems(KBS) using advanced NLP library*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/creating-smart-knowledge-base-systems-kbs-using-advanced-nlp-library-b5c21dfafcd1

Ramesh, V.A.M.A.N.A.N., Parkavi, P. and Ramar, K., 2013. *Predicting student performance: a statistical and data mining approach*. International journal of computer applications, 63(8), pp. 35-39.

Realizeit, 2015. *Realizeit Adaptive Learning Systems*. http://realizeitlearning.com/

Reich, J 2014, *MOOC Completion and Retention in the Context of Student Intent*, EDUCAUSE review, viewed 28 August 2019, https://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent

Reyes, J.A., 2015. *The skinny on big data in education: Learning analytics simplified*. TechTrends, 59(2), pp. 75-80.

Riener, C. and Willingham, D., 2010. *The myth of learning styles*. Change: The magazine of higher learning, *42*(5), pp. 32-35.

Rienties, B. and Rivers, B.A., 2014. *Measuring and understanding learner emotions: Evidence and prospects.* Learning Analytics Review, 1, pp. 1-28.

Rienties, B. and Toetenel, L., 2016. *The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules*. Computers in Human Behavior, *60*, pp. 333-341.

Rienties, B., Boroowa, A., Cross, S., Farrington-Flint, L., Herodotou, C., Prescott, L., Mayles, K., Olney, T., Toetenel, L. and Woodthorpe, J., 2016a. *Reviewing three case-studies of learning analytics*

interventions at the open university UK. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge LAK '16 (pp. 534-535). ACM.

Rienties, B., Boroowa, A., Cross, S., Kubiak, C., Mayles, K. and Murphy, S., 2016b. *Analytics4Action Evaluation Framework: A Review of Evidence-Based Learning Analytics Interventions at the Open University UK*. Journal of Interactive Media in Education, 2016(1): 2, pp 1–11

Rienties, B., Nguyen, Q., Holmes, W. and Reedy, K., 2017. *A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK*. Interaction Design and Architecture (s), 33, pp. 134-154.

Rodríguez-Triana, M.J., Martínez-Monés, A. and Villagrá-Sobrino, S., 2016. *Learning Analytics in Small-Scale Teacher-Led Innovations: Ethical and Data Privacy Issues.* Journal of Learning Analytics, 3(1), pp. 43-65.

Rouse, M 2018, *Knowledge-based systems (KBS)*, SearchCIO, viewed 28 August 2019, https://searchcio.techtarget.com/definition/knowledge-based-systems-KBS

Russell, S.J. and Norvig, P. 2016, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,

Sammut, C. and Webb, G.I., 2017. *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated.

Schott, M 2019, *K-Nearest Neighbors (KNN) Algorithm for Machine Learning*, Capitol One Tech, https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26

Schuler, R.S., Jackson, S.E. & Tarique, I., 2011. *Global talent management and global talent challenges: Strategic opportunities for IHRM*. Journal of World Business, 46(4), pp. 506–516.

Schweighofer, P. and Ebner, M., 2015. *Aspects to be considered when implementing technology-enhanced learning approaches: A literature review*. Future Internet, 7(1), pp. 26-49.

Schwendimann, B.A., Rodriguez-Triana, M.J., Vozniuk, A., Prieto, L.P., Boroujeni, M.S., Holzer, A., Gillet, D. and Dillenbourg, P., 2016. *Perceiving learning at a glance: A systematic literature review of learning dashboard research*. IEEE Transactions on Learning Technologies, 10(1), pp. 30-41.

Sclater, N. and Bailey, P., 2018. *Code of practice for learning analytics*. Viewed 10 October 2019, https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics

Sclater, N., Peasgood, A. and Mullan, J., 2016a. *Learning analytics in Higher Education*. Viewed 10 October 2019, https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v3.pdf

Sclater, N., Peasgood, A. and Mullan, J., 2016b. *Case study A: Traffic lights and interventions: Signals at Purdue University*. JISC, viewed 10 October 2019, https://analytics.jiscinvolve.org/wp/files/2016/04/CASE-STUDY-A-Purdue-University.pdf

Sclater, N. and Mullan, J., 2017. *Learning analytics and student success–Assessing the evidence*. Viewed 10 October 2019, http://repository.jisc.ac.uk/6560/1/learning-analytics_and_student_success.pdf

Sclater, N 2017, *Effective Learning Analytics Using data and analytics to support students Notes and presentations from the 11th Jisc Learning Analytics Network event at Aston University*, Jisc, viewed 28 August 2019, https://analytics.jiscinvolve.org/wp/2017/09/11/notes-and-presentations-from-the-11th-jisc-learning-analytics-network-event-at-aston-university/

Seif, G 2018, *A Guide to Decision Trees for Machine Learning and Data Science*, Towards Data Science, viewed 28 August 2019, https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956

Shale-Hester, T 2019, *Artificial intelligence to prevent traffic jams months in advance*, Auto Express, viewed 28 August 2019, https://www.autoexpress.co.uk/car-news/107731/artificial-intelligence-to-prevent-traffic-jams-months-in-advance

Shannon, C.E. and Weaver, W., 1998. *The mathematical theory of communication*. University of Illinois press.

Sharples, M., Adams, A., Ferguson, R., Mark, G., McAndrew, P., Rienties, B., Weller, M. and Whitelock, D., 2014. *Innovating pedagogy 2014: exploring new forms of teaching, learning and assessment, to guide educators and policy makers.* The Open University, UK.

Simpson, O., 2006. *Predicting student success in open and distance learning*. Open Learning: The Journal of Open, Distance and e-learning, 21(2), pp. 125-138.

Simpson, O., 2013. *Student retention in distance education: are we failing our students?*. Open Learning: The Journal of Open, Distance and e-learning, 2*8*(2), pp. 105-119.

Sivakumar, N. and Praveena, R., 2015. *Determining optimized learning path for an e-learning system using ant colony optimization algorithm*. International Journal of Computer Science & Engineering Technology, 6(2), pp. 61-66.

Slade, S. and Prinsloo, P., 2013. *Learning analytics: Ethical issues and dilemmas*. American Behavioral

Scientist, 57(10), pp. 1510-1529.

Slade, S. and Tait, A., 2019. *Global guidelines: Ethics in learning analytics, v*iewed 10 October 2019, https://www.learntechlib.org/p/208251/

Smith, L.I., 2002. *A tutorial on principal components analysis*. University of Otago, New Zealnd.

Soni, D 2018, *Introduction to k-Nearest-Neighbors, Towards Data Science*, viewed 28 August 2019, https://towardsdatascience.com/introduction-to-k-nearest-neighbors-3b534bb11d26

StatSoft, I., 2013. *Electronic statistics textbook*. Statsoft. Tulsa, OK, viewed 10 October 2019, https://www.freetechbooks.com/electronic-statistics-textbook-t932.html

Stubbs, M., Martin, I. and Endlar, L., 2006. *The structuration of blended learning: putting holistic design principles into practice*. British journal of educational technology, 37(2), pp. 163-175.

Tan, P.N., 2018. *Introduction to data mining*. Pearson Education India.

Tempelaar, D.T., Rienties, B. and Giesbers, B., 2015. *In search for the most informative data for feedback generation: Learning Analytics in a data-rich context*. Computers in Human Behavior, 47, pp. 157-167.

The Glossary of Educational Reform 2013, *Locus of Control*, viewed 28 August 2019, https://www.edglossary.org/locus-of-control/

The Guardian 2019, *Artificial intelligence (AI)*, viewed 28 August 2019, https://www.theguardian.com/technology/artificialintelligenceai

The Oxford Dictionary 2019, *Overview artificial intelligence*, viewed 28 August 2019, https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960

Truong, H.M., 2016. *Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities.* Computers in human behavior, *55*, pp. 1185-1193.

UK Government 1998, *Data Protection Act 1998*, viewed 12 December 2019, http://www.legislation.gov.uk/ukpga/1998/29/contents

UK Government 2015, *English indices of deprivation 2015*, viewed 28 August 2019, https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

UK Government 2018, *Guide to the General Data Protection Regulation (GDPR)*, viewed 28 August 2019, https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation

Umer, R., Susnjak, T., Mathrani, A. and Suriadi, S., 2018. *A learning analytics approach: Using online weekly student engagement data to make predictions on student performance*. *In 2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1-5). IEEE.

University of Hertfordshire 2019a, *Home Page*, viewed 28 August 2019, https://www.herts.ac.uk/

University of Hertfordshire 2019b, *Refunds*, viewed 28 August 2019, https://www.herts.ac.uk/study/fees-and-funding/refunds

Vasquez, F 2017, *Deep Learning made easy with Deep Cognition*, Medium, viewed 28 August 2019, https://becominghuman.ai/@faviovazquez

Vitiello, M., Walk, S., Helic, D., Chang, V. and Guetl, C., 2018. *User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC*. J. UCS, 24(8), pp. 1131-1150.

Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2015. *The potential for using artificial intelligence techniques to improve e-learning systems*. In ECEL 2015 Conference proceedings, pp. 762-770.

Wakelam, E., Davey, N., Sun, Y., Jefferies, A., Alva, P. and Hocking, A., 2016, May. *The Mining and Analysis of Data with Mixed Attribute Types.* In Proceedings: IMMM 2016: Sixth International Conference on Advances in Information Mining and Management. IARIA, pp 32-37.

Wakelam, E., Jefferies, A., Davey, N. and Sun, Y., 2020. *The potential for student performance prediction in small cohorts with minimal available attributes*. British Journal of Educational Technology, 51(2), pp. 347-370.

Walker, R., Voce, J., Jenkins, M., Strawbridge, F., Barrand, M., Hollinshead, L., Craik, A., Latif, F., Sherman, S., Brown, V. and Smith, N., 2018. 2018 Survey of Technology Enhanced Learning for Higher Education in the UK. Oxford: Universities and Colleges Information Systems Association (UCISA)

Wang, F. & Hannafin, M.J., 2005. *Design-based research and technology-enhanced learning environments.* Educational Technology Research and Development, 53(4), pp. 5–23.

Wang, Y.S., Wang, H.Y. and Shee, D.Y., 2007. *Measuring e-learning systems success in an organizational context: Scale development and validation.* Computers in Human Behavior, 23(4), pp. 1792-1808.

Wen, D., Graf, S., Lan, C.H., Anderson, T. and Dickson, K., 2007. *Supporting web-based learning*

through adaptive assessment. FormaMente Journal, 2(1-2), pp. 45-79.

Wilson, A., Watson, C., Thompson, T.L., Drew, V. and Doyle, S., 2017. *Learning analytics: challenges and limitations.* Teaching in Higher Education, 22(8), pp. 991-1007.

Wolff, A., Zdrahal, Z., Nikolov, A. and Pantucek, M., 2013. *Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment.* In Proceedings of the third international conference on learning analytics and knowledge LAK '13 (pp. 145-149). ACM.

Wolff, A., Zdrahal, Z., Herrmannova, D., Kuzilek, J. and Hlosta, M., 2014. *Developing predictive models for early detection of at-risk students on distance learning modules.* Open University UK.

Wong, B.T. and Li, K.C., 2018, July. *Learning analytics intervention: A review of case studies*. In 2018 International Symposium on Educational Technology (ISET) (pp. 178-182). IEEE.

Wu, J.H. and Wang, Y.M., 2006. *Measuring KMS success: A respecification of the DeLone and McLean's model*. Information & Management, 43(6), pp. 728-739.

Yau, J.Y.K., Ifenthaler, D. and Mah, D.K., 2018. *Utilizing learning analytics for study success: A systematic review*. Proceedings of EdMedia 2018 - World Conference on Educational Media and Technology

Young, J., Shaxson, L., Jones, H., Hearn, S., Datta, A. and Cassidy, C., 2014. *A guide to policy engagement and influence.* Overseas Development Institute, London.

Zhang, Z., 2016. *Introduction to machine learning: k-nearest neighbors*. Annals of translational medicine, 4(11).

Appendix A: University of Hertfordshire Researcher Development Programme (RDP) courses

- How to be an Effective Researcher
- Exploring and Organising Your Literature
- Teaching for Research Students
- Risk Management
- Registration and Doctoral Review Assessment
- Becoming a member of your chosen discipline
- Research Oriented Writing Skills
- Critical Reading
- Literature searching: Systematic searching using online resources
- Research Integrity
- Plagiarism
- Raising the Visibility of Your Research
- Turnitin
- Questionnaire Design
- What's the Story? (Poster Presentation)
- An Innocents Guide to Intellectual Property
- Getting Published and Promoting your Research
- Literature Review
- Relationships in Data
- Thesis What Thesis
- Applying for Ethical Approval for your Research Project
- Research Data Management
- The British PhD and How to Bag One
- Using the Research Information System

Appendix B: University of Hertfordshire Ethics Approval

University of Hertfordshire UH

HEALTH SCIENCE ENGINEERING & TECHNOLOGY ECDA

ETHICS APPROVAL NOTIFICATION

TO	Edward Wakelam
CC	Amanda Jefferies
FROM	Dr Simon Trainis, Health, Sciences, Engineering & Technology ECDA Chair
DATE	13th October 2017

Protocol number: cCOM/PGR/UH/02965

Title of study: *An Investigation to consider whether the analysis of university collected student data can be used to predict students at risk and therefore provide tutors with potential intervention support.*

Your application for ethics approval has been accepted and approved with the following conditions by the ECDA for your School and includes work undertaken for this study by the named additional workers below:

Approval Conditions:

The principle investigator can only use the modules he teaches on for this study.

This approval is valid:

From: 13/10/2017

To: 01/09/2018

Additional workers: no additional workers named

Please note:

Your application has been conditionally approved. You must ensure that you comply with the conditions noted above as you undertake your research. You are required to complete and submit an EC7 Protocol Monitoring Form once this study is complete. Available via the Ethics Approval StudyNet Site via the 'Application Forms' page http://www.studynet1.herts.ac.uk/ptl/common/ethics.nsf/Teaching+Documents?Openview&count=9999&restricttocategory=Application+Forms

If your research involves invasive procedures you are required to complete and submit an EC7 Protocol Monitoring Form, and your completed consent paperwork to this ECDA once your study is complete.

Failure to comply with the conditions will be considered a breach of protocol and may result in disciplinary action which could include academic penalties. Additional documentation requested as a condition of this approval protocol may be submitted via your supervisor to the Ethics Clerks as it becomes available. All documentation

relating to this study, including the information/documents noted in the conditions above, must be available for your supervisor at the time of submitting your work so that they are able to confirm that you have complied with this protocol.

Approval applies specifically to the research study/methodology and timings as detailed in your Form EC1/EC1A. Should you amend any aspect of your research, or wish to apply for an extension to your study, you will need your supervisor's approval and must complete and submit form EC2. In cases where the amendments to the original study are deemed to be substantial, a new Form EC1A may need to be completed prior to the study being undertaken.

Should adverse circumstances arise during this study such as physical reaction/harm, mental/emotional harm, intrusion of privacy or breach of confidentiality this must be reported to the approving Committee immediately. Failure to report adverse circumstance/s would be considered misconduct.

Ensure you quote the UH protocol number and the name of the approving Committee on all paperwork, including recruitment advertisements/online requests, for this study.

Students must include this Approval Notification with their submission.

Appendix C: University of Hertfordshire Refund Policy (University of Hertfordshire, 2019b)

Academic Year 2018/19

These refund and liability dates do not apply to UH Online or Research Students.

Semester A

Category of Student	Withdrawal Dates	Fee Liability	Refund
Full and part-time Home, EU, International Undergraduate and Postgraduate	Between 17th September 2018 to 30th September 2018	0% of Tuition Fees	Full Refund
	Between 1st October 2018 and 6th January 2019	25% of Tuition Fees	25% of full tuition fees paid
	Between 7th January 2019 and 27th April 2019	50% of Tuition Fees	50% of full tuition fees paid
	On or after the 28th April 2019	100% of Tuition Fees	No Refund

Semester B

Category of Student	Withdrawal Dates	Fee Liability	Refund
Full and part-time Home, EU, International Undergraduate and Postgraduate	Between 14th January 2019 to 24th February 2019	0% of Tuition Fees	Full Refund
	Between 25th February 2019 and 11th April 2019	25% of Tuition Fees	25% of full tuition fees paid
	Between 12th April 2019 and 16th May 2019	50% of Tuition Fees	50% of full tuition fees paid*
	On or after 17th May 2019	100% of Tuition Fees	No Refund

Semester C

Category of Student	Withdrawal Dates	Fee Liability	Refund
Full and part-time Home, EU, *International Undergraduate and Postgraduate	Between 20th May 2019 to 30th May 2019	0% of Tuition Fees	Full Refund
	Between 31st May 2019 and 3rd October 2019	25% of Tuition Fees	25% of full tuition fees paid
	Between 4th October 2019 and 2nd January 2020	50% of Tuition Fees	50% of full tuition fees paid
	On or after the 3rd January 2020	100% of Tuition Fees	No Refund

Appendix D: Students' Knowledge Levels on DC Electrical Machines Dataset (Kahraman et al. 2013)

STG	SCG	STR	LPR	PEG	UNS
0	0	0	0	0	very_low
0.08	0.08	0.1	0.24	0.9	**High**
0.06	0.06	0.05	0.25	0.33	Low
0.1	0.1	0.15	0.65	0.3	Middle
0.08	0.08	0.08	0.98	0.24	Low
0.09	0.15	0.4	0.1	0.66	Middle
0.1	0.1	0.43	0.29	0.56	Middle
0.15	0.02	0.34	0.4	0.01	very_low
0.2	0.14	0.35	0.72	0.25	Low
0	0	0.5	0.2	0.85	**High**
0.18	0.18	0.55	0.3	0.81	**High**
0.06	0.06	0.51	0.41	0.3	Low
0.1	0.1	0.52	0.78	0.34	Middle
0.1	0.1	0.7	0.15	0.9	**High**
0.2	0.2	0.7	0.3	0.6	Middle
0.12	0.12	0.75	0.35	0.8	**High**
0.05	0.07	0.7	0.01	0.05	very_low
0.1	0.25	0.1	0.08	0.33	Low
0.15	0.32	0.05	0.27	0.29	Low
0.2	0.29	0.25	0.49	0.56	Middle
0.12	0.28	0.2	0.78	0.2	Low
0.18	0.3	0.37	0.12	0.66	Middle
0.1	0.27	0.31	0.29	0.65	Middle
0.18	0.31	0.32	0.42	0.28	Low
0.06	0.29	0.35	0.76	0.25	Low
0.09	0.3	0.68	0.18	0.85	**High**
0.04	0.28	0.55	0.25	0.1	very_low
0.09	0.255	0.6	0.45	0.25	Low

0.08	0.325	0.62	0.94	0.56	**High**
0.15	0.275	0.8	0.21	0.81	**High**
0.12	0.245	0.75	0.31	0.59	Middle
0.15	0.295	0.75	0.65	0.24	Low
0.1	0.256	0.7	0.76	0.16	Low
0.18	0.32	0.04	0.19	0.82	**High**
0.2	0.45	0.28	0.31	0.78	**High**
0.06	0.35	0.12	0.43	0.29	Low
0.1	0.42	0.22	0.72	0.26	Low
0.18	0.4	0.32	0.08	0.33	Low
0.09	0.33	0.31	0.26	0	very_low
0.19	0.38	0.38	0.49	0.45	Middle
0.02	0.33	0.36	0.76	0.1	Low
0.2	0.49	0.6	0.2	0.78	**High**
0.14	0.49	0.55	0.29	0.6	Middle
0.18	0.33	0.61	0.64	0.25	Middle
0.115	0.35	0.65	0.27	0.04	very_low
0.17	0.36	0.8	0.14	0.66	Middle
0.1	0.39	0.75	0.31	0.62	Middle
0.13	0.39	0.85	0.38	0.77	**High**
0.18	0.34	0.71	0.71	0.9	**High**
0.09	0.51	0.02	0.18	0.67	Middle
0.06	0.5	0.09	0.28	0.25	Low
0.23	0.7	0.19	0.51	0.45	Middle
0.09	0.55	0.12	0.78	0.05	Low
0.24	0.75	0.32	0.18	0.86	**High**
0.18	0.72	0.37	0.29	0.55	Middle
0.1	0.6	0.33	0.42	0.26	Low
0.2	0.52	0.36	0.84	0.25	Middle
0.09	0.6	0.66	0.19	0.59	Middle
0.18	0.51	0.58	0.33	0.82	**High**
0.08	0.58	0.6	0.64	0.1	Low

0.09	0.61	0.53	0.75	0.01	Low
0.06	0.77	0.72	0.19	0.56	Middle
0.15	0.79	0.78	0.3	0.51	Middle
0.2	0.68	0.73	0.48	0.28	Low
0.24	0.58	0.76	0.8	0.28	Middle
0.25	0.1	0.03	0.09	0.15	very_low
0.32	0.2	0.06	0.26	0.24	very_low
0.29	0.06	0.19	0.55	0.51	Middle
0.28	0.1	0.12	0.28	0.32	Low
0.3	0.08	0.4	0.02	0.67	Middle
0.27	0.12	0.37	0.29	0.58	Middle
0.31	0.1	0.41	0.42	0.75	**High**
0.29	0.15	0.33	0.66	0.08	very_low
0.3	0.2	0.52	0.3	0.53	Middle
0.28	0.16	0.69	0.33	0.78	**High**
0.255	0.18	0.5	0.4	0.1	very_low
0.265	0.06	0.57	0.75	0.1	Low
0.275	0.1	0.72	0.1	0.3	Low
0.245	0.1	0.71	0.26	0.2	very_low
0.295	0.2	0.86	0.44	0.28	Low
0.32	0.12	0.79	0.76	0.24	Low
0.295	0.25	0.26	0.12	0.67	Middle
0.315	0.32	0.29	0.29	0.62	Middle
0.25	0.29	0.15	0.48	0.26	Low
0.27	0.1	0.1	0.7	0.25	Low
0.248	0.3	0.31	0.2	0.03	very_low
0.325	0.25	0.38	0.31	0.79	**High**
0.27	0.31	0.32	0.41	0.28	Low
0.29	0.29	0.4	0.78	0.18	Low
0.29	0.3	0.52	0.09	0.67	Middle
0.258	0.28	0.64	0.29	0.56	Middle
0.32	0.255	0.55	0.78	0.34	Middle

0.251	0.265	0.57	0.6	0.09	very_low
0.288	0.31	0.79	0.23	0.24	Low
0.323	0.32	0.89	0.32	0.8	**High**
0.255	0.305	0.86	0.62	0.15	Low
0.295	0.25	0.73	0.77	0.19	Low
0.258	0.25	0.295	0.33	0.77	**High**
0.29	0.25	0.29	0.29	0.57	Middle
0.243	0.27	0.08	0.42	0.29	Low
0.27	0.28	0.18	0.48	0.26	Low
0.299	0.32	0.31	0.33	0.87	**High**
0.3	0.27	0.31	0.31	0.54	Middle
0.245	0.26	0.38	0.49	0.27	Low
0.295	0.29	0.31	0.76	0.1	Low
0.29	0.3	0.56	0.25	0.67	Middle
0.26	0.28	0.6	0.29	0.59	Middle
0.305	0.255	0.63	0.4	0.54	Middle
0.32	0.27	0.52	0.81	0.3	Middle
0.299	0.295	0.8	0.37	0.84	**High**
0.276	0.255	0.81	0.27	0.33	Low
0.258	0.31	0.88	0.4	0.3	Low
0.32	0.28	0.72	0.89	0.58	**High**
0.329	0.55	0.02	0.4	0.79	**High**
0.295	0.59	0.29	0.31	0.55	Middle
0.285	0.64	0.18	0.61	0.45	Middle
0.265	0.6	0.28	0.66	0.07	very_low
0.315	0.69	0.28	0.8	0.7	**High**
0.28	0.78	0.44	0.17	0.66	Middle
0.325	0.61	0.46	0.32	0.81	**High**
0.28	0.65	0.4	0.65	0.13	Low
0.255	0.75	0.35	0.72	0.25	Low
0.305	0.55	0.5	0.11	0.333	Low
0.3	0.85	0.54	0.25	0.83	Middle

0.325	0.9	0.52	0.49	0.76	**High**
0.312	0.8	0.67	0.92	0.5	**High**
0.299	0.7	0.95	0.22	0.66	**High**
0.265	0.76	0.8	0.28	0.28	Low
0.255	0.72	0.72	0.63	0.14	Low
0.295	0.6	0.72	0.88	0.28	Middle
0.39	0.05	0.02	0.06	0.34	Low
0.4	0.18	0.26	0.26	0.67	Middle
0.45	0.04	0.18	0.55	0.07	very_low
0.48	0.12	0.28	0.7	0.71	**High**
0.4	0.12	0.41	0.1	0.65	Middle
0.41	0.18	0.33	0.31	0.5	Middle
0.38	0.1	0.4	0.48	0.26	Low
0.37	0.06	0.32	0.78	0.1	Low
0.41	0.09	0.58	0.18	0.58	Middle
0.38	0.01	0.53	0.27	0.3	Low
0.33	0.04	0.5	0.55	0.1	very_low
0.42	0.15	0.66	0.78	0.4	Middle
0.44	0.08	0.8	0.22	0.56	Middle
0.39	0.15	0.81	0.22	0.29	Low
0.42	0.21	0.87	0.56	0.48	Middle
0.46	0.2	0.76	0.95	0.65	**High**
0.365	0.243	0.19	0.24	0.35	Low
0.33	0.27	0.2	0.33	0.1	very_low
0.345	0.299	0.1	0.64	0.13	Low
0.48	0.3	0.15	0.65	0.77	**High**
0.49	0.245	0.38	0.14	0.86	**High**
0.334	0.295	0.33	0.32	0.3	Low
0.36	0.29	0.37	0.48	0.13	very_low
0.39	0.26	0.39	0.77	0.14	Low
0.43	0.305	0.51	0.09	0.64	Middle
0.44	0.32	0.55	0.33	0.52	Middle

0.45	0.299	0.63	0.36	0.51	Middle
0.495	0.276	0.58	0.77	0.83	**High**
0.465	0.258	0.73	0.18	0.59	Middle
0.475	0.32	0.79	0.31	0.54	Middle
0.348	0.329	0.83	0.61	0.18	Low
0.385	0.26	0.76	0.84	0.3	Middle
0.445	0.39	0.02	0.24	0.88	**High**
0.43	0.45	0.27	0.27	0.89	**High**
0.33	0.34	0.1	0.49	0.12	very_low
0.4	0.33	0.12	0.3	0.9	**High**
0.34	0.4	0.38	0.2	0.61	Middle
0.38	0.36	0.46	0.49	0.78	**High**
0.35	0.38	0.32	0.6	0.16	Low
0.41	0.49	0.34	0.21	0.92	**High**
0.42	0.36	0.63	0.04	0.25	Low
0.43	0.38	0.62	0.33	0.49	Middle
0.44	0.33	0.59	0.53	0.85	**High**
0.4	0.42	0.58	0.75	0.16	Low
0.46	0.44	0.89	0.12	0.66	Middle
0.38	0.39	0.79	0.33	0.3	Low
0.39	0.42	0.83	0.65	0.19	Low
0.49	0.34	0.88	0.75	0.71	**High**
0.46	0.64	0.22	0.22	0.6	Middle
0.44	0.55	0.11	0.26	0.83	**High**
0.365	0.68	0.1	0.63	0.18	Low
0.45	0.65	0.19	0.99	0.55	**High**
0.46	0.78	0.38	0.24	0.89	**High**
0.37	0.55	0.41	0.29	0.3	Low
0.38	0.59	0.31	0.62	0.2	Low
0.49	0.64	0.34	0.78	0.21	Low
0.495	0.82	0.67	0.01	0.93	**High**
0.44	0.69	0.61	0.29	0.57	Middle

0.365	0.57	0.59	0.55	0.25	Low
0.49	0.9	0.52	0.9	0.47	**High**
0.445	0.7	0.82	0.16	0.64	Middle
0.42	0.7	0.72	0.3	0.8	**High**
0.37	0.6	0.77	0.4	0.5	Middle
0.4	0.61	0.71	0.88	0.67	**High**
0.6	0.14	0.22	0.11	0.66	Middle
0.55	0.1	0.27	0.25	0.29	Low
0.68	0.19	0.19	0.48	0.1	very_low
0.73	0.2	0.07	0.72	0.26	Low
0.78	0.15	0.38	0.18	0.63	Middle
0.55	0.1	0.34	0.3	0.1	very_low
0.59	0.18	0.31	0.55	0.09	very_low
0.64	0.09	0.33	0.65	0.5	Middle
0.6	0.19	0.55	0.08	0.1	very_low
0.69	0.02	0.62	0.3	0.29	Low
0.78	0.21	0.68	0.65	0.75	**High**
0.62	0.14	0.52	0.81	0.15	Low
0.7	0.18	0.88	0.09	0.66	Middle
0.75	0.015	0.78	0.31	0.53	Middle
0.55	0.17	0.71	0.48	0.11	very_low
0.85	0.05	0.91	0.8	0.68	**High**
0.78	0.27	0.13	0.14	0.62	Middle
0.8	0.29	0.06	0.31	0.51	Middle
0.9	0.26	0.19	0.58	0.79	**High**
0.76	0.258	0.07	0.83	0.34	Middle
0.72	0.32	0.48	0.2	0.6	Middle
0.6	0.251	0.39	0.29	0.3	Low
0.52	0.288	0.32	0.5	0.3	Low
0.6	0.31	0.31	0.87	0.58	**High**
0.51	0.255	0.55	0.17	0.64	Middle
0.58	0.295	0.62	0.28	0.3	Low

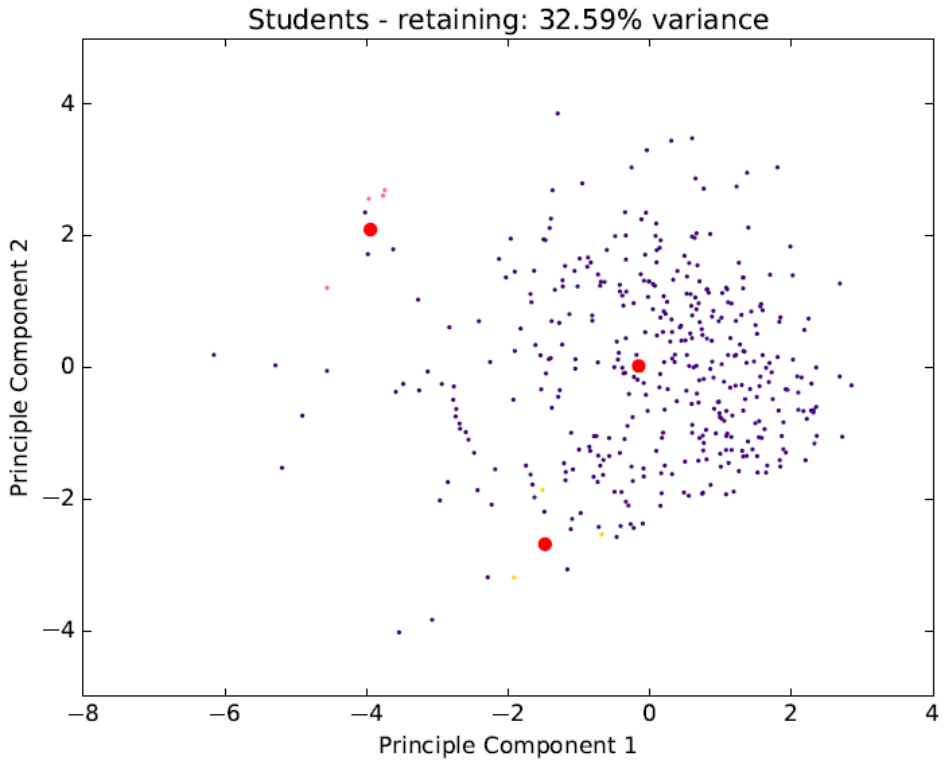
0.61	0.258	0.56	0.62	0.24	Low
0.77	0.267	0.59	0.78	0.28	Middle
0.79	0.28	0.88	0.2	0.66	Middle
0.68	0.27	0.78	0.31	0.57	Middle
0.58	0.299	0.73	0.63	0.21	Low
0.77	0.29	0.74	0.82	0.68	**High**
0.71	0.475	0.13	0.23	0.59	Middle
0.58	0.348	0.06	0.29	0.31	Low
0.88	0.335	0.19	0.55	0.78	**High**
0.99	0.49	0.07	0.7	0.69	**High**
0.73	0.43	0.32	0.12	0.65	Middle
0.61	0.33	0.36	0.28	0.28	Low
0.51	0.4	0.4	0.59	0.23	Low
0.83	0.44	0.49	0.91	0.66	**High**
0.66	0.38	0.55	0.15	0.62	Middle
0.58	0.35	0.51	0.27	0.3	Low
0.523	0.41	0.55	0.6	0.22	Low
0.66	0.36	0.56	0.4	0.83	**High**
0.62	0.37	0.81	0.13	0.64	Middle
0.52	0.44	0.82	0.3	0.52	Middle
0.5	0.4	0.73	0.62	0.2	Low
0.71	0.46	0.95	0.78	0.86	**High**
0.64	0.55	0.15	0.18	0.63	Middle
0.52	0.85	0.06	0.27	0.25	Low
0.62	0.62	0.24	0.65	0.25	Middle
0.91	0.58	0.26	0.89	0.88	**High**
0.62	0.67	0.39	0.1	0.66	Middle
0.58	0.58	0.31	0.29	0.29	Low
0.89	0.68	0.49	0.65	0.9	**High**
0.72	0.6	0.45	0.79	0.45	Middle
0.68	0.63	0.65	0.09	0.66	Middle
0.56	0.6	0.6	0.31	0.5	Middle

0.54	0.51	0.55	0.64	0.19	Low
0.61	0.78	0.69	0.92	0.58	**High**
0.78	0.61	0.71	0.19	0.6	Middle
0.54	0.82	0.71	0.29	0.77	**High**
0.5	0.75	0.81	0.61	0.26	Middle
0.66	0.9	0.76	0.87	0.74	**High**

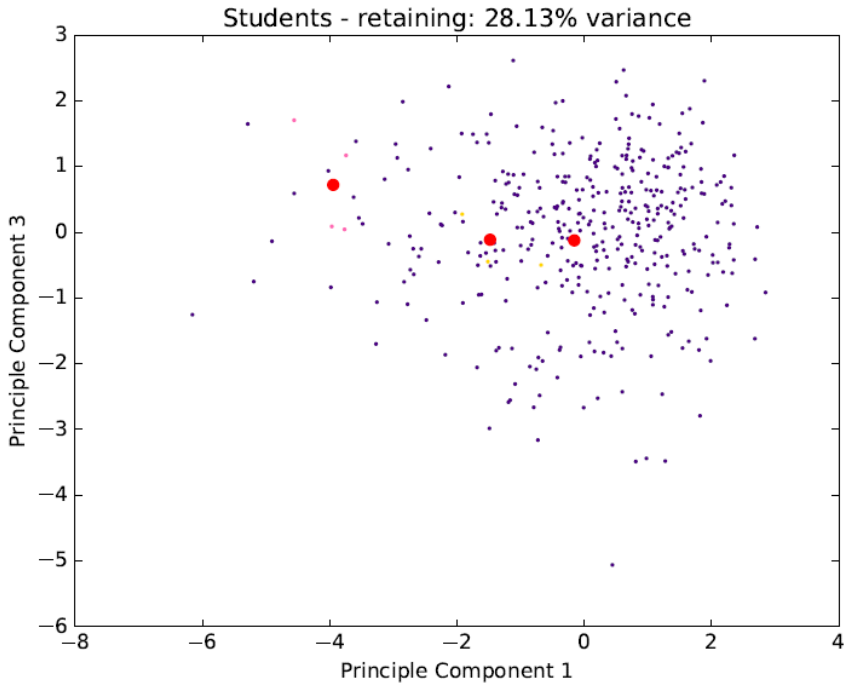
Appendix E: Portuguese Student Dataset Full Analyses

Mathematics Students: GNG (Normalised data) Principal Components Analysis

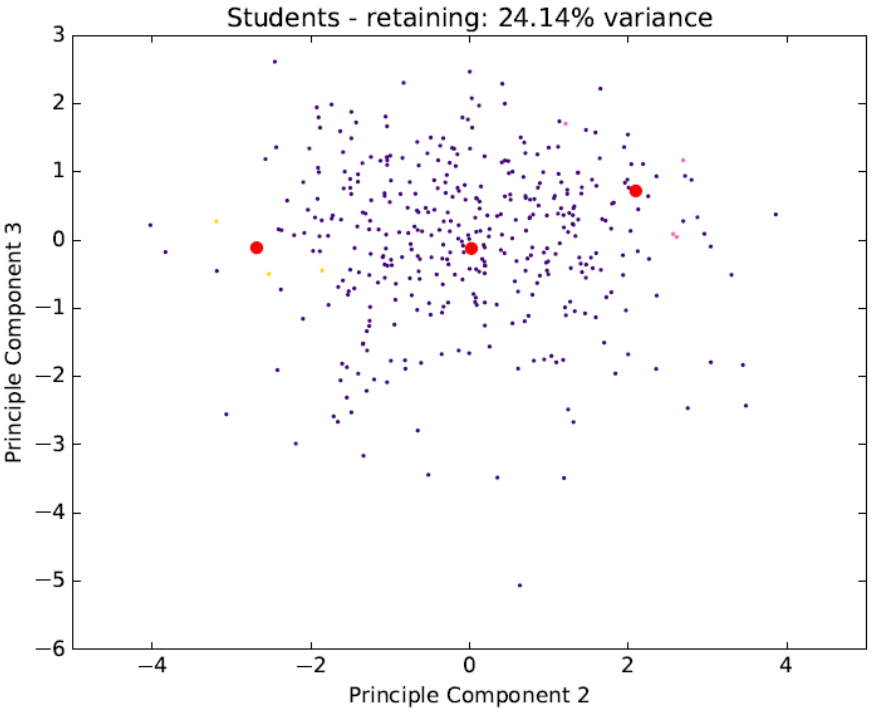
Scatter Plot PC1 v PC2



Scatter Plot PC1 v PC3

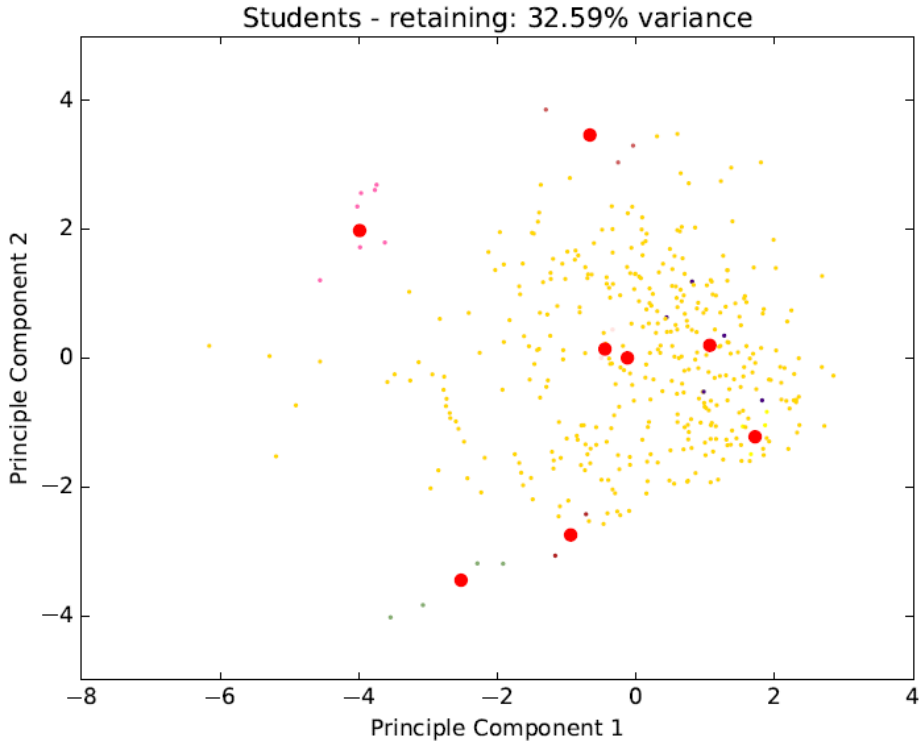


Scatter Plot PC2 v PC2

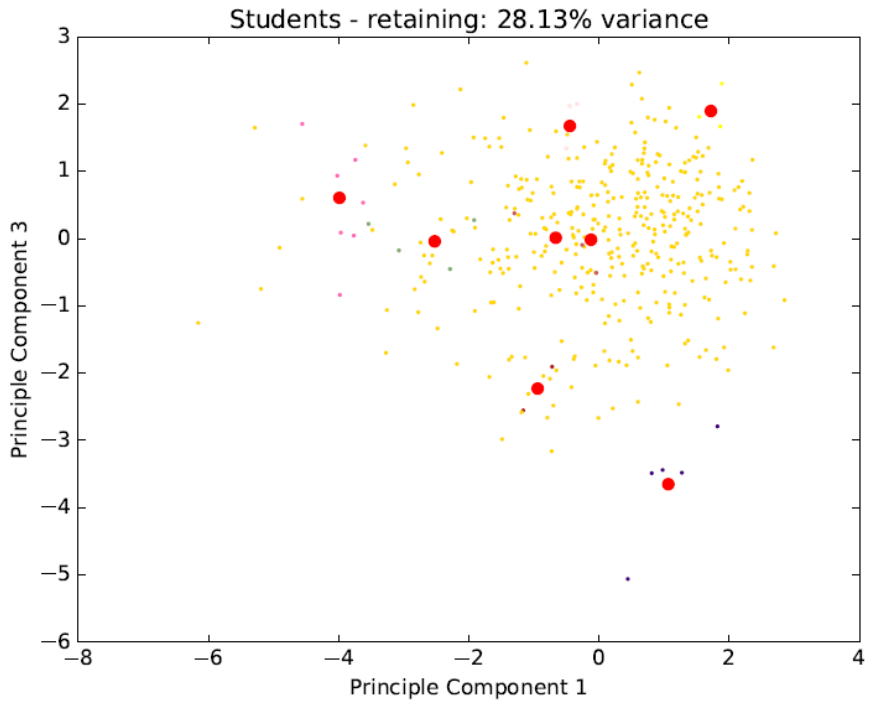


Mathematics Students: GNG (PCA data) Principal Components Analysis

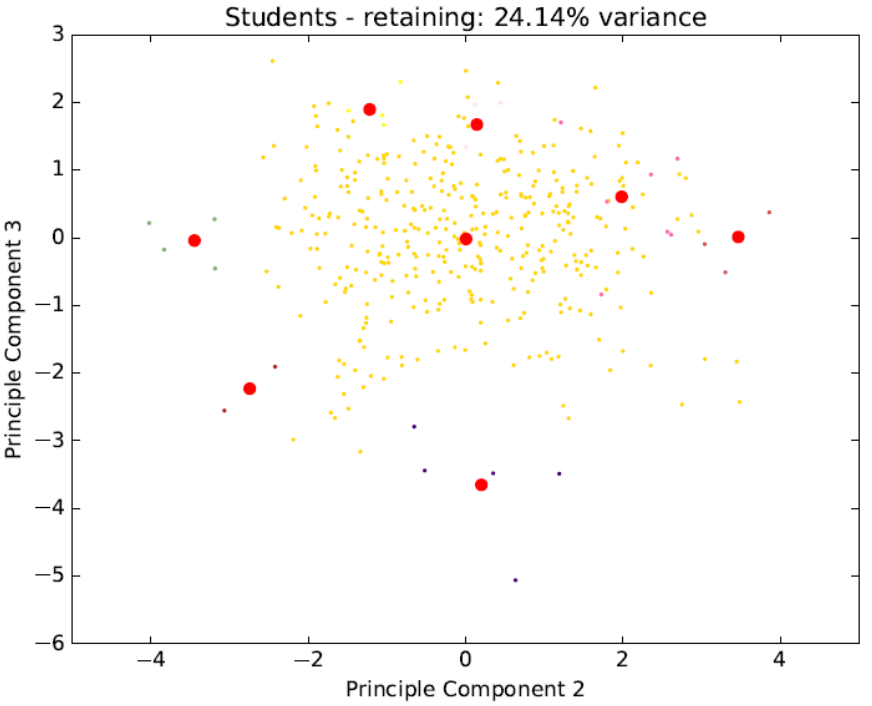
Scatter Plot PC1 v PC2



Scatter Plot PC1 v PC3

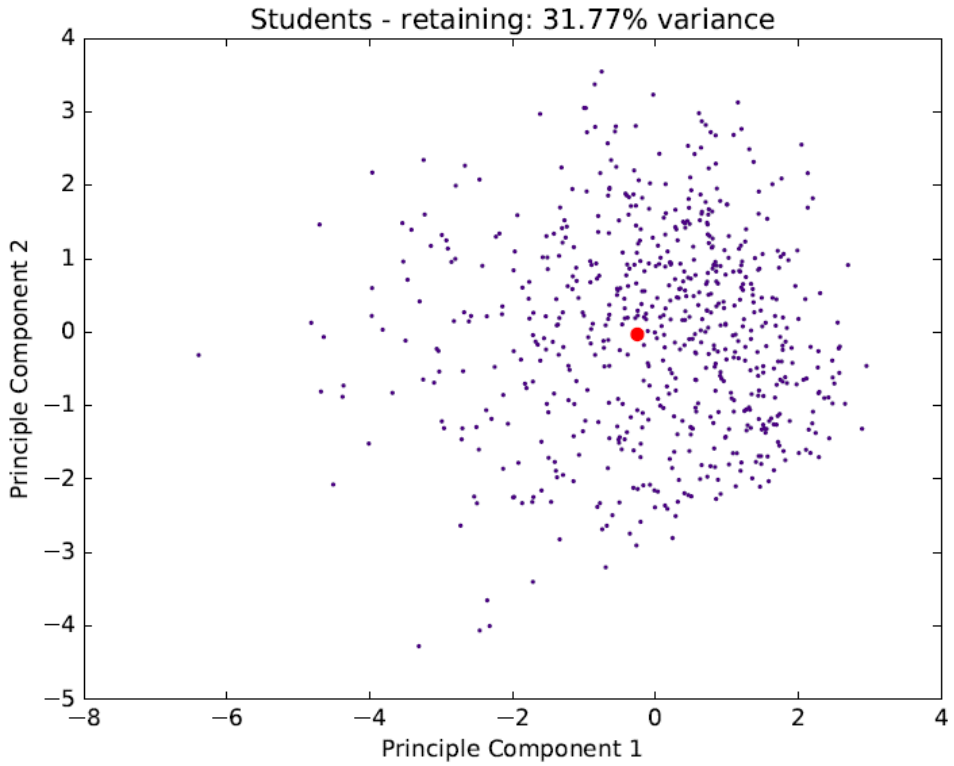


Scatter Plot PC2 v PC2

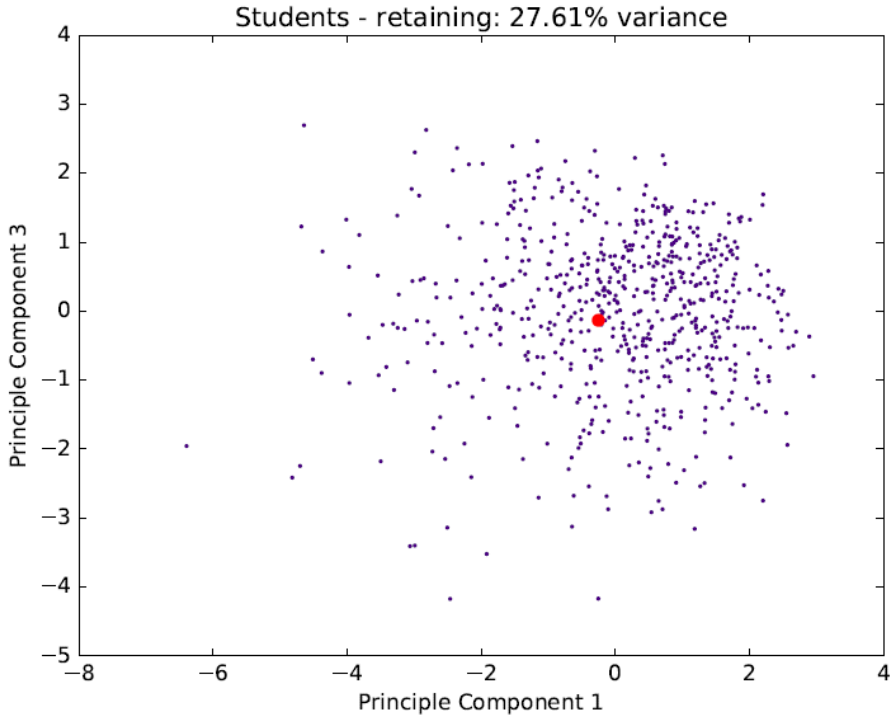


Language Students: GNG (Normailsed data) Principal Components Analysis

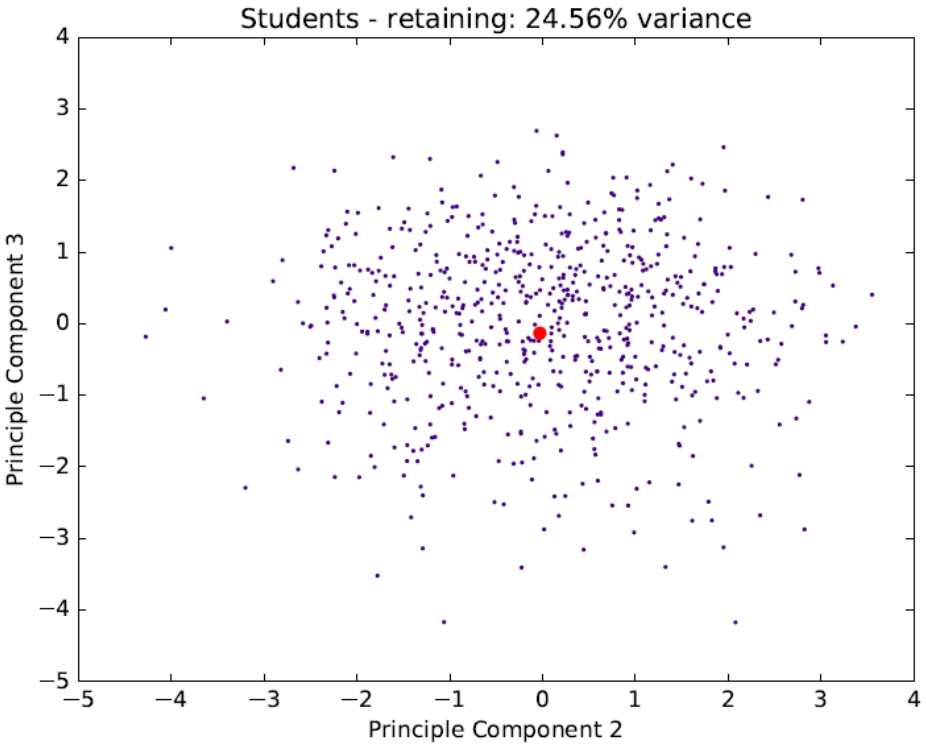
Scatter Plot PC1 v PC2



Scatter Plot PC1 v PC3

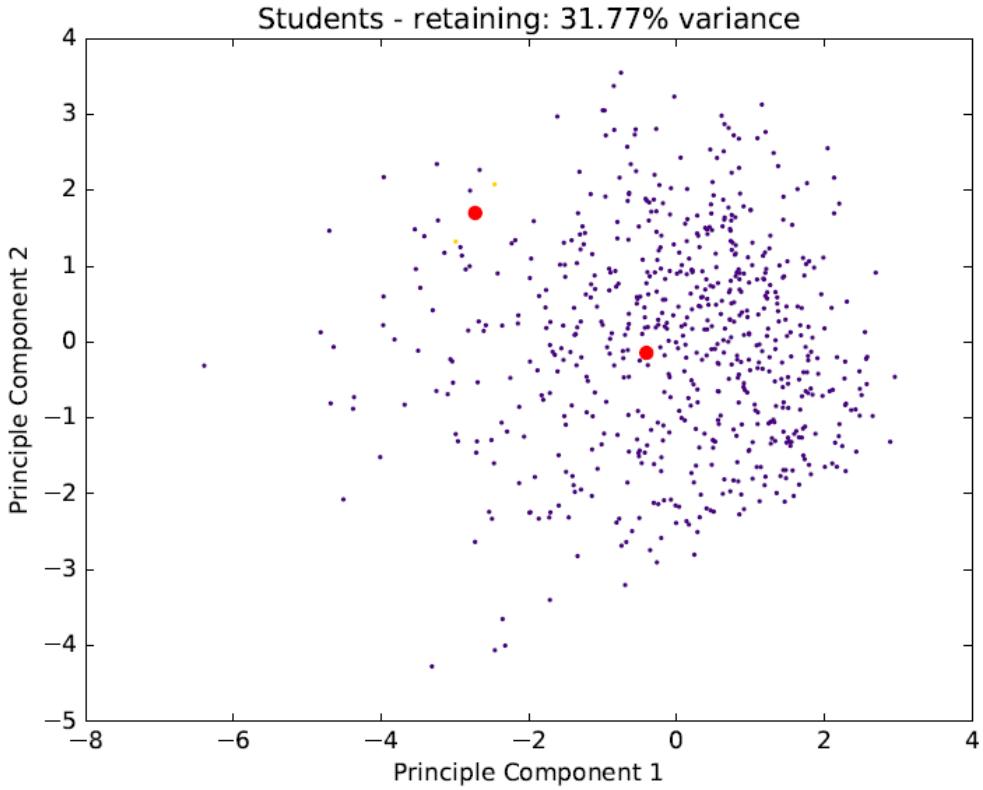


Scatter Plot PC2 v PC2

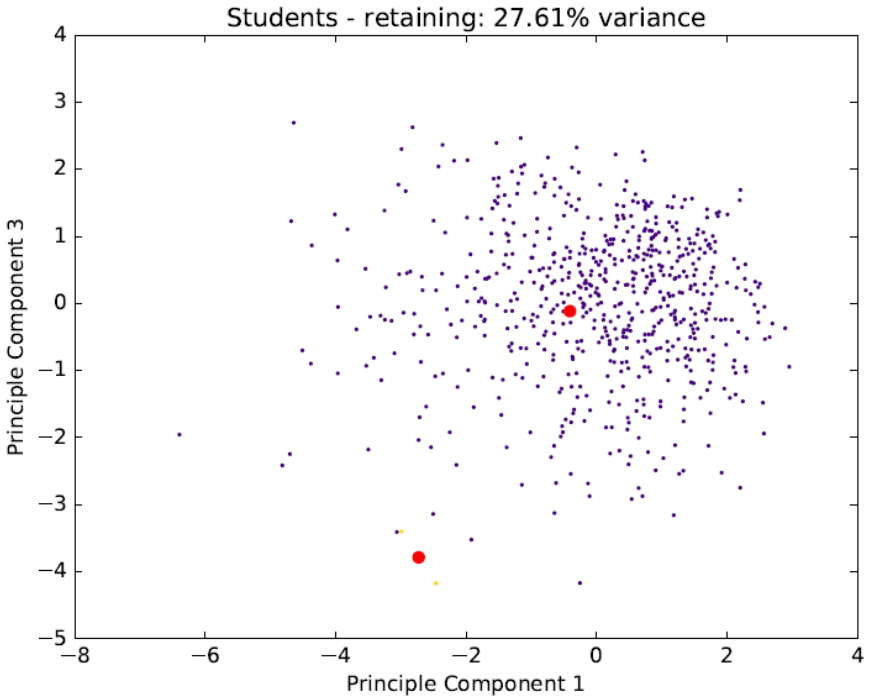


Language Students: GNG (PCA data) Principal Components Analysis

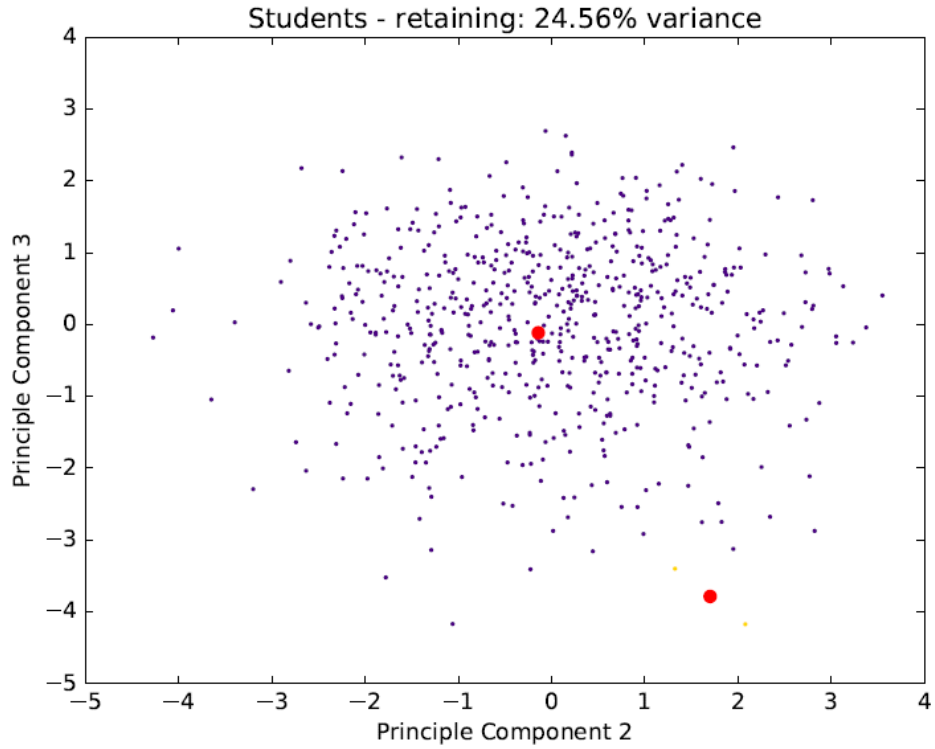
Scatter Plot PC1 v PC2



Scatter Plot PC1 v PC3



Scatter Plot PC2 v PC2



Appendix F: Portuguese Student Dataset Attribute Chi-square Analyses

Mathematics Students

Chi Square Value	School	Gender	Address	Famsize	Pstatus	Mjob	Fjob	Reason	Guardian
School		0.059619	30.92309	1.661988	0.833034	4.110855	8.411387	12.46533	2.330878
Gender	0.059619		0.320933	3.189682	0.217078	17.48356	4.747741	4.939352	2.125939
Address	30.92309	0.320933		2.074623	0.715886	13.24533	2.131105	9.648347	2.67488
Famsize	1.661988	3.189682	2.074623		8.841532	3.495522	3.38038	0.430565	0.370875
Pstatus	0.833034	0.217078	0.715886	8.841532		3.344336	4.682709	0.525325	4.883143
Mjob	4.110855	17.48356	13.24533	3.495522	3.344336		73.3809	22.94628	13.59872
Fjob	8.411387	4.747741	2.131105	3.38038	4.682709	73.3809		15.84185	17.65457
Reason	12.46533	4.939352	9.648347	0.430565	0.525325	22.94628	15.84185		4.510204
Guardian	2.330878	2.125939	2.67488	0.370875	4.883143	13.59872	17.65457	4.510204	
Schoolsup	7.718648	7.551948	0.241219	0.324039	0.704695	5.994006	5.984691	0.312507	1.824446
Famsup	10.74957	9.08082	0.22569	5.034214	0.143598	9.585135	6.772568	6.670319	0.079568
Paid	0.115276	6.586003	1.101183	0.076122	0.851711	12.83429	3.516168	12.05219	1.798817
Activities	5.402126	3.936832	1.041956	5.06E-06	3.743599	7.215537	2.163945	6.949786	0.717322
Nursery	3.148272	0.026578	1.402566	4.116711	3.243343	9.302842	5.924464	1.666443	12.05765
Higher	0.23038	9.013019	0.725395	0.013317	0.65543	8.848236	3.636967	10.23743	0.167847
Internet	7.048016	0.76865	18.57315	0.000205	1.939594	28.86134	4.284409	2.445501	1.400999
Romantic	1.45538	4.111434	0.010917	0.467286	0.646327	2.357142	2.248781	3.042658	6.229845

Mathematics Students (Continued)

Chi Square Value	Schoolsup	Famsup	Paid	Activities	Nursery	Higher	Internet	Romantic
School	7.718648	10.74957	0.115276	5.402126	3.148272	0.23038	7.048016	1.45538
Gender	7.551948	9.08082	6.586003	3.936832	0.026578	9.013019	0.76865	4.111434
Address	0.241219	0.22569	1.101183	1.041956	1.402566	0.725395	18.57315	0.010917
Famsize	0.324039	5.034214	0.076122	5.06E-06	4.116711	0.013317	0.000205	0.467286
Pstatus	0.704695	0.143598	0.851711	3.743599	3.243343	0.65543	1.939594	0.646327
Mjob	5.994006	9.585135	12.83429	7.215537	9.302842	8.848236	28.86134	2.357142
Fjob	5.984691	6.772568	3.516168	2.163945	5.924464	3.636967	4.284409	2.248781
Reason	0.312507	6.670319	12.05219	6.949786	1.666443	10.23743	2.445501	3.042658
Guardian	1.824446	0.079568	1.798817	0.717322	12.05765	0.167847	1.400999	6.229845
Schoolsup		4.328459	0.170126	0.836997	0.834627	1.172647	0.037038	2.57346
Famsup	4.328459		33.95304	0.000889	1.40008	4.014649	4.237975	0.061127
Paid	0.170126	33.95304		0.180596	4.121138	14.14174	9.262462	0.012105
Activities	0.836997	0.000889	0.180596		0.002946	3.677104	0.935375	0.152527
Nursery	0.834627	1.40008	4.121138	0.002946		1.164779	0.024215	0.298605
Higher	1.172647	4.014649	14.14174	3.677104	1.164779		0.163962	4.410166
Internet	0.037038	4.237975	9.262462	0.935375	0.024215	0.163962		2.998126
Romantic	2.57346	0.061127	0.012105	0.152527	0.298605	4.410166	2.998126	

Portuguese Language Students

Chi Square Value	School	Gender	Address	Famsize	Pstatus	Mjob	Fjob	Reason	Guardian
School		4.476302	44.68353	0.321356	0.513194	37.87644	21.36587	52.52394	2.902137
Gender	4.476302		0.422099	6.259082	2.716758	18.29724	4.381946	3.706348	1.004701
Address	44.68353	0.422099		1.380038	5.812352	27.22644	5.379133	19.25855	0.805415
Famsize	0.321356	6.259082	1.380038		37.26052	2.576552	4.921267	2.909908	0.259848
Pstatus	0.513194	2.716758	5.812352	37.26052		2.601546	6.19235	3.172359	18.9528
Mjob	37.87644	18.29724	27.22644	2.576552	2.601546		134.3821	29.70167	23.90056
Fjob	21.36587	4.381946	5.379133	4.921267	6.19235	134.3821		20.1095	20.93092
Reason	52.52394	3.706348	19.25855	2.909908	3.172359	29.70167	20.1095		5.657144
Guardian	2.902137	1.004701	0.805415	0.259848	18.9528	23.90056	20.93092	5.657144	
Schoolsup	9.873003	8.025522	0.209245	2.064809	0.058029	5.754457	7.799878	3.473028	1.116558
Famsup	2.635098	10.87828	0.020186	1.029016	0.067556	7.84983	7.82029	6.460178	1.801887
Paid	0.040559	4.081215	0.603061	1.638991	0.164555	1.503778	0.918204	6.508251	3.329578
Activities	5.095086	10.09316	0.055864	0.141958	6.693433	11.12185	3.071918	18.64628	1.161284
Nursery	0.014088	1.233879	0.212077	6.579405	0.695002	7.574448	2.869689	1.661415	6.776574
Higher	12.02375	2.193326	3.818601	0.013275	0.335204	25.84951	8.851248	10.42579	25.47137
Internet	37.53393	2.819415	20.05642	0.115793	2.317283	59.58003	8.17804	14.35037	0.53832
Romantic	3.387014	7.873406	0.621232	0.704018	1.880436	4.925732	1.182811	3.36457	11.85505

Portuguese Language Students (Continued)

Chi Square Value	Schoolsup	Famsup	Paid	Activities	Nursery	Higher	Internet	Romantic
School	9.873003	2.635098	0.040559	5.095086	0.014088	12.02375	37.53393	3.387014
Gender	8.025522	10.87828	4.081215	10.09316	1.233879	2.193326	2.819415	7.873406
Address	0.209245	0.020186	0.603061	0.055864	0.212077	3.818601	20.05642	0.621232
Famsize	2.064809	1.029016	1.638991	0.141958	6.579405	0.013275	0.115793	0.704018
Pstatus	0.058029	0.067556	0.164555	6.693433	0.695002	0.335204	2.317283	1.880436
Mjob	5.754457	7.84983	1.503778	11.12185	7.574448	25.84951	59.58003	4.925732
Fjob	7.799878	7.82029	0.918204	3.071918	2.869689	8.851248	8.17804	1.182811
Reason	3.473028	6.460178	6.508251	18.64628	1.661415	10.42579	14.35037	3.36457
Guardian	1.116558	1.801887	3.329578	1.161284	6.776574	25.47137	0.53832	11.85505
Schoolsup		3.689858	1.065144	0.593707	0.206682	4.728313	0.436769	5.772422
Famsup	3.689858		5.770886	0.035854	0.501544	4.726664	3.354263	0.355311
Paid	1.065144	5.770886		2.808336	0.493182	0.377328	0.657255	0.217557
Activities	0.593707	0.035854	2.808336		1.023872	1.308849	4.403863	2.146998
Nursery	0.206682	0.501544	0.493182	1.023872		1.17804	0.033266	0.342852
Higher	4.728313	4.726664	0.377328	1.308849	1.17804		3.211522	6.410989
Internet	0.436769	3.354263	0.657255	4.403863	0.033266	3.211522		0.787407
Romantic	5.772422	0.355311	0.217557	2.146998	0.342852	6.410989	0.787407	

Appendix G: University of Hertfordshire, Strategic IT Management Module Full Analysis

Analysis Technique	Prediction Success Measure	EVS1	EVS2	EVS3	Gp Presn	Indiv Rep	Module Result	Ave Error	Ave Success
Decision Tree Regression	Relative % Error	28%	67%	43%	26%	36%	10%	35%	65%
	Mean Squared Error	0.0767	0.1489	0.1051	0.0411	0.0603	0.0137		
	Correlation Coefficient	-0.0912	-0.4518	0.0706	-0.0224	0.1732	0.7386		
Decision Tree Classification	Relative % Error	35%	17%	61%	9%	26%	4%	25%	75%
	Mean Squared Error	0.1071	0.1739	0.6087	0.087	0.2609	0.043		
	Correlation Coefficient	-0.249	0.2655	-0.3367	-0.0455	-0.15	0.6908		
Decision Tree Regression (Combined StudyNet Clicks)	Relative % Error	23%	68%	43%	4%	31%	12%	30%	70%
	Mean Squared Error	0.0459	0.1435	0.1019	0.0127	0.0603	0.0158		
	Correlation Coefficient	0.2754	-0.5090	-0.0426	0.7853	0.1732	0.6942		
K Nearest Neighbour, K=1	Relative % Error	29%	46%	48%	10%	30%	12%	29%	71%
	Mean Squared Error	0.0806	0.0969	0.1464	0.0216	0.0611	0.0213		
	Correlation Coefficient	0.042	0.0843	0.2329	0.558	0.0262	0.5363		
K Nearest Neighbour, K=1 (Combined StudyNet Clicks)	Relative % Error	27%	54%	49%	11%	43%	19%	34%	66%
	Mean Squared Error	0.0736	0.1101	0.1426	0.0247	0.0838	0.0315		
	Correlation Coefficient	-0.0295	-0.0083	-0.1433	0.4638	0.2651	0.1394		
K Nearest Neighbour, K=2	Relative % Error	26%	51%	34%	14%	26%	11%	27%	73%
	Mean Squared Error	0.0527	0.0982	0.0781	0.0261	0.046	0.0217		
	Correlation Coefficient	-0.1536	-0.34	0.0701	0.3899	0.1541	0.5424		
K Nearest Neighbour, K=2 (Combined StudyNet Clicks)	Relative % Error	26%	42%	37%	11%	31%	19%	28%	72%
	Mean Squared Error	0.0634	0.0755	0.0841	0.0229	0.0586	0.032		
	Correlation Coefficient	-0.0295	0.1093	0.0683	0.5106	-0.0404	0.0455		
K Nearest Neighbour, K=3	Relative % Error	26%	45%	26%	26%	28%	11%	27%	73%
	Mean Squared Error	0.0527	0.0842	0.0591	0.0334	0.0532	0.0181		
	Correlation Coefficient	-0.0876	-0.3019	0.2973	0.146	-0.1536	0.7391		
K Nearest Neighbour, K=3 (Combined StudyNet Clicks)	Relative % Error	24%	40%	32%	12%	27%	18%	26%	75%
	Mean Squared Error	0.0613	0.0669	0.0692	0.0028	0.0526	0.0289		
	Correlation Coefficient	-0.3137	0.0535	0.1928	0.5069	-0.0402	0.2075		
Random Forest	Relative % Error	20%	44%	30%	19%	29%	9%	25%	75%
	Mean Squared Error	0.0341	0.0657	0.0756	0.0359	0.0461	0.0191		
	Correlation Coefficient	0.4165	0.1443	0.0648	0.1352	0.1711	0.5985		
Random Forest (Combined StudyNet Clicks)	Relative % Error	20%	50%	35%	10%	29%	14%	26%	74%
	Mean Squared Error	0.0465	0.0922	0.0726	0.0189	0.0542	0.0196		
	Correlation Coefficient	0.0438	-0.2986	0.1289	0.62	0.0732	0.579		

Appendix H: Intelligent Learning/Training Systems

Intelligent Learning/Training Systems in the Education Sector

No.	System	Home page/Web link
1	**ActiveMath**	http://activemath.com/
2	**ALEKS**	https://www.aleks.com/
3	**Algebra Tutor**	http://act-r.psy.cmu.edu/papers/Lessons_Learned.html
4	**Andes Physics Tutor**	http://www.andestutor.org/
5	**Aplia**	https://www.cengage.com/aplia/
6	**ASPIRE**	http://aspire.cosc.canterbury.ac.nz/
7	**AutoTutor**	http://ace.autotutor.org/IISAutotutor/index.html
8	**Betty's Brain**	https://wp0.vanderbilt.edu/oele/bettys-brain/
9	**Carnegie Learning**	https://www.carnegielearning.com/
10	**CIRCSIM-Tutor**	http://www.cs.iit.edu/~circsim/
11	**COLLECT-UML**	https://link.springer.com/chapter/10.1007/11554028_64
12	**DreamBox**	http://www.dreambox.com/
13	**EER-Tutor**	https://ictg.cosc.canterbury.ac.nz:8005/eer-tutor/login
14	**ESC101-ITS**	https://www.cse.iitk.ac.in/users/karkare/MTP/2014-15/mohit2015parsing.pdf
15	**eSpindle**	https://www.learnthat.org/
16	**eTeacher**	http://www.eteacher-project.eu/
17	**Grockit**	https://www.crunchbase.com/organization/grockit
18	**Knewton**	https://www.knewton.com/
19	**Knowledge Sea II**	http://www.pitt.edu/~taler/KnowledgeSea.html
20	**KnowRe**	https://www.knowre.com/
21	**LearnSmart**	https://www.mheducation.com/prek-12/platforms.html

22	**Mathematics Tutor**	https://link.springer.com/article/10.1007/s40593-014-0023-y
23	**Mathspring**	http://mathspring.org/
24	**Memorangapp**	https://www.memorangapp. com/
25	**MyLab, Mastering**	https://www.pearsonmylabandmastering.com/global/
26	**PlanetSherston**	http://www.learnanywhere.co.uk/planet-sherston/
27	**PrepMe**	http://prepme.com/
28	**PrepU**	https://thepoint.lww.com/template/rendertemplatebyinstanceid/-18
29	**REALP**	https://www.cmu.edu/cmtoday/education_innovation/cognitive-learning-innovative-practice/
30	**Scootpad**	https://www.scootpad.com/
31	**SmartTutor**	http://www.smarttutor.com/
32	**Snapwiz**	http://www.snapwiz.com/
33	**SpellBEE**	https://www.spellbeeinternational.com/
34	**SQL-Tutor**	http://www.cosc.canterbury.ac.nz/tanja.mitrovic/sqltw-its.htm
35	**Why2-Atlas**	https://dl.acm.org/citation.cfm?id=744057
36	**ZOSMAT**	Atatürk University

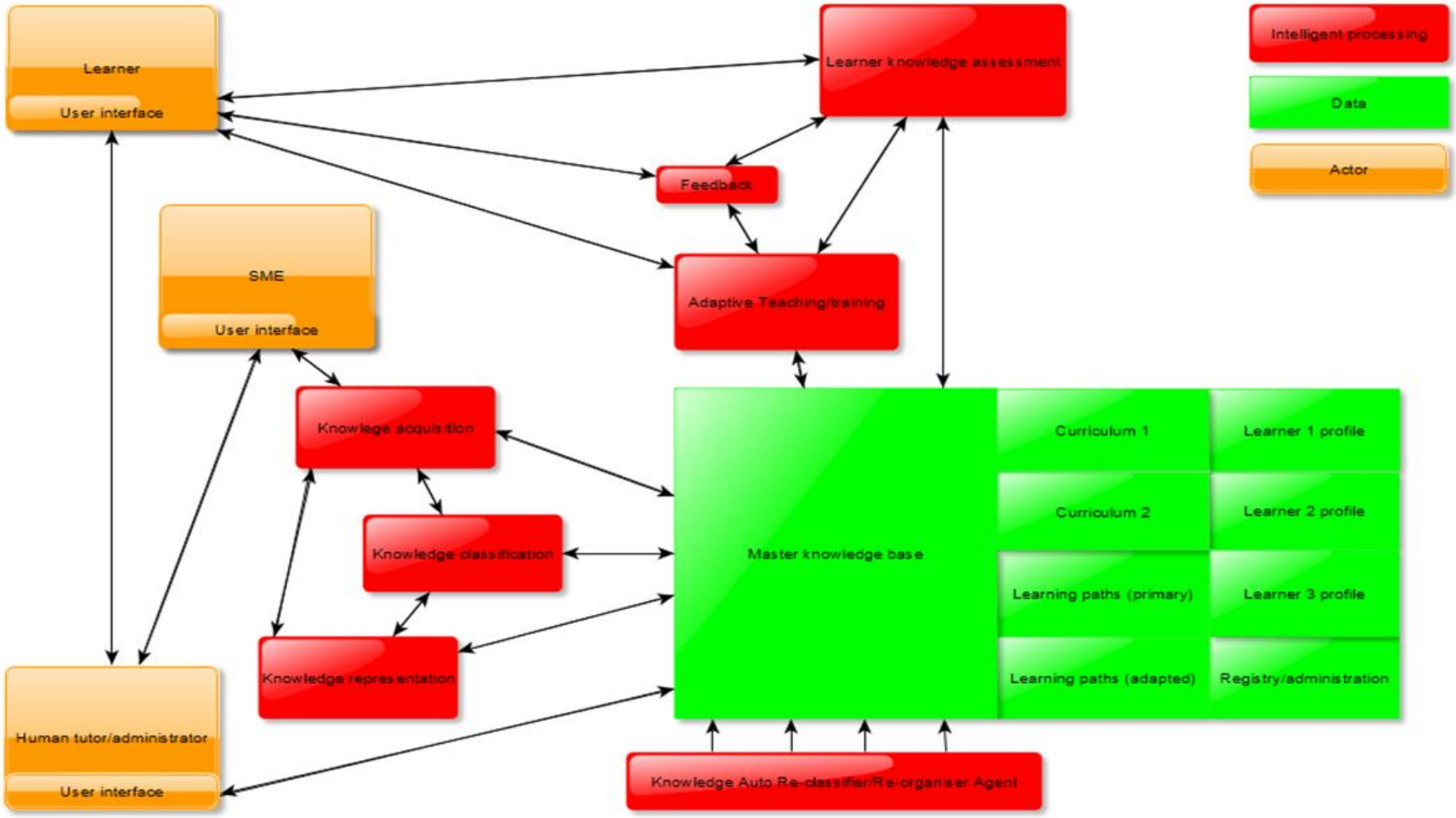
Intelligent Learning/Training Systems in the Commercial Sector

No.	**System**	**Home page/Web link**
1	**3KEYMASTER**	https://www.ws-corp.com/default.asp?PageID=25&PageNavigation=Instructor-Station
2	**Cerego Global**	https://www.cerego.com/
3	**CODES**	https://www.researchgate.net/figure/Music-prototyping-edition-in-CODES_fig1_220999324
4	**CogBooks**	https://www.cogbooks.com/
5	**SHERLOCK**	https://apps.dtic.mil/dtic/tr/fulltext/u2/a201748.pdf

Intelligent Learning/Training Systems in the Education & Commercial Sector

No.	System	Home page/Web link
1	**Adaptive 3.0 Learning Platform**	https://www.fulcrumlabs.ai/adaptive-learning-3-0/
2	**Alelo**	https://www.alelo.com/
3	**aNewSpring**	https://www.anewspring.com/
4	**Cardiac Tutor**	https://artiteacher.wordpress.com/2018/05/16/cardiac-tutor/
5	**Desire2Learn, LeaP**	https://www.d2l.com/en-eu/products/learning-environment/
6	**ELM-ART**	http://www.contrib.andrew.cmu.edu/~plb/ITS96.html
7	**Generalized Intelligent Framework for Tutoring (GIFT)**	https://www.gifttutoring.org/projects/gift/wiki/Overview
8	**Navigate 2**	http://www.jblnavigate.com/2
9	**OER & Competency Learning Platform**	https://www.bnedloudcloud.com/competency-learning-platform/
10	**Oracle Intelligent Tutoring System (OITS)**	https://philpapers.org/rec/ALDDAE-2
11	**Realizeit**	http://realizeitlearning.com/
12	**Smart Sparrow**	https://www.smartsparrow.com/

Appendix I: Adaptive Learning System Conceptual Framework (Wakelam et al., 2015)



Appendix J: The Potential for Using Artificial Intelligence Techniques to Improve e-learning Systems

European Conference on E-learning (ECEL), October 2015

The Potential for Using Artificial Intelligence Techniques to Improve e-learning Systems

Edward Wakelam, Amanda Jefferies, A, Neil Davey, Yi Sun.
University of Hertfordshire, Hatfield, UK
e.wakelam@herts.ac.uk
a.l.jefferies@herts.ac.uk
n.davey@herts.ac.uk
y.2.sun@herts.ac.uk

Abstract: There has been significant progress in the development of techniques to deliver more effective e-learning systems in both education and commerce but our research has identified very few examples of comprehensive learning systems that exploit contemporary artificial intelligence (AI) techniques. We have surveyed existing intelligent learning/training systems and explored the contemporary AI techniques which appear to offer the most promising contributions to e-learning. We have considered the non-technological challenges to be addressed and considered those factors which will allow step change progress. With the convergence of several of the required components for success increasingly in place we believe that the opportunity to make this progress is now much stronger.

We present a description of the fundamental components of an adaptive learning system designed to fulfil the objectives of the teacher and to develop a close relationship with the learner, monitoring and adjusting the teaching based upon a wide variety of analyses of their knowledge and performance. This is an important area for future research with the opportunity to deliver significant value to both education and commerce. The development of improved learning systems in conjunction with trainers, teachers and subject matter experts will provide benefits to educational institutions and help commercial organisations to face critical challenges in the training, development and retention of the key skills required to address new, emerging technologies and business models.

Keywords: Adaptive learning systems, evaluation of intelligent tools, adoption of e-learning by teachers and learners, education and career training, artificial intelligence

1. Introduction

There appears to be considerable potential to make significant steps forward in the application of Artificial Intelligence (AI) to learning systems. A variety of AI techniques (Russell & Norvig 2002) can be applied in real-time to analyse learner behaviour, tailor learning components to learner abilities and knowledge, and to exploit the very large quantities of subject and student data available in both the education and commercial sectors. The development of learning systems in conjunction with trainers, teachers and subject matter experts will provide benefits to institutions across the board, from career/vocational development, re-validation and re-training through to Higher Education and school. This potential has existed for some time, and while research to date has found a variety of work discussing and modelling how individual AI techniques can be applied to different aspects of learning systems and student achievement (for example Gligora Marković, et al., 2014) very few examples of comprehensive learning systems that exploit AI techniques have been identified to date.

Bridging the gap between emerging techniques in AI and Machine Learning (ML) described in section 2 and the essential pedagogy (the theory and practice of education) has proven to be a significant challenge (Jenkins, et al., 2014). However, we believe that the opportunity to make step change progress is now much stronger with the convergence of several of the required components for success increasingly in place. These are:

- The availability of appropriate learning platforms, with almost all learners owning computing devices both inside and outside of the learning setting.

- The increasing quantity and quality of the data (subject and analytics) available to the analytical learning systems using AI.
- The technology (hardware and supporting software) is now powerful enough to handle and exploit the quantity and complexity of data and algorithms necessary for success.
- Institutions are putting more emphasis into this area – exploiting e-learning opportunities and looking for efficiency gains (Johnson 2014).
- Learners are increasingly interested in learning and developing their knowledge on-line at least in parallel with the traditional classroom/campus model.

As a result, the deployment of AI and ML techniques in Technology Enhanced Learning (TEL) has the potential for accelerated growth and adoption. In particular, exploring how AI and ML techniques can be applied to the development of adaptive learning systems, this includes the classification and representation of subject matter knowledge. The latter refers to the organisation of the subject knowledge and the rules and the processes which connect them into a logical structure that:

- Is comprehensive and efficient for the learning system, as well as for the creation, validation and future manipulation by the subject matter expert (SME).
- Is capable of incorporating all the relevant interconnections between the information in a similar way to the way our own brains do.
- Allows the learning system itself to automatically self-organise and search for further connections and rules (Mo et al. 2012).

The aim of this paper is to identify ways in which current research is addressing how contemporary artificial intelligence techniques can be used to improve technology enhanced learning.

2. An Overview of the Literature

In this section the current status and best practice in the four foundational areas of this research: Pedagogy; Technology Enhanced learning; Relevant Artificial Intelligence and Machine Learning Techniques; Survey of Intelligent Learning/Training Systems**;** are discussed:

Pedagogy

Pedagogy continues to be a major area of research with significant on-going work into the field of Technology Enhanced Learning, alongside increased understanding of the behaviours and needs of both learner and tutor (Jenkins, et al., 2014). The latest in the Open University series of Innovating Pedagogy reports (Sharples, et al., 2014) identifies ten innovations that are expected to transform education, from threshold concepts and bricolage to learning to learn and learning design informed by analytics. This body of work, including a very wide variety of field trials and extensive data provides a firm foundation upon which to analyse existing TEL techniques, approaches and learning systems, and to identify the critical factors necessary for the successful definition, design and development of step-forward adaptive learning systems including subject matter knowledge classification. For example, modelling student performance and applying learning analytics is critical to the review of any application of pedagogical concepts (Tempelaar, et al., 2015).

An exploration of the latest pedagogical research confirms the breadth and depth of formal understanding of the art and science of education available to the designers of learning systems, albeit with continuing adjustments being made to educational best practice. It would be impractical to incorporate every component of available research conclusions and recommended approaches and it is therefore important to focus upon those which are fundamental, and wherever possible allow real time decision making based upon incisive learner interaction and individual based learning history and data to determine the system approach.

An aspect of the development of any learning system is an understanding of the variety of individual learning styles (Graf 2007). Graf's paper illustrates the considerable variety of research and opinion on an individual's learning criteria. Basing an approach on all of these would be very challenging, while any non-formal method of deciding which ones to select could

result in a flawed approach. Therefore, in designing an effective adaptive learning system we can choose one of two distinct approaches:

- Incorporating a formal method of automatically detecting the learner's learning style (Feldman, et al., 2014).
- Allowing the system to explore and exploit the actual learning style being displayed by the learner by capturing and analysing all and any parametric data (e.g. even including the colours of the content) available to the system, i.e. collecting as much data as possible to allow the algorithms to decide what's best for the specific learner. This is the approach taken by Realizeit (Realizeit 2015) which has proven successful in their adaptive learning product.

A learner's cognitive style (the way an individual thinks, perceives and remembers information) is another key pedagogical concept where there is some evidence that exploiting an understanding of these concepts has improved student learning achievement (Chipman 2010). This is an area for research and potential exploitation, although it is important to note that there has been conflicting evidence on whether cognitive style makes any difference when designing Adaptive Learning Systems (Mampadi et al. 2011).

Technology Enhanced Learning

The field of TEL has been the subject of much research and practice, in a very wide range of techniques and approaches ranging from classroom management and collaborative learning to MOOCs and gamification (Glover 2013). An analysis of TEL research published between 2009 and 2014 (Schweighofer & Ebner 2015) recorded 4567 papers, dealing with aspects from demographical differences to learner/teacher issues and technical infrastructure. The majority of these papers focus upon Higher Education with only 38 papers addressing business.

However, the commercial world is facing critical challenges in the training, development and retention of key skills, exacerbated by new, emerging technologies and business models, giving organisations business critical dependencies on the relevant subject matter experts (SMEs) and on leadership/talent development (Bhatia & Kaur 2014). These challenges are presenting a major threat in many organisations, limiting business opportunities and weakening their ability to compete (Schuler et al. 2011). Developments in TEL and in particular in the progress of adaptive learning systems already explored in HE (Lilley & Piper, 2009) have the potential to make a dramatic difference in addressing these challenges.

Commercial organisations are increasingly automating their training programmes to allow them to be delivered globally, asynchronously and electronically. These training modules can be stand-alone or part of a classroom based blended learning package and are ideal for situations where a large number of geographically separated learners are targeted. Typically, these modules are delivered as on-line question and answer based dialogues, presenting the learner with explanatory information, occasionally including video material, followed by marked exercises. The learner repeats the course until the pass level is reached and at each subsequent re-take the questions are varied from a set database.

In the UK Higher Education (HE) sector, progress in the numbers of on-line courses available to students has been modest in recent years (see Table 1), giving rise to concerns that the investments in TEL are not addressing pedagogical needs (Jenkins, et al., 2014). As identified by Jenkins "supplementary use of the web to support module delivery remains the most common use of TEL" and as can be seen from the table, fully online modules are a very small proportion.

Table 1: Proportion of all modules or units of study in the TEL environment in use across the UK HE sector (Walker et al., 2014)

Sector mean	2014	2012	2010	2008	2005	2003
Category A – web supplemented	39%	39%	46%	48%	54%	57%
Category Bi – web dependent, content	27%	29%	26%	24%	16%	13%
Category Bii – web dependent, communication	9%	10%	17%	13%	10%	10%
Category Biii – web dependent, content and communication	21%	18%	18%	13%	13%	13%

Category E – fully online	3%	3%	3%	4%	6%	5%

The 2014 summative HE Academy report on flexible technologies (Barnett 2014) observed that the drive towards greater flexibility is now being influenced by a combination of the marketisation of HE, the demands of students as consumers, the potential of new technologies and the apparent potential for making HE available to a wider audience at lower unit costs.

Recent analysis of 4567 TEL publications between 2009 and 2013 (Schweighofer & Ebner 2015) recognises the breadth and depth of on-going research into TEL approaches, summarising key aspects to be taken into account in TEL implementation. These analyses show learners' aspects, including learning behaviour, strategy and style, as well as interaction and participation, as the largest focus of research in the more technologically focused publications.

In the future it is likely that it will be the demands and imperatives of the students and/or the commercial learners that prove to be a major driver in TEL adoption, not only for its educational merit, but in order to enable them to support the stresses of combining work, study and personal life (Jefferies & Hyde 2010, Fabris 2015). Intensified by trends in social media, the integration of on-line, hybrid and collaborative learning alongside the rise of data driven learning and assessment are all strong pressures for increasing the adoption of TEL in HE (Johnson 2014).

Relevant Artificial Intelligence and Machine Learning Techniques

In parallel, there has been considerable progress in the field of Artificial Intelligence (AI) and its related subjects with substantial on-going research in both the academic and commercial worlds. Since early 2014 the level of media interest in the field has noticeably increased with articles in the news such as: 2029, the year when robots will have the power to outsmart their makers (Kurzweil 2014) and Driverless cars trialled on UK roads for first time in four towns and cities (Dearden 2015). This steady increase in public awareness (albeit in more populist topics) will facilitate a more open approach to considering AI as a practical tool in real life activities, and in respect of this research in its application to learning systems in both educational and commercial areas.

Of particular relevance to learning systems are continued developments in Machine Learning (ML), which aims to determine how to perform important tasks by generalizing from examples (Hastie et al, 2005). This includes data mining which is a technique for analysing and extracting data, correlations and patterns from large data sets and turning it into useful information. Other commonly used techniques are:

- Neural networks, which are composed of a large number of highly connected processing nodes working in unison to solve specific problems.
- Support Vector Machine (SVM) which allows us to classify data in a way in which we can then analyse new data points to confidently identify which solution space they fit within.
- Decision trees which allow us to create a tree-like picture of decisions and alternative next steps and to determine a strategy to reach a defined goal.

Other AI techniques to be considered are:

- Knowledge Based Systems (sometimes referred to as Expert Systems), which use a set of rules to solve problems based upon stored expert knowledge (Höver & Steiner 2009).
- Fuzzy logic which allows us to use degrees of truth/accuracy in data analysis rather than the black or white ones and zeroes or yes and no's traditionally used in systems (Benabdellah 2014).
- Roulette wheel algorithms which select the best fitting solutions to problems combined with fuzzy logic have been deployed to maximise learning path choice (Benabdellah 2014) and to predict student motivation (Sivakumar & Praveena 2015).
- Ant Colony optimisation is an algorithm for establishing the optimal paths in data and processes in a similar way to how ants behave (Sivakumar & Praveena 2015).

These techniques are critical for exploiting the very large subject matter and student/learner data sets now available in order to develop powerful new learning systems. These data sets are no longer capable of real-time analysis by using manual or orthodox IT techniques due to:

- The very large quantity of data that is available to be captured and exploited.
- The level of complexity of the interdependencies of large numbers of data classes/attributes, requiring multi-dimensional analysis (Tempelaar, et al., 2015).

Suitable techniques for continued research and development are grouped under Adaptive Learning Systems (ALS), Intelligent Tutor Systems (ITS), Cognitive Systems and Predictor/Recommender Systems. The line between Intelligent Tutoring Systems and Adaptive Learning Systems has become increasingly blurred. In the past ITSs tended to be subject matter specific, developing from what can be described as "flowcharted learning" into increasingly sophisticated systems deploying AI techniques. The field of adaptive learning has allowed these systems to develop a close relationship with the learner, monitoring and adjusting the teaching and creating idealised learning paths based upon a wide variety of analyses of their knowledge and performance (Marengo, et al., 2015). This level of automated judgement is made by understanding the learner profile, their learning style and their base knowledge of the subject area (Marengo, et al., 2015).

In designing adaptive learning systems there are a significant number of potential techniques and models which can be deployed. Recent research into the prevalence of these show learner and domain knowledge modelling, adaptivity and content presentation as the most prevalent in learning systems, with cognitive style almost the least characterised (Gligora Marković, et al., 2014). In the US there is positive evidence of the increasing adoption of such systems. As discussed in section 3 below, the challenges are mainly organisational and not technological (Oxman & Wong 2014). The first commercial successes in learning systems in the US came from cognitive tutoring systems which delivered high school mathematics to over 475,000 students in 2007 (Raley 2012), showing that students performed 15-25% and 50-100% respectively better than the control group on skill knowledge and problem solving

Additionally, some progress has been made in the area of adaptive learning systems in the commercial area, with research into the benefits and risk areas from the learner's point of view. The results indicated a positive response to the alignment of adaptive learning to job roles and career paths, while removing the time wasted on non-relevant learning material. The research also reinforced the criticality of the input and capture of the expert knowledge (Höver & Steiner 2009).

Survey of Intelligent Learning/Training Systems

A number of successful, although mostly niche, systems have been developed and are in place in the field, alongside a variety of prototypes. As can be seen in Table 2, systems in the education sector dominate.

Table 2: Survey of "Intelligent" Learning/Training Systems Identified

Sector	Quantity	Percentage
Education sector	32	78%
Commercial/Public sector	3	7%
Both	6	15%
Total	41	100%

Of those surveyed, 17 (41%) have been developed by universities or as collaborative projects between university and industry. We estimate that approximately half (46%) are adaptive learning systems the details of which are shown in Tables 3, 4 and 5.

Adaptive learning systems adjust the learning experience based upon the student's progress, increasing the level of difficulty when they're progressing well, and slowing down if they need further instruction. The greatest progress appears to be where the knowledge base being addressed is embodied in comprehensively curated areas of knowledge, for example, STEMM subjects including mathematics and physics, and English education.

Table 3: Intelligent Learning/Training Systems in the Education sector

System	Developed by	Type	Key words
ActiveMath [P, J, S]	DFKI & Saarland University	Adaptive learning	Educational data mining. Natural Language Processing. Collaborative. STEMM.
ALEKS [P, J, S, U]	New York University and the University of California, Irvine	Adaptive learning	Web based. Knowledge space theory. STEMM, Accounting.
Algebra Tutor [S]	Carnegie Mellon	Intelligent tutoring	Artificial intelligence, cognitive, human computer interaction. Computer programming, STEMM.
Andes Physics Tutor [S, U]	Arizona State University	Intelligent tutoring	Highly interactive. STEMM.
Aplia [U, Po]	Stanford university	Adaptive learning	On-line homework system. Multiple subjects - STEMM, accounting, English, history, finance.
ASPIRE [J, U]	University of Canterbury (New Zealand)	Intelligent tutoring	Authoring. Develops web tutoring systems.
AutoTutor [U]	University of Memphis	Intelligent tutoring	Natural language. Speech engine. Newtonian physics, Introductory computer literacy.
Betty's Brain [P, S]	Vanderbilt & Stanford Universities	Cognitive	Metacognitive skills. STEMM.
Carnegie Learning [S]	Carnegie Mellon University	Adaptive learning Cognitive	Pedagogy. Cognitive science. Research led. STEMM.
CIRCSIM-Tutor [U]	Sponsored by US Naval Research Office	Intelligent tutoring	Dialogue based, natural language. Medicine.
DreamBox [P, J]	DreamBox	Adaptive learning	Game-like environment based. STEMM.
ESC101-ITS [U]	The Indian Institute of Technology, Kanpur, India	Intelligent tutoring	Programming.
eSpindle [P, J, S]	LearnThat	Personalised learning	US Spelling Bee system. Spelling.
eTeacher [S, U]	eTeacher	Adaptive learning	Intelligent agent. On-line assisted learning. System engineering course.
Grockit [S]	Kaplan	Adaptive learning	Collaborative. Game-like environment. STEMM.
Knewton [S, U]	Knewton	Adaptive learning	Content agnostic. Psychometrics and cognitive learning theory, Inference engine.
System	**Developed by**	**Type**	**Key words**
Knowledge Sea II [U, Po]	University of Pittsburgh	Adaptive learning	Computer programming.
KnowRe [J, S]	KnowRe	Adaptive learning	Game-like environment based. STEMM.
Mathematics Tutor [J, S]	University of Massachusetts	Adaptive learning	STEMM.
Mathspring [P, J, S]	Univ of Massachusetts	Adaptive learning	Intelligent tutoring. Math.
Memorangapp. [U, Po]	MIT	Memory reinforcement.	Spaced repetition. Medicine.
MyLab, Mastering [U, Po]	Pearson	Adaptive learning	On-line learning. Multiple subjects.
PlanetSherston [P]	Sherston	Personalised learning	Game play learning.
PrepMe [S]	Stanford, University of Chicago, CalTech	Adaptive learning	Virtual classroom. STEMM.
PrepU [U, Po]	PrepU, collaboration with UCLA	Adaptive learning	Quiz engine. STEMM.
REALP [J, S]	Worcester Polytechnic Institute, Carnegie Mellon	Personalised learning	Based upon a tool designed to investigate the development time for tutoring systems. Reading comprehension.
Scootpad [P, J, S]	Scootpad	Adaptive learning	Behaviour tracking. Prediction. STEMM.
SmartTutor [A]	University of Hong Kong	Adaptive learning	Personalised on-line distance

			learning. Generic.
Snapwiz [U, Po]	Wiley	Adaptive learning	Collaborative. STEMM, Languages, Business, Social Science.
SpellBEE [P, J, S]	Brandeis University	Artificial Intelligence Machine learning	Education research tool.
Why2-Atlas [U]	UCLA	Natural language	Textual analysis system. STEMM.
ZOSMAT [J,S]	Atatürk University	Intelligent tutoring	Classroom based. STEMM.

[Key: Primary, Junior, Secondary, University, Postgraduate, Adult]

Table 4: Intelligent Learning/Training Systems in the Commercial/Public sector

System	Developed by	Type	Key words
aNewSpring	aNewSpring	Adaptive learning	Corporate Learning Management System. Blended and hybrid learning
CODES	Universidade Federal do Rio Grande do Sul	Learning system	Web-based. Musical prototyping specific for non-musicians.
SHERLOCK	University of Pittsburgh	Intelligent Tutoring System	Decision trees. Student competence and performance model. USAF technician specific.

Table 5: Intelligent Learning/Training Systems in the Education & Commercial sector

System	Developed by	Type	Key words
Alelo	University of Southern California	Virtual Role-Play simulations	Pedagogical agents as social actors. Multimedia. Cyberlearning.
Cardiac Tutor	University of Massachusetts Medical School	Adaptive learning/Intelligent tutoring	Real time simulation. Knowledge based. Medicine, cardiology specific.
Desire2Learn, LeaP	Brightspace	Adaptive learning	Predictive analytics.
ELM-ART	Freiburg University of Education	Adaptive learning	Web-based. LISP programming specific
Realizeit	CCKF/Realizeit	Adaptive learning	Content agnostic. Supervised & Unsupervised learning. Classification trees. Fuzzy Logic.
Smart Sparrow	University of New South Wales in Sydney	Adaptive learning Intelligent tutoring Data mining	Educational data mining. Content agnostic.

These systems are dominated by those focussed upon the education sector, but we should expect increasing interest from the commercial world, since individuals will be faced with a number of different careers during their working life as industries are created, evolve and disappear. The development of new and more intelligent methods of supporting these aspirations will become very important to both individuals and organisations, presenting the opportunity to deliver significant value, in terms of reducing training and re-validation costs, in accelerating training delivery and in considerable enhancement of people's personal experience in learning.

In terms of organizational & geographical traction, analysis of existing systems can be summarised as follows:

- The field of education is leading the way in both research and in the development of learning/training systems:
  - Primary, secondary, university education, with STEMM the most popular subject areas. (Table 3).
  - MOOCS have made rapid progress, however the completion rates are less than 7% (Jordan 2014).
- Business/vocational research and learning/training systems are currently running a poor second (Tables 4 and 5) with Medicine appearing more often than others in the area of applying intelligent techniques to areas including diagnosis and training.

- The requirement for distance learning appears to be an early TEL driver.
- Geographically, traction is the highest in the US, followed by the UK, followed by Europe, with Australia showing up intermittently in searches.

3. Challenges to be addressed and related discussion

While the adoption of TEL continues to gain traction, there are a number of organisational/non-technological challenges that must steadily be addressed and in particular kept in mind in the design, development and deployment of these systems:

Organisational
- Systems can be expensive both to develop and to implement.
- Organisational conservatism – the prevailing attitude of "what we have works fine..", and the need to evidence benefits.
- Requires the cooperation and support of individuals across both organisations and organisational levels (Barnett 2014).

Administrative/political:
- Integration of TEL into the existing curriculum (Oxman & Wong 2014).
- Overcoming resistance from competing methods and their champions.

The needs and concerns of the teacher/trainer:
- Teacher/trainer resistance – the need for persistence while under significant pressure to deliver improved student grade performance dealing with high workloads (Wang & Hannafin 2005).
- Requires the cooperation and input of domain subject matter experts.

The needs and concerns of the student/learner:
- Ensuring student/learner motivation and early identification of disenchantment (Oxman & Wong 2014).
- Continuous feedback to ensure the maintenance of a continuously accurate student model (progress measurement, learning rates, proven alternative learning paths).

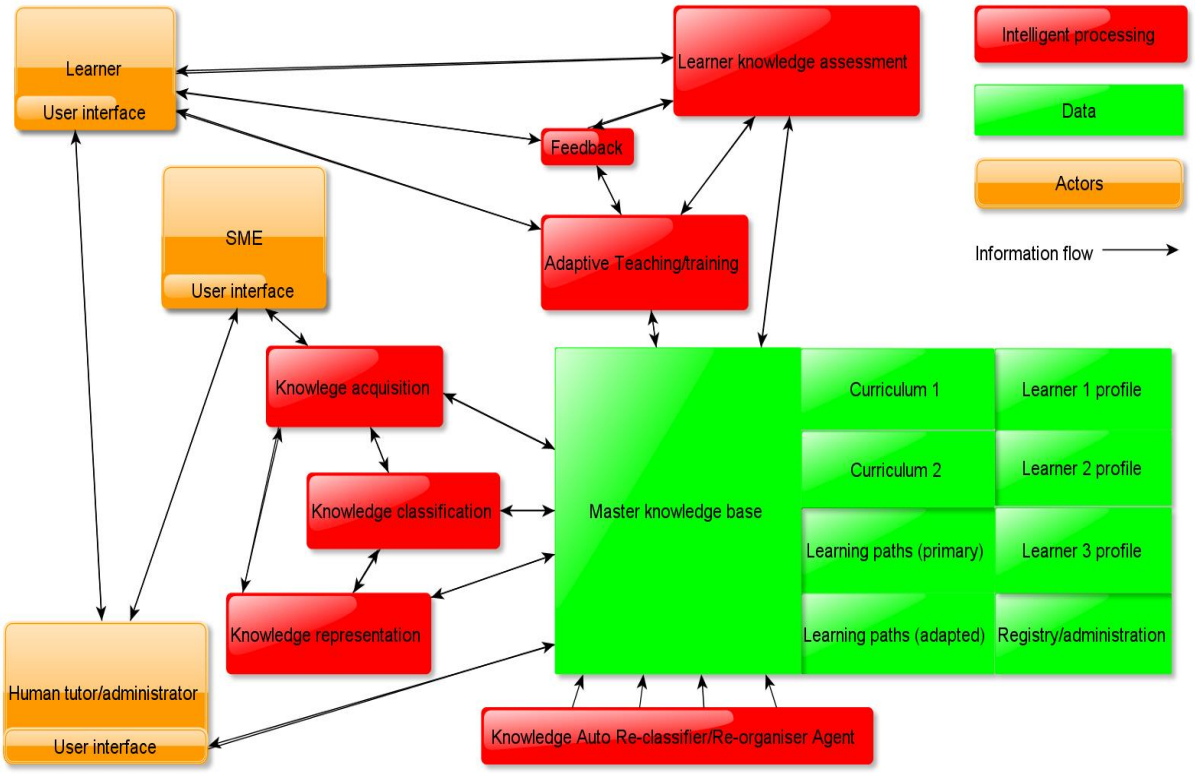
Technical
- The modelling of such a complex cognitive task.
- Incorporating the essential pedagogy. For example, effective feedback to the learner and very careful use of hints to ensure that deep learning is developed.
- Integration with all user platforms - mobile, fixed, on-line/off-line, social.
- Ability to exploit rapidly developing technologies/platforms.
- Necessity of systematic and regular update of domain subject matter.

4. Conclusion

We have identified the scope for contemporary AI techniques to be used in the development of adaptive learning systems and have undertaken a thorough review of existing intelligent learning/training systems in both education and commercial sectors. While some progress has been made there is scope for further work.

Accordingly, we have put together a conceptual framework for an Adaptive Learning System, including all major components as shown in Figure 1.

Figure 1: Adaptive Learning System Conceptual Framework showing human intervention (actors), intelligent processing, data structures and information flows



Future work comprises the establishment of the important features that determine the success of learning systems from the pedagogical perspective based upon research and recent practice. Initial work will be to pilot an analysis of student performance using existing data which we will then use to develop an adaptive learning system. We shall then refine the conceptual framework in line with the latest and emerging pedagogical and AI/ML research and design, implement, test and evaluate an adaptive learning system using contemporary AI techniques.

5. References

Barnett, R. (2014). Conditions of Flexibility Securing a more responsive Higher Education system. *HEA Report.*

Benabdellah, N.C., (2014). Ant Colony Algorithm and new Pheromone to Adapt Units Sequence Learner's Profiles. *International Journal of Computer Science and Applications,*, 12.

Bhatia, A., Kaur, L., (2014). Global Training & Development trends & Practices: An Overview. *International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-8)*, 3(8).

Chipman, S.E.F., (2010). Applications in Education and Training: A Force Behind the Development of Cognitive Science. *Topics in Cognitive Science*, 2, pp. 386–397.

Dearden, L., (2015). Driverless cars trialled on UK roads for first time in four towns and cities. The Independent.

Fabris, C., (2015). Social Networking and Social Support: Does It Play a Role in College Social Integration? The Chronicle of Higher Education.

Feldman, J., Monteserin, A., & Amandi, A. (2014). Automatic detection of learning styles: state of the art. *Artificial Intelligence Review*, 1-30.

Gligora Marković, M., Alen J., Bozidar, K., (2014). A Prevalence Trend of Characteristics of Intelligent and Adaptive Hypermedia E-learning Systems. *WSEAS Transactions on Advances in Engineering Education* 11, 80-101.

Glover, I., 2013. Play as you learn : gamification as a technique for motivating learners Play As You Learn : Gamification as a Technique for Motivating Learners. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telemcommunications 2013*. pp. 1998–2008.

Graf, S. (2007). Adaptivity in learning management systems focussing on learning styles (Doctoral dissertation, Vienna University of Technology).

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*. doi:10.1007/BF02985802

Höver, K.M. & Steiner, C.M., (2009). Adaptive Learning Environments: A Requirements Analysis in Business Settings. *International Journal of Advanced Corporate Learning*, 2(3), pp. 27–33.

Jefferies, A. & Hyde, R., (2010). Building the future students' blended learning experiences from current research findings. *Electronic Journal of e-learning*, 8, pp. 133 – 140.

Jenkins, M., Walker, R., Voce, J., (2014). Achieving flexibility? The rhetoric and reality of the role of learning technologies in UK Higher Education.

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2014). *NMC horizon report: 2014 K* (pp. 1-52).

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distance Learning*.

Kurzweil, R., (2014). 2029: the year when robots will have the power to outsmart their makers. The Guardian.

LACE 2019, Learning Analytics Community Exchange Home Page , viewed 28 Augst 2019, http://www.laceproject.eu

Lilley, M., & Pyper, A. (2009). The application of the flexilevel approach for the assessment of computer science undergraduates. In Human-Computer Interaction. Interacting in Various Application Domains (pp. 140-148). Springer Berlin Heidelberg.

Mampadi, F. et al., (2011). Design of adaptive hypermedia learning systems: A cognitive style approach. *Computers & Education*, 56, pp. 1003–1011.

Marengo, A., Pagano, A., Monopoli,G (2015) Adaptive System Prototype: Automated and Customised Learning Experience, *INTED2015 Proceedings*, pp. 4536-4544.

Mo, S., & Zeng, J. (2012). Particle swarm optimisation based on self-organising topology driven by fitness with different links removing strategies. *International Journal of Innovative Computing and Applications*, *4*(2), 119-132.

Oxman, S. & Wong, W., (2014). White Paper: Adaptive Learning Systems. , (February).

Raley, N. (2012). Intelligent Tutoring Systems: A Literature Synthesis.

Realizeit (2015). Realizeit Adaptive Learning Systems. http://realizeitlearning.com/

Russell, S., & Norvig, P., (2002). Artificial Intelligence: A Modern Approach. Prentice Hall Series in Artificial Intelligence. Prentice Hall. Second edition.

Schweighofer P, Ebner M. (2015). Aspects to Be Considered when Implementing Technology-Enhanced Learning Approaches: A Literature Review. *Future Internet*. 2015; 7(1):26-49.

Sharples, M., Adams, A., Ferguson, R., Gaved, M., McAndrew, P., Rienties, B., Weller,M., & Whitelock, D. (2014). Innovating Pedagogy 2014: Open University Innovation Report 3.

Schuler, R.S., Jackson, S.E. & Tarique, I., (2011). Global talent management and global talent challenges: Strategic opportunities for IHRM. *Journal of World Business*, 46(4), pp. 506–516.

Sivakumar, N., Praveena, R. (2015). Dertermining Optimized Learning Path for an E-learning system using Ant Colony Optimization Algorithm, Milton Keynes: The Open University. *International Journal of Computer Science & Engineering Technology, Vol. 6 No. 02 Feb 2015.*

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2014). In search for the most informative data for feedback generation: Learning Analytics in a data-rich context. *Computers in Human Behavior*.

Walker, R., Voce, J., Nicholls, J, Swift, E., Ahmed, J., Horrigan, S., & Vincent, P. (2014). *2014 Survey of Technology Enhanced Learning for Higher Education in the UK*. Universities and Colleges Information Systems Association, Oxford, UK.

Wang, F. & Hannafin, M.J., (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), pp. 5–23.

Appendix K: The Mining and Analysis of Data with Mixed Attribute Types

Ed Wakelam, Neil Davey, Yi Sun, Amanda Jefferies, Parimala Alva, Alex Hocking
School of Computer Science
University of Hertfordshire
Hatfield, UK
e-mail: {e.wakelam, n.davey, y.2.sun, a.l.jefferies, p.alva, a.hocking3}@herts.ac.uk

***Abstract*— Mining and analysis of large datasets has become a major contributor to the exploitation of Artificial Intelligence in a wide range of real life challenges, including education, business intelligence and research. In the field of education, the mining, extraction and exploitation of useful information and patterns from student data provides lecturers, trainers and organisations with the potential to tailor learning paths and materials to maximize teaching efficiency and to predict and influence student success rates. Progress in this important area of student data analytics can provide useful techniques for exploitation in the development of adaptive learning systems. Student data often includes a combination of nominal and numeric data. A large variety of techniques are available to analyse numeric data, however there are fewer techniques applicable to nominal data. In this paper, we summarise our progress in applying a combination of what we believe to be a novel technique to analyse nominal data by making a systematic comparison of data pairs, followed by numeric data analysis, providing the opportunity to focus on promising correlations for deeper analysis.**

Keywords-Data Mining; Educational Data Mining; Data Analytics; Numeric, Nominal Data Analysis; Dimensionality reduction; Knowledge Extraction.

INTRODUCTION

We are initially investigating the potential to apply Artificial Intelligence (AI) techniques to improve E-learning systems in both educational and business settings [1], In particular, we are focussing upon how learning systems can be designed to adapt to individual students during the learning activity. This adaptability would enable the E-learning system to monitor and adjust the teaching based upon a wide variety of analyses of the knowledge and performance of the student. In order to achieve this, we are investigating how student attributes may be analysed and deployed.

Our first steps have been to perform a variety of analyses on open source published student data [2] in order to identify factors which correlate with student performance [3]. Significant advances in the field of data mining [4] are providing opportunities for tools to be deployed in analysing education data [5]. There have also been continued developments in Machine Learning (ML), which aims to determine how to perform important tasks by generalizing from examples [6].

These results may then be used to improve the design of adaptive learning systems [7] using contemporary AI techniques.

In section II, we discuss each of the types of student features relevant to our research: Categorical, comprising Nominal and Ordinal, and Measurement (Quantitative). Section III introduces the open source student dataset which we have used to explore applicable analysis techniques. In section IV, we describe our experimental analysis of this data, summarising our results in section V. Finally, we discuss our conclusions in section VI including further work already underway and recommendations for future work.

EXISTING DATA ANALYSIS TECHNIQUES

Categorical Data

- *Nominal Features*

Nominal data is data where the feature values are labels such as male/female or yes/no. There are a number of statistical techniques available to analyse nominal datasets, notably Chi-square and Cramer's V [8]. Each has its own limitations, for example, sensitivity to sample size and a stronger than justified evidence of correlations [9].

In the case of nominal data, it is not possible to compare attributes directly in order to search for correlations. However, we can compare the correspondence between groupings of attributes and we have explored the use of what we believe to be a novel technique to do so. In this case, we have chosen to compare correlations between pairs of attributes [10]. Future work is underway to apply alternative nominal data analysis techniques to our data in order to compare our results and to identify the strengths and weaknesses of our technique.

- *Ordinal Features*

Ordinal data is a type of categorical data in which order is important. The originators of our dataset do not categorise any of the student data captured in their study as ordinal.

Measurement (Quantitative) Data

There are a variety of statistical techniques available to analyse quantitative (numeric) datasets. In this case we have selected to use Principal Components Analysis (PCA) to reduce the dimensionality of our data and Growing Neural Gas (GNG) to identify potentially interesting clusters of data. GNG [11] has been successfully used to identify clusters in data for many applications such as the analysis of Hubble Space Telescope images [12] and automatic landmark extraction in images [13]. PCA and GNG have also been successfully combined for intrusion detection [14].

PORTUGUESE STUDENT DATASET

In order to investigate the predictive accuracy of student achievement data was taken from a set of students from a Portuguese study [15]. This data consists of information taken from two Portuguese secondary schools and each student has 33 attributes. The data includes three labels: first period grade,

second period grade and final grade. The subjects are Mathematics (395 students) and Portuguese Language (649 students) and the data was collected during the 2005-2006 academic year. The attributes comprise 16 numeric (including the labels: first period, second period and final performance grades) and 17 nominal (Tables I and II).

EXAMPLES OF THE NUMERIC ATTRIBUTES

Identifier	Description
Age	Student's age (numeric: from 15 to 22)
Absences	Number of school absences (numeric: from 0 to 93)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

EXAMPLES OF THE NOMINAL ATTRIBUTES

Identifier	Description
Gender	Student's gender (binary: "F" - female or "M" - male)
Mjob	Mother's job (nominal: "teacher", "health" care related, civil "services" (e.g., admin or police), "at_home" or "other")
Romantic	With a romantic relationship (binary: yes or no)

For consistency we have adopted the original attribute types as used in the Portuguese study, although there are a small number of the attributes defined as numeric which could be considered as ordinal.

EXPERIMENTAL ANALYSIS

Analysis of Nominal Data

Our method is to compare the correspondence between pairs of our nominal data attributes. To illustrate, the technique, here is a worked example of a dataset of 4 students, each with 2 nominal attributes (Table III).

EXAMPLE DATASET

Student	Attribute 1 (a1)	Attribute 2 (a2)
s1	p	x
s2	p	y
s3	q	z
s4	p	y

After setting a counter to zero we compare every possible pairing of student attribute values in the attribute 1 column of Table III with the corresponding pair in the attribute 2 column. If the selected pair from attribute 1 have the same value and the corresponding pair from attribute 2 also have the same value then we increment the counter by 1. Similarly if they both have different values then we increment the counter by 1. Otherwise, we decrement the counter by 1 (see Table IV).

So, for example, looking at step 1 below, the values of attribute 1 are both "p" (i.e., the same), whereas the values of attribute 2 are "x" and "y" (i.e., different), so we decrement the counter by 1. However, looking at step 2, the values of attribute 1 are "p" and "q" (different), and the values of attribute 2 are "x" and "z" (different), so we increment the counter by 1.

STEP BY STEP PROCESS

Step	Student pairing	a1	a2	Score	Cumulative counter
1	(s1 s2)	(p p)	(x y)	-1	-1
2	(s1 s3)	(p q)	(x z)	+1	0
3	(s1 s4)	(p p)	(x y)	-1	-1
4	(s2 s3)	(p q)	(y z)	+1	0
5	(s2 s4)	(p p)	(y y)	+1	1
6	(s3 s4)	(q p)	(z y)	+1	2

We repeat this process for all combinations of attribute values and the resultant counter totals are used to populate a correlation matrix. This is done by inserting the counter total into the correlation matrix cell which corresponds to the respective attribute. Obviously, each attribute fully correlates with itself resulting in identical values across the matrix diagonal. We normalise our resulting matrix by dividing all entries by this value to keep all correlation matrix values between -1 and +1 (see Table V).

NORMALISED CORRELATION MATRIX FOR ILLUSTRATIVE EXAMPLE 1

	a1	a2
a1	1	$^1/_3$

	a1	a2
a2	$^1/_3$	1

Positive values represent positive correlations between the respective attributes, negative values represent negative correlations and the magnitude of the value represents the strength of the correlation.

For example, where there are a high proportion of student pairs where the corresponding attributes, such as Mother's job and gender are correspondingly the same or different this will result in a relatively higher correlation value (for example, 1/3 in Table V) between the two attributes.

For each attribute, we evaluate its correlation with all other attributes and find the mean value over all these correlations. As a first indicator of interesting attributes, particular attention was paid to those correlations where the magnitude of the mean value was high in comparison to the mean values of other attributes. Those correlations where the magnitude was above the mean for that attribute then provided additional correlations for consideration.

We applied the technique to each of the Mathematics and Portuguese Language datasets in turn. For each dataset, we were then able to identify those pairs of attributes that were most strongly correlated – whether positively or negatively. This enabled us to consider the potential influences on student behaviours.

We were also able to compare the correlations in the Mathematics dataset with those in the Portuguese Language dataset.

Using the correlation matrix generated by this technique we then produced corresponding PC1 v PC2 scatter plots for each of our Mathematics and Portuguese Language student datasets in order to visualize potential clusters for future analysis and comparison with any clusters identified in our numeric data. In order to visualize and more easily identify potential clusters we produced a PCA scatter plot for each of the four final grade intervals (using final grades 0-5, 6-10, 11-15, 16-20 as our labels) for each student dataset.

Analysis of Measurement Data

After normalisation of the Mathematics and Portuguese Language student numeric datasets, respectively (by subtracting the mean and dividing by the standard deviation) we performed a linear Principal Component Analysis (PCA), plotting each of the leading three principle components, PC1 v PC2, PC2 v PC3, PC1 v PC3. In each Figure, the amount of variance accounted for by the respective principal components is reported. For example, in Figure 1 PC1 and PC2 account for 26% of the total information in the data.

In each case a visual inspection suggested possible clusters. In order to try and identify these clusters we applied GNG, with key parameters set to 50 training runs and a maximum of 200 nodes. This technique [16] identified a small number of clusters and their respective centroids as well as allowing us to identify the actual students in each cluster.

RESULTS

We are looking to identify interesting correlations in our student data attributes, providing the opportunity to focus on promising correlations for deeper analysis.

Nominal data

- *Mathematics students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VI and VII respectively.

HIGHEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Higher Education wish	0.23
School	0.19
Parent cohabitation	0.18

LOWEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Paid tutor	0.008
Gender	0.006
Extra-curricular activity	0.003

Our results show potential correlations may exist between the student's wish to take Higher Education and other nominal attributes - the school attended and parent cohabitation status, followed by receipt of extra educational support, Mother's job, access to the internet, the reason for choice of school and nursery school attendance.

Mother's job also shows potential correlations with other factors, including the wish for Higher Education, parent cohabitation, school attended, educational support and choice of school.

Paid extra tuition does not correlate strongly with other factors, even parent's jobs, which we might have expected. This is also true for students receiving

educational support from within the family. However, future analyses may show that such extra tuition correlates with student performance measured by their grades.

Internet access also shows potential correlations with a number of factors, including the wish for Higher Education, school attended, parent cohabitation, address, the level of educational support by the school and Mother's job.

Factors which show very low correlations with others are the level of extra-curricular activities, whether the student was male or female and paid tutoring, followed by romantic relationships, Father's job, and family size.

- *Portuguese Language students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VIII and IX respectively.

HIGHEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	**Mean value**
Paid tutor	0.20
Higher Education wish	0.18
Parent cohabitation	0.16

LOWEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	**Mean value**
Family education support	0.02
Gender	0.01
Extra-curricular activity	0.003

Our results show potential correlations may exist between paid tutoring, the student's wish to take Higher Education and parent cohabitation followed by educational support and Mother's job.

Paid extra tuition shows potential correlations with a number of other factors including the level of educational support, the wish for Higher Education, parent cohabitation, and Mother's job. This is also true for extra educational support provided by the school, correlating with the use of paid tutors, parent cohabitation, and Mother's job.

Mother's job shows potential correlation with the use of paid tutoring, educational support, parent cohabitation and attendance at a nursery school.

Internet access only correlated modestly with other factors for Portuguese Language students.

Factors which show very low correlations with others are the level of extra-curricular activities, student gender and family educational support, followed by romantic interest, guardian, Father's job and school attended.

- *Comparisons between Mathematics and Portuguese Language analysis results*

The wish to take Higher Education shows potential correlation with Mother's job, cohabitation status and receipt of extra educational support for both sets of students.

In both cases Mother's job correlates with other factors. In contrast, Father's job, along with romantic relationships and extra-curricular activities shows very low correlations with other factors in both sets.

Additional educational support provided by the school also shows potential correlation with a number of other factors in both sets.

In comparison with Portuguese Language students, paid extra tuition in the case of Mathematics students does not correlate strongly with other factors.

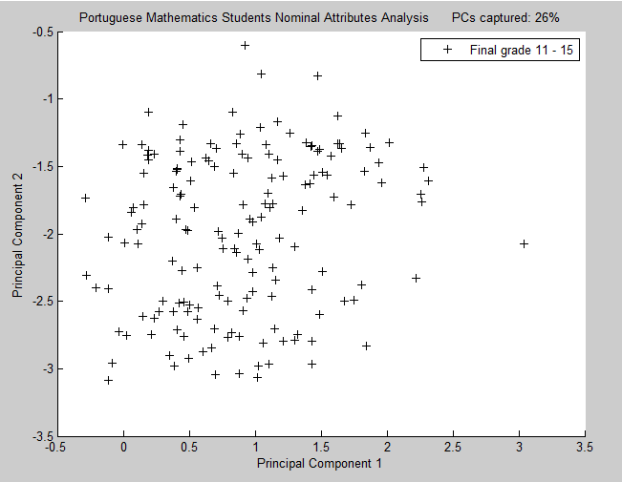
Interestingly, gender, considered to be an influential factor, does not correlate well with other attributes in either set.

In the case of Mathematics students, internet access shows potential correlations with a number of factors, such as the wish to take further education, school attended, and parent cohabitation. However, in the case of Portuguese Language students, internet access shows only modest correlations.

- *Principal Component Analysis*

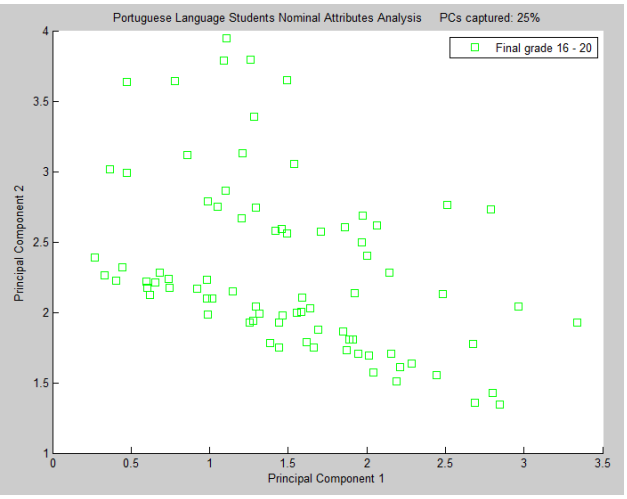
As described in section 1, above, a PCA projection will allow visualization of multi-dimensional data in a two dimensional representation. For each dataset the initial PCA plot including all final grades proved too challenging to visualize and so we produced four plots, one for each of the four final grade intervals. We have included one example from each dataset. Principle component analysis of our Mathematics and Portuguese Language student data shows no evidence of potential clustering.

For example, a PC1 v PC2 nominal data plot of Mathematics students' achieving final grades of between 11 and 15 (Figure 1).



Mathematics nominal data PC1 v PC2 Final Grades 11-15

A further example shows a PC1 v PC2 nominal data plot of Portuguese Language students' achieving grades of between 11 and 15 (Figure 2). This data plot appears to exhibit a lower boundary delineation which we believe to be a result of a predominance of very narrow variances in the attribute values in this particular dataset.

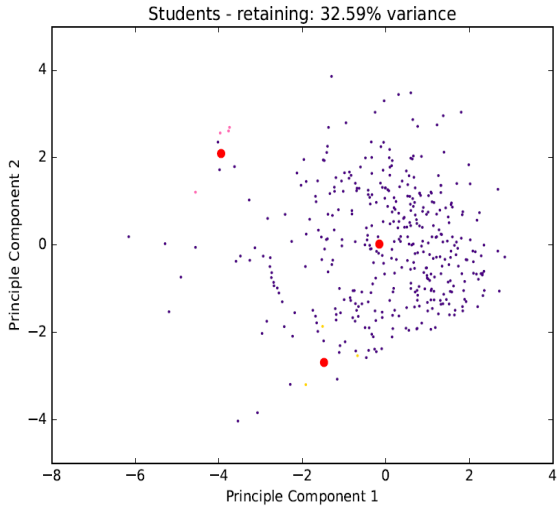


Portuguese Lang nominal data PC1 v PC2 Final Grades 16-20

Measurement data

- Mathematics students

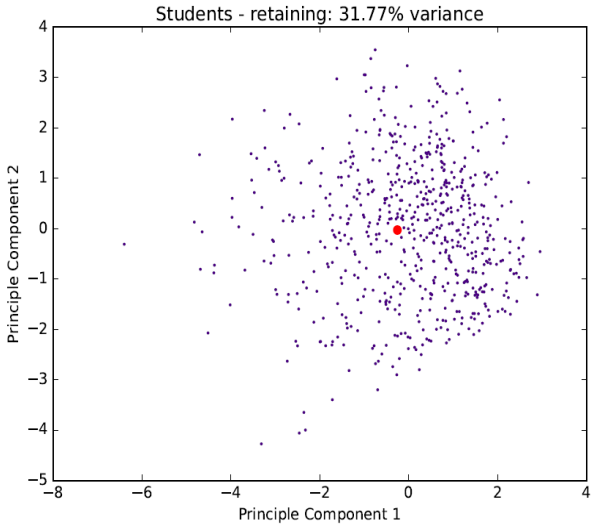
GNG identified modest clustering in each of the PC1, PC2, PC3 comparisons For example, in Figure 3 we can see that three clusters have been identified. The centroids are shown in red and in each case the students in each cluster are identified in order to for look for potential correlations with the results of our nominal data analysis.



Mathematics students numeric data PC1 v PC2 scatter plot

- Portuguese Language students

GNG did not identify useful clustering in either of the PC1, PC2, PC3 comparisons. In all cases only one cluster was identified, for example, in Figure 4. As above, the centroids are shown in red.



Portuguese Lang. students numeric data PC1 v PC2 scatter plot

We repeated the GNG analysis, adjusting the key parameters, increasing the number of training runs from 50 to 100 and maximum nodes from 200 to 600. However, this did not result in improvement. Further work is underway to identify alternative techniques to identify potential clustering in the Portuguese Language student numeric data, such as Curvilinear Component Analysis (CCA).

CONCLUSION AND NEXT STEPS

In this paper, we have taken the first steps in exploring a mixed attribute type (numeric and

nominal) dataset provided by real student data with the objective of identifying useful potential correlations between attributes.

We have applied a novel approach to the analysis of the nominal data, comparing the correspondence between pairs of nominal attributes.

We then investigated if the analysis would identify interesting information in the dataset, which to some extent it did. Our PCA plot of the Mathematics nominal data showed no evidence of clustering. Further work is underway to apply a non-linear visualization method in order to investigate potential clustering.

We then applied numeric data analysis techniques to identify clustering and potential correlations in our numeric attributes identifying some potentially interesting patterns.

In the case of our Mathematics student data using Principle Component Analysis followed by the GNG technique we were able to identify some clustering of the data, however the corresponding analysis of our Portuguese Language student data did not identify useful clusters.

Further work is underway to analyse and make comparisons between the numeric and nominal datasets to identify correlations, and subsequently to use these analyses to develop methods to predict student performance.

From the educational perspective, this would then allow us to perform follow up analyses on the extent to which different attributes can influence student achievement.

Future work includes the application of alternative nominal data analysis techniques to our nominal student data in order to compare the results and evaluate the advantages and disadvantages of these techniques in comparison with those of the technique deployed.

The novel nominal data analysis technique may provide a useful additional tool in the analysis of nominal data. We have shared the technique and corresponding MATLAB code with colleague researchers to gain further feedback on its usage and ideas on how to increase the sophistication of the method. Please contact us for a copy of the code.

REFERENCES

E. Wakelam, A. Jefferies, N. Davey and Y. Sun, “The potential for using artificial intelligence techniques to improve E-learning systems”, 2015.

P. Cortez and A. Silva, “Using data mining to predict secondary school student performance”, in A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), Porto, Portugal, April, 2008, pp. 5-12.https://archive.ics.uci.edu/ml/datasets/Student+Performance#

V. Ramesh, P. Parkavi and K. Ramar, "Predicting student performance: a statistical and data mining approach". International journal of computer applications 63, no. 8, 2013, pp. 0975 – 8887.

P. Bhalchandra et al, "Prognostication of student’s performance: An hierarchical clustering strategy for educational dataset." In Computational Intelligence in Data Mining—Volume 1, Springer India, 2016, pp. 149-157.

D. Fatima, S. Fatima, “A survey on research work in educational data mining.” IOSR Journal of Computer Engineering (IOSR-JCE), 17, 2015.

T. Hastie, R. Tibshirani, J. Friedman and J. Franklin, “The elements of statistical learning: data mining, inference and prediction.” The Mathematical Intelligencer. doi:10.1007/BF02985802, 2005.

D. Clow, "An overview of learning analytics." *Teaching in Higher Education*, 2013, pp. 683-695.

A. Agresti, “Categorical data analysis.” Vol. 996. New York: John Wiley & Sons, 1996.

P. M. Bentler,, and D. G. Bonett, “Significance tests and goodness of fit in the analysis of covariance structures.” Psychological bulletin, 88(3), 588, 1980.

P. Ashrafi, “Predicting the absorption rate of chemicals through mammalian skin using Machine Learning algorithms.” (Ph.D. unpublished). University of Hertfordshire, 2016.

B. Fritzke, "A growing neural gas network learns topologies." Advances in neural information processing systems 7, 1995, pp. 625-632.

A. Hocking, J. Geach, Y. Sun, N. Davey, N. Hine, "Unsupervised image analysis & galaxy categorisation in multi-wavelength Hubble space telescope images", Proceedings of the ECMLPKDD 2015 Doctoral Consortium (ECML 2015), 2015, pp. 105-114.

E. Fatemizadeh, C. Lucas and H. Soltanian-Zadeh, "Automatic landmark extraction from image data using modified growing neural gas network." Information Technology in Biomedicine, IEEE Transactions on 7, no. 2, 77-85, 2003.

G. Liu, and X. Wang, "An integrated intrusion detection system by using multiple neural networks." IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 22-27.

P. Cortez, and A. Silva, “Using data mining to predict secondary school student performance.” In the Proceedings of 5th Annual Future Business Technology Conference, 2008, pp. 5–12.

A. Parimala, “Using machine learning and computer simulations to analyse neuronal activity in the cerebellar nuclei during absence epilepsy.” (Ph.D. unpublished). University of Hertfordshire, 2015.

Appendix L: The Potential for Student Performance Prediction in Small Cohorts with Minimum Available Attrubutes

British Journal of Educational Technology *Vol 0 No 0 2019* 1–24
doi:10.1111/bjet.12836

The potential for student performance prediction in small cohorts with minimal available attributes

Edward Wakelam, Amanda Jefferies, Neil Davey and Yi Sun

Edward Wakelam is a PhD student and lecturer at the University of Hertfordshire researching the application of data mining techniques to learning analytics. Amanda Jefferies is professor of Technology-Enhanced Learning at the University of Hertfordshire. Neil Davey is a principal lecturer and research fellow at the University of Hertfordshire in the area of applied machine learning. Yi Sun is a senior research fellow at the University of Hertfordshire in the areas of applied machine learning and data visualisation. Address for correspondence: Edward Wakelam, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK. Email: e.wakelam@herts.ac.uk

Abstract

The measurement of student performance during their progress through university study provides academic leadership with critical information on each student's likelihood of success. Academics have traditionally used their interactions with individual students through class activities and interim assessments to identify those "at risk" of failure/withdrawal. However, modern university environments, offering easy on-line availability of course material, may see reduced lecture/tutorial attendance, making such identification more challenging. Modern data mining and machine learning techniques provide increasingly accurate predictions of student examination assessment marks, although these approaches have focussed upon large student populations and wide ranges of data attributes per student. However, many university modules comprise relatively small student cohorts, with institutional protocols limiting the student attributes available for analysis. It appears that very little research attention has been devoted to this area of analysis and prediction. We describe an experiment conducted on a final-year university module student cohort of 23, where individual student data are limited to lecture/tutorial attendance, virtual learning environment accesses and intermediate assessments. We found potential for predicting individual student interim and final assessment marks in small student cohorts with very limited attributes and that these predictions could be useful to support module leaders in identifying students potentially "at risk."

Introduction and motivation for study

An ability to predict individual student performance at appropriate points during a module, in particular their likely intermediate and final assessment marks, may provide module leadership with useful guidance on which individuals are "at risk" of failure or withdrawal. This information may then give lecturers and tutors an opportunity to make timely supportive interventions designed to increase the student's likelihood of success. The identification of students "at risk" of failure or withdrawal has become increasingly important to academics, tutors, support staff and institutions, for a variety of reasons. For the students themselves, the failure to achieve

Practitioner Notes

What is already known about this topic

- Learning analytics is a powerful tool in analysing student progress and predicting student outcomes.
- The majority of learning analytics research has focussed upon large data sets comprising of large student cohorts with a significant number of student attributes.
- The analysis of learning analytics data provides course leadership with the ability to identify students "at risk" of failure or withdrawal to allow the opportunity to make positive and timely interventions.
- The aggregation of learning analytics allows course leadership to potentially identify opportunities to improve course presentation and execution.

What this paper adds

Exploration of the potential for predicting student performance in small student cohorts where student data are limited by availability and/or institutional regulation.

- There is some potential for predicting student performance where the student cohort is small and student data are limited to attendance, virtual learning environment accesses and interim assessments. Prediction accuracy is similar to that achieved with large data sets.
- The analyses performed supported module leadership in identifying the need for timely student interventions.
- Random Forest and K-Nearest Neighbours machine learning techniques produced the most accurate prediction results.

Implications for practice and/or policy

Learning analytics can provide institutions with useful supporting data in small student cohort settings, where the availability of individual student data is restricted.

- Machine learning analyses can be provided alongside traditional institutional student performance measures currently made available to module leadership.
- Institutional restrictions often placed on student data availability and privacy are not necessarily a barrier to the deployment of learning analytics.
- The adoption of these methods requires appropriate additions to documented institutional intervention policy.

their potential is a waste, as are the consequent limitations on their future career development. Sometimes worse is the personal stress and trauma they consequently face, alongside the financial impact and the potential consequential effect on their families. In the UK for example, in academic year 2015/16, 6.4% of UK domiciled full-time entrants did not continue in their studies after their first year (Higher Education Statistics Agency, 2018). In Australia and the US, these figures are worse with attrition rates of over 21% (Australian Government Department of Education and Training, 2016) and over 25% (National Center for Education Statistics, 2017). For institutions, the financial impacts can be very significant, compounded by the consequential effects of published statistical measures of student drop-out rates and student satisfaction scores. Universities operate a sliding scale of refund levels to be applied should a student leave the course. In the case of the author's own university, the cost of refunds of full time UK and EU undergraduate student withdrawals can be as high as £27,750, and over 20% higher for non UK/EU students. This is

based upon the recognition that in the vast majority of cases the university place cannot be filled by a suitable replacement and is based upon current annual fees of £9,250. Universities operate in a very competitive environment, and pay considerable attention to their place in league tables and how they may improve their position. The student satisfaction score is an integral part of each institution's overall score and is therefore an area of focus for university management and policies. In the modern HE/University system student non-attendance at lectures and tutorials remains high (Marburger, 2001; Mearman, Pacheco, Webber, Ivlevs, & Rahman, 2014) as course material has increasingly become available on-line and accessible to students 24/7. This reduction in face time between educators and students makes it increasingly difficult for tutors to identify students "at risk" who are struggling with the material or failing to engage. It has always been the case that students are able to request additional lecture/tutorial and face to face time with their tutors.

The application of machine learning techniques to predict student outcomes has made significant progress in recent years (Ashraf, Anwer, & Khan, 2018), providing academic leadership with useful information upon which to consider positive and timely interventions. These techniques have been applied in academic environments where so called "big data" is available (Daniel, 2015). We understand big data to be large student populations and a wide selection of data points (attributes) per student. For example, in the case of the OU, the machine learning analysis operates upon over 32,000 students and 27 attributes per student. However, many university courses/modules are comprised of relatively small student cohorts, often less than 30. A recent study, based upon 67 UK universities, found average class sizes of approximately 20 students (Huxley, Mayo, Peacey, & Richardson, 2018). In addition, academic institutions have legal and ethical obligations to maintain the privacy of individual student data (Corrin *et al.*, 2019), and therefore restrict its availability to prediction algorithms. Opportunities to make useful predictions based upon relatively small student cohorts combined with limited student attributes could provide educators with the capability to identify students "at risk" and make timely supportive interventions.

Research questions and problem formulation

Our study focusses upon two research questions:

Small student cohorts and limited student attributes

Is it possible and useful to predict student performance on courses comprising of relatively small student cohorts, where a very limited set of student attributes are readily available for analysis?

While there is evidence to show that predictions based upon large cohorts can provide educators with useful support in identifying students "at risk" (Heuer & Breiter, 2018), there is little evidence of the value that can be derived where cohorts are small, in this case 23 students. Given that most institutions have a significant number of courses which comprise of these smaller cohorts (Huxley *et al.*, 2018), more research could prove of value. In this case, the data are limited to lecture/tutorial attendance, virtual learning environment (VLE) accesses and five formal interim assessments. This question is critical given institutional protocols and concerns regarding the privacy of student data (see Section "The opportunity to make interventions"), coupled with the ethics of analysing and then taking subsequent action from the results. It is also the case that universities are confused as to whether in providing this data students are in fact giving prior (and legally supportable) approval for their inclusion in learning analytics (LA), and furthermore whether this entitles the institutions to categorise students and to be the catalyst/basis for interventions (Sclater & Bailey, 2015). This is equally true in the case of analytics based upon large student cohorts. Our research question addresses the combination of both small student cohorts and limited attributes.

The opportunity to make interventions

How useful would these analyses be in order to provide course leaders with the opportunity to make timely supportive interventions at appropriate points during the module? In this example, there is a relatively even spread of formal assessments throughout the duration of the course, including two at an early stage.

Experiment design

We apply and compare three machine learning techniques, Decision Tree (DT), K-Nearest Neighbours (KNN) and Random Forest (RF) analyses to analyse and predict student performance, applied at appropriate points during module delivery. These points were selected to coincide with intermediate assessments. DT, KNN and RF methods were selected given their ability to perform well when some values are missing (Quinlan, 2014) and their widespread core use in LA research (Ashraf *et al*, 2018). Given that our experiment is designed to analyse student performance breakdown, missing values may be expected. In the case of our experiment, missing values occur where a student chooses not to take part in an interim assessment. For example, only the highest two of the three multiple choice assessments (see Table 5) count towards the student's final mark and in some cases students who scored highly in the first two of these assessments chose to not sit the third. After module completion, we applied RF analysis retrospectively at each intermediate assessment point to make overall module score predictions and evaluate their accuracy.

Literature review

We have structured our literature review to address each of the two research questions in turn.

Small student cohorts and limited student attributes

An inability to identify and consequently successfully support students "at risk" of failure or withdrawal presents two serious threats to universities. Firstly, the consequences of already budgeted student fees disappearing from university revenues are significant as can be seen by the percentages of student withdrawals. For example, the UK Higher Education Statistics Agency (HESA, 2018) performance indicators show that the percentage of full-time students not continuing after one year of study who started in 2015/16 was 6.4%. In the case of part-time students, the figure was 34.2%. In the case of American University students, Lin, Yu and Chen (Lin, Yu, & Chen, 2012) noted that predicted retention probability decreases from around 70% for a representative full-time student to 57% for a part-time student. In the case of open, distance environments retention and progression has been established to be a greater issue than for traditional full-time campus-based students Simpson (2006, 2013). Secondly, student satisfaction scores are an integral part of the scoring mechanism that determines a university's place in national and global rankings. The impact of these scores on rankings has been shown to be greater for more able students, for universities with entry standards in the upper-middle tier, and for subject departments facing more competition from other universities (Gibbons, Neumayer, & Perkins, 2015).

The prediction of student outcomes is a core component of the field of LA. LA is defined as " the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Ferguson, 2012). The overwhelming focus on LA in higher education has been devoted to the analysis of "big data" (Ashraf *et al.*, 2018) where the data comprises very large student cohorts and a large number of student data items.

In our experiment, we are interested in the application of LA for the prediction of intermediate and final student assessment marks, where the student cohort is small and with limited attributes. In order to provide ourselves with appropriate benchmarks for comparison we now discuss published comparative prediction accuracies across a variety of techniques, applied to large student cohorts with multiple student attributes.

In the case of large data sets, a variety of student attributes are used in the analyses summarised, including personal and admission data as well as previous educational records (Table 1) cited from Ashraf *et al.*, 2018.

A comparison of various data mining techniques (Ashraf *et al*, 2018) to predict student module marks using regression methods demonstrates achieved student prediction accuracy levels ranging from 50% to 97%. Accuracy is measured as the percentage accuracy of the prediction versus the actual student result. Accuracy levels are shown by algorithm (Table 2) and by summary attributes and algorithm (Table 3) cited from Ashraf *et al.*, 2018. These analyses included student numbers in excess of 10,000 and 77 attributes in some cases.

In their analysis of LA and interventions publications between 2007 and 2018, Wong and Li selected 23 case studies highlighting the measured benefits of LA in distance learning institutions (Wong & Li, 2018).

Table 1: Student attributes

Criteria	*Details*
Student demographic information	Age, gender, region, residence, guardian info
Previous results	Cleared certificates, scholarships and results
Grades	Recent assignment results, quizzes, final exam, CGPA, attendance
Social network details	Interaction with social media websites
Extra-curricular activities	Games partitions, sports, hobbies
Psychometric factor	Behaviour, absence, remarks

Table 2: Prediction accuracy by algorithm

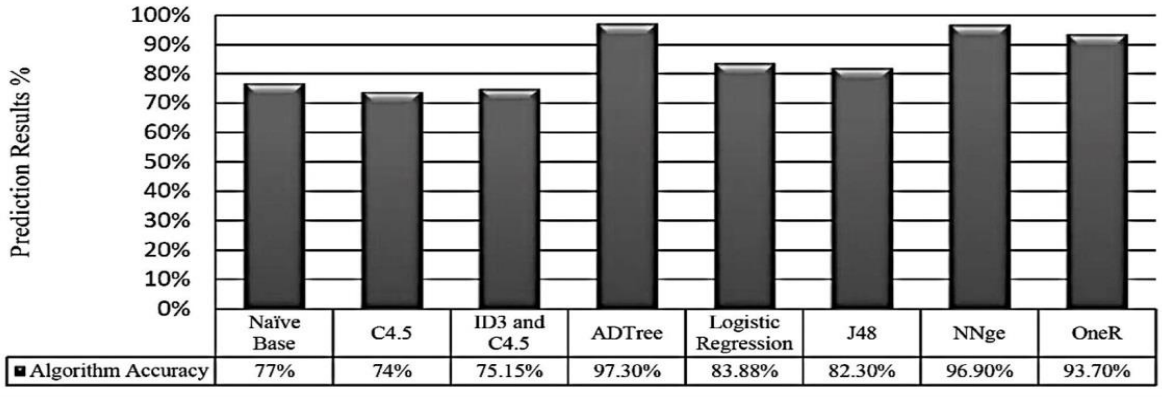
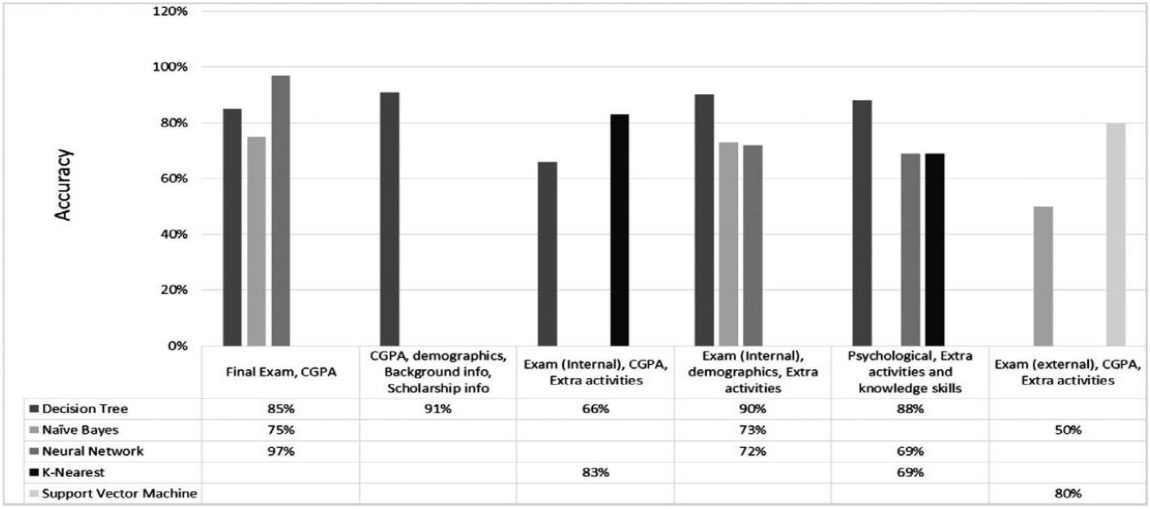


Table 3: Prediction accuracy by summary attributes and algorithm



	Final Exam, CGPA	CGPA, demographics, Background info, Scholarship info	Exam (Internal), CGPA, Extra activities	Exam (Internal), demographics, Extra activities	Psychological, Extra activities and knowledge skills	Exam (external), CGPA, Extra activities
Decision Tree	85%	91%	66%	90%	88%	
Naïve Bayes	75%			73%		50%
Neural Network	97%			72%	69%	
K-Nearest			83%		69%	
Support Vector Machine						80%

There is some evidence that interim assessment as part of the overall course assessment is a strong predictor of student success (Sclater, Peasgood, & Mullan, 2016). Case studies included in this report also identify a student's VLE accesses as a better predictor of success than their historical or demographic data. As with the majority of research conducted, these case studies measured very positive impacts from resulting interventions. A recent study (Heuer & Breiter, 2018) analysing student VLE activity across 22 courses and 32,593 OU students found student VLE accesses to be an important indicator of student performance. An experiment conducted on 200 students over a two-year period at Manchester Metropolitan University made extensive use of VLE usage to determine how to improve the design of learning environments (Stubbs, Martin, & Endlar, 2006).

The opportunity to make interventions

The objective of LA in this instance is to offer tutors the opportunity to identify and support the need to make timely interventions where a student's success is potentially "at risk." The LA cycle is shown in Figure 1 below (Ferguson & Clow, 2017).

In the UK the Open University (OU) is a world leader in the collection, intelligent analysis and use of large scale student analytics. It provides academic staff with systematic and high quality actionable analytics for student, academic and institutional benefit (Rienties, Nguyen, Holmes, Reedy, 2017). Rienties and Toetenel's, 2016 study (Rienties & Toetenel, 2016) identifies the importance of the linkage between LA outcomes, student satisfaction, retention and module learning design. These analytics are often provided through dashboards tailored for each of academics and students (Schwendimann *et al.*, 2017).

The OU's world-class Analytics4Action initiative (Rienties, Boroowa, Cross, Farrington-Flint *et al.*, 2016) supports the university-wide approach to LA. In particular, the initiative provided valuable insights into the identification of students and modules where interventions would be beneficial, analysing over 90 large-scale modules over a two-year period. Analytics4Action identifies

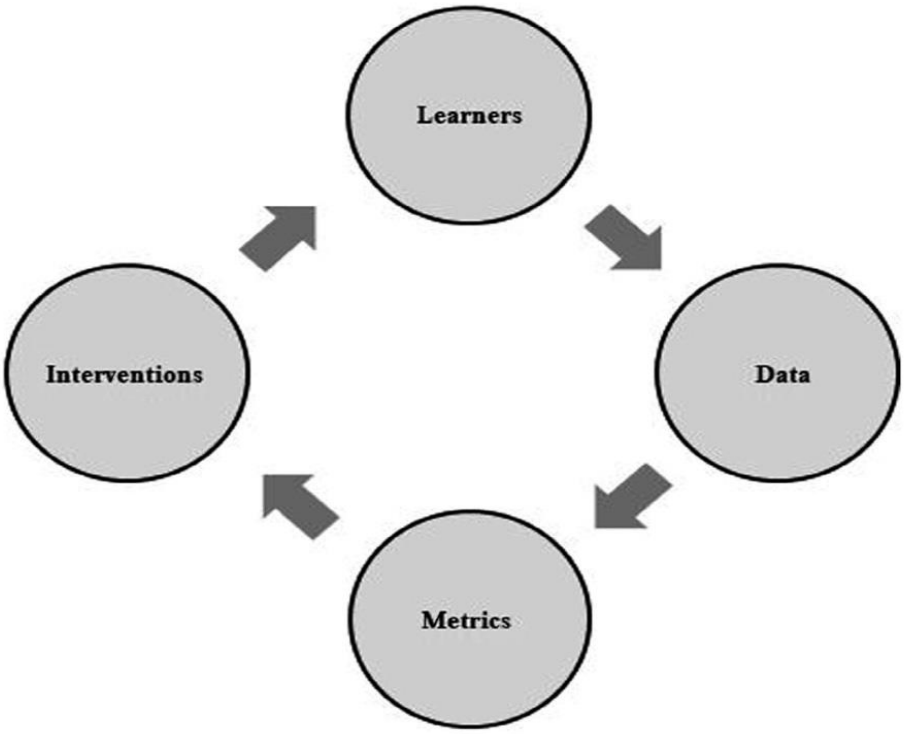


Figure 1: The learning analytics cycle

six phases for teachers and institutions to follow to successfully convert LA outcomes into actionable and impact-measurable interventions (Rienties, Boroowa, Cross, Kubiak *et al.*, 2016). The deployment of LA establishes the need and opportunity for student and module interventions (Clow, 2012). The study concludes that the faster the feedback loop to students, the more effective the outcomes. This is often an iterative process allowing institutions to understand and address systematic issues. Choi and colleagues (Choi, Lam, Li, & Wong, 2018) summarise the pros and cons of alternative intervention methods (Table 4), their study highlighting the benefits to staff faced with limited time and resources.

It is important to note that LA also provide institutions with the opportunity to address systematic issues with individual modules, as referenced above "module interventions" (Clow, 2012). Legal, ethical and moral considerations in the deployment of LA and interventions are key challenges to institutions. They include informed consent, transparency to students, the right to challenge the accuracy of data and resulting analyses and prior consent to intervention processes and their execution (Slade & Tait, 2019). These are well documented in a number of research papers including Pardo and Siemens (2014) and de Freitas *et al.* (2015). In addition, a comprehensive literature review of 86 publications was commissioned by Jisc (formerly the Joint Information Systems Committee, who provide UK universities and colleges with shared digital infrastructure and services including LA), to discuss the challenges faced by institutions and provide the background for a future code of practice (Sclater & Bailey, 2015). A discussion on ethical and data privacy issues in LA based on three studies in higher education and primary school contexts

Table 4: Pros and cons for the commonly-used intervention methods

Method	*Pros*	*Cons*
Email	• Least expensive • Allows personalisation via mail merge	• Students may easily overlook the message due to too many spam emails
Phone call	• Good for emergency matters – two-way synchronous communications	• Students may not be available and sometimes feel offended
Instant messaging	• Preferred communication channel for many students	• More costly than email as it requires one-to-one communications
LMS post & news	• Facilitates many-to-many asynchronous communications	• Requires students to login to the LMS and may overlook the posts and news
Group consultation	• Effective communication • Good for timid students	• Usually needs making appointments in advance and expensive for instructors
Face-to-face consultation	• Effective communication • One-to-one consultation	• Most expensive and usually needs to make appointments in advance
Video recording	• Effective instruction • Not restricted by time	• Substantial initial effort to record the instructions
Peer review	• Encourages critical evaluation • Students can learn from each other	• Requires good question design • Often conducted in class
E-tutorial	• Supplementary instructions available 24/7 (e.g. MyMathLab and MyStatLab developed by Pearson Publishing) • Suitable for highly motivated students	• May incur a price for students or instructors

(Rodríguez-Triana, Martínez-Monés, & Villagrá-Sobrino, 2016), specifically focusses on tutor-led approaches. Legislation has been in place for over two decades, specifically the European Data Protection Directive 1995 and the UK Data Protection Act 1998. More recently, General Data Protection Regulation (Guide to the General Data Protection Regulation, 2018) sets out the legal data protection principles which institutions and organisations are responsible for adhering to. In addition, despite their algorithmic accuracy intentions, there is growing research into the potential for machine learning approaches to introduce bias, such as class, gender and ethnicity (Wilson *et al.*, 2017).

Module description

The selected course instance is a Level 6 (Final Year undergraduate) Computer Science module, duration 15 weeks (including a 3 week vacation period and 2 weeks allocated for submission and review of each of the two final assessments) comprising five intermediate summative assessments and no final examination. Each week students are expected to attend a two-hour lecture and one-hour tutorial. During the course of the module, there are 10 lectures and 9 tutorials. Three EVS (Electronic Voting System) in-class tests are included, with the best two results counting towards the final overall module assessment (see Table 5). The module has a profile of early

Table 5: Module assessments

Week no.	*Name*	*Description*	*Number of weeks to complete assessment*	*Submit on week no.*	*Result publication week no.*	*Percentage contribution to final result*
1	EVS1	Multiple choice	Immediate	4	4	5%
2	EVS2	Multiple choice	Immediate	6	6	5%
3	EVS3	Multiple choice	Immediate	10	10	5%
4	Group Presentation	Group work and presentation	6	11	12	40%
5	Individual Report	Technical Report	8	15	18	50%

"low stakes" assessments with "higher stakes" assessments later in the module. The module VLE comprises of eight sections, including the course guide for example, however student focus was overwhelmingly on the News and Teaching sections.

Note that only the highest two scores of the three EVS results contribute to the final result.

Dataset description

The student cohort is 23. For each student the attributes collected comprise attendance at lectures/tutorials, VLE accesses and intermediate assessment results spread throughout the module (Table 6). Ethics approval limited analysis to dynamic data collected during course execution. Static attributes such as gender, age, prior academic results were not included.

Methodology

We applied three machine learning techniques, DT (regression), KNN and RF to predict student assessment marks, using only their attendance, VLE accesses, and intermediate summative assessments results. The aim of these techniques is to create a model that takes these input values to predict the value of a target variable, in this case the students' assessment marks.

Summary of Machine learning techniques

Decision Tree

DTs are a tree-like model of successive decisions, where each leaf in the tree is a decision, with its corresponding probability, followed by a consequential branch leading to the next leaf, ultimately leading to a prediction (Horning, 2013).

K-Nearest Neighbours

KNN iteratively searches for the most similar (nearest) data points in a given data set, allowing classification of the target data point and consequently prediction (Zhang, 2016).

Random Forest

RF analysis is an ensemble prediction method which uses multiple DTs and averaging their individual results in order to predict a target variable (Horning, 2013).

Design of experiments to meet research questions

Commencing at module registration, each student's attendance at lectures and tutorials was recorded, both as a simple count and as a percentage of overall module tutorials/lectures to date. As well as cumulative attendance, we recorded the delta increases between the measurement points, which were selected to coincide with intermediate assessments. A continuous count of

Table 6: Student attributes

Attribute	*Data range*
Lecture/tutorial attendance	1–19
Delta increase in attendance from prior period	1%–100%
Cumulative VLE News section accesses	0–unlimited
Cumulative VLE Teaching section accesses	0–unlimited
Cumulative VLE accesses	0–unlimited
EVS1 result	0%–100%
EVS2 result	0%–100%
EVS3 result	0%–100%
Group presentation result	0%–100%
Individual report result	0%–100%

individual student "accesses" on items in the VLE was maintained. Of the 8 sections of the VLE, 99% of student accesses were in only 2 sections, News and Teaching. The News section included all module announcements and weekly reminders of tasks to complete. The Teaching section included all course material. For the purposes of the experiment we included each of these two section accesses in our analyses. Intermediate and final assessment results were recorded for each student. This resulted in the data set shown in Table 6. For each analysis point, each of DT, KNN and RF analyses were carried out and the resultant predictions compared with actual student results and the level of accuracy measured. These analyses included the overall module result at module completion. Regression methods were selected to enable the prediction of an actual assessment mark, as opposed to classification methods which would simply predict a pass or fail. This data mining method is often used in the construction of predictive models (Daniel, 2015). The measurement methods used were percentage relative error/accuracy, Mean Squared Error (MSE) and Correlation Coefficient (CC). Prediction accuracies between the analysis methods were compared. We then repeated our analyses combining the two VLE section accesses (see Table 6) into one total in order to determine sensitivity. The progressive prediction results at each assessment point were shared with the module leader for consideration of potential interventions during module delivery. To provide module leadership with data which could potentially support their choice of intervention approach, tabular and graphical comparative analyses of attendance, VLE accesses and intermediate assessment results were also provided. Additionally, we repeated the prediction analyses at each assessment point, based upon the assessment results data alone, excluding attendance and VLE "accesses" in order to compare results. Upon availability of the overall module result after module completion, we were able to revisit our collected data at each assessment point and perform overall module result prediction analyses at each point. We selected RF for these analyses given that it delivered the most accurate predictions in our earlier analyses. Upon module completion, the correlation between all assessments, including overall module results was investigated.

Performance measurement

Percentage relative accuracy is measured as the percentage accuracy of the prediction compared to the actual student result. This permitted a direct comparison with the measurement method used by Ashraf *et al.*, 2018 which compared the results of various data mining techniques, as described in Section "The opportunity to make interventions." MSE measures how close a prediction (regression) line is to the set of actual data points, by calculating the distances from the points to the prediction line (distances are the "errors"), squaring them and calculating their average (mean). The squaring removes any negative signs as well as giving more weight to the larger differences. CC measures how strongly variables are related to each other by dividing their covariance by the product of their standard deviations. A CC of +1 indicates a perfect positive correlation, which means that as variable X increases, variable Y increases and while variable X decreases, variable Y decreases. A CC of −1 indicates a perfect negative correlation. For the purposes of identifying the strongest overall correlations for each analysis technique we calculate the average using absolute CC values.

Experimental results

Research question 1

The value and usefulness of prediction based upon small student cohorts (in this case <30) and where organisational barriers limit the availability of student data. In this case, our data only includes Attendance, VLE accesses and assessment marks. We summarise our results under each of machine learning analyses and traditional statistical methods.

Machine learning analyses

For each of three prediction accuracy measures, Relative % Accuracy, MSE and CC, we present the results of each of DT, KNN and RF analyses, carried out at each assessment point (Tables 7–9). In each case, this includes both the analyses results where VLE News and Teaching accesses are included as separate attributes and where they are combined as one attribute. Prediction accuracy is calculated as 100%—Absolute value of (Actual assessment result—predicted result)/100%. The results of each technique are then discussed.

The overall module result is an arithmetic combination of the intermediate assessment results (see Table 5) and, therefore, we would expect all the prediction methods at the module result assessment point to deliver the most accurate results. This is clearly the case with accuracy between 81% and 91%, averaging 86%. The less than 100% accuracy in each case may be explainable by a combination of inaccuracies in the prediction techniques used and the influence of attendance and VLE access data. RF and KNN (K = 3) with VLE accesses combined delivered the highest average prediction each with accuracies of 75%. Importantly for potential intervention opportunities, predictions at each of the intermediate assessment points using these analysis techniques, although mixed (between 56% and 88%) were promising in several cases, with accuracies at 70% or above at 9 of the 12 points. The least accurate results were delivered by DT Regression and KNN (K = 1) with VLE accesses combined, averaging 65% and 66%, respectively.

As with our Relative % Error measure, the most accurate prediction results (in the case of MSE these are the closest results to zero) are as expected at the overall module result assessment point. At this point, MSE values are between 0.01 and 0.03. Similarly to Relative % Error measure, RF and KNN (K = 3) with VLE accesses combined delivered the most accurate prediction results, excluding the overall module result predictions, with average MSE values of 0.046 and 0.047 respectively. The least accurate results were delivered by KNN (K = 1) with VLE accesses combined, DT and KNN (K = 1) with average MSE values of 0.08, 0.07 and 0.07 respectively.

As with average % accuracy and MSE, CC prediction results are strongest at the overall module result assessment point, with CC values between 0.05 and 0.74. However, in the case of CC, it is DT with VLE accesses combined that delivers our strongest prediction results with an average CC of 0.4, followed by RF with VLE accesses combined and KNN, K = 3 each with an average CC of 0.29. The least accurate results were delivered by KNN, K = 1 and K = 2, with VLE accesses combined giving us CC values of 0.13 and 0.17 respectively. The remaining analysis techniques delivered promising prediction results with CC values between 0.22 and 0.28. In order to investigate the corresponding effect of attendance and VLE access data, we repeated the analyses using only the assessments and excluding all other data. The results were mixed with only very small variations leading us to believe that inaccuracies in the prediction techniques themselves are the major contributor. We present an illustrative subset of the results (Table 10).

Results including all attributes are shown first and results using the assessment results only (ie excluding attendance and VLE accesses) are shown second. We can see that the comparative results are mixed. Recommendations for further work include investigating the predictive effect of cumulative multi-year analyses on the inclusion of attendance and VLE accesses data. After module completion, we performed an overall module result prediction analysis at each assessment point, using RF analysis (Table 11).

We obtained average student final result prediction accuracies of between 82% and 86% using RF analyses. However, the variance between individual student predictions and their actual final result at each assessment point was high, with accuracies ranging from 11% to 99% (Table 12). MSE and CC accuracies performed in line with relative % accuracy analyses.

Table 7: Prediction accuracy measured by relative % accuracy

Relative % accuracy	*EVS1*	*EVS2*	*EVS3*	*Group presentation*	*Individual report*	*Module result*	*Average % Accuracy*	*Ave % Accuracy (Excl. Module result)*
Decision Tree	72%	33%	57%	74%	64%	90%	65%	60%
Regression								
Decision Tree	77%	32%	57%	96%	69%	88%	70%	66%
Regression (Combined VLE Clicks)								
K Nearest	71%	54%	52%	90%	70%	88%	71%	67%
Neighbour, K = 1								
K Nearest	73%	46%	52%	89%	57%	81%	66%	63%
Neighbour, K = 1 (Combined VLE Clicks)								
K Nearest	74%	49%	66%	86%	74%	89%	73%	70%
Neighbour, K = 2								
K Nearest	74%	58%	63%	89%	69%	81%	72%	71%
Neighbour, K = 2 (Combined VLE Clicks)								
K Nearest	74%	55%	74%	74%	72%	89%	73%	70%
Neighbour, K = 3								
K Nearest	76%	60%	68%	88%	73%	82%	75%	73%
Neighbour, K = 3 (Combined VLE Clicks)								
Random Forest	80%	56%	70%	81%	71%	91%	75%	72%
Random Forest (Combined VLE Clicks)	80%	50%	65%	90%	71%	86%	74%	71%

Table 8: Prediction accuracy measured by mean squared error

Mean Squared Error	EVS1	EVS2	EVS3	Group presentation	Individual report	Module result	Ave MSE	Ave MSE (Excl. Module result)
Decision Tree Regression	0.0767	0.1489	0.1051	0.0411	0.0603	0.0137	0.0743	0.0743
Decision Tree Regression (Combined VLE Clicks)	0.0459	0.1435	0.1019	0.0127	0.0603	0.0158	0.0634	0.0634
K Nearest Neighbour, K = 1	0.0806	0.0969	0.1464	0.0216	0.0611	0.0213	0.0713	0.0713
K Nearest Neighbour, K = 1 (Combined VLE Clicks)	0.0736	0.1101	0.1426	0.0247	0.0838	0.0315	0.0777	0.0777
K Nearest Neighbour, K = 2	0.0527	0.0982	0.0781	0.0261	0.046	0.0217	0.0538	0.0538
K Nearest Neighbour, K = 2 (Combined VLE Clicks)	0.0634	0.0755	0.0841	0.0229	0.0586	0.032	0.0561	0.0561
K Nearest Neighbour, K = 3	0.0527	0.0842	0.0591	0.0334	0.0532	0.0181	0.0501	0.0501
K Nearest Neighbour, K = 3 (Combined VLE Clicks)	0.0613	0.0669	0.0692	0.0028	0.0526	0.0289	0.0470	0.0470
Random Forest	0.0341	0.0657	0.0756	0.0359	0.0461	0.0191	0.0461	0.0461
Random Forest (Combined VLE Clicks)	0.0465	0.0922	0.0726	0.0189	0.0542	0.0196	0.0507	0.0507

Table 9: Prediction accuracy measured by correlation coefficient

Correlation Coefficient	*EVS1*	*EVS2*	*EVS3*	*Group presentation*	*Individual report*	*Module result*	*Ave CC*	*Ave CC (Excl. Module result)*
Decision Tree Regression	−0.0912	−0.4518	0.0706	−0.0224	0.1732	0.7386	0.2580	0.1618
Decision Tree Regression (Combined VLE Clicks)	0.2754	−0.5090	−0.0426	0.7853	0.1732	0.6942	0.4133	0.3571
K Nearest Neighbour, K = 1	0.042	0.0843	0.2329	0.558	0.0262	0.5363	0.2466	0.1887
K Nearest Neighbour, K = 1 (Combined VLE Clicks)	−0.0295	−0.0083	−0.1433	0.4638	0.2651	0.1394	0.1749	0.1820
K Nearest Neighbour, K = 2	−0.1536	−0.34	0.0701	0.3899	0.1541	0.5424	0.2750	0.2215
K Nearest Neighbour, K = 2 (Combined VLE Clicks)	−0.0295	0.1093	0.0683	0.5106	−0.0404	0.0455	0.1339	0.1516
K Nearest Neighbour, K = 3	−0.0876	−0.3019	0.2973	0.146	−0.1536	0.7391	0.2876	0.1973
K Nearest Neighbour, K = 3 (Combined VLE Clicks)	−0.3137	0.0535	0.1928	0.5069	−0.0402	0.2075	0.2191	0.2214
Random Forest	0.4165	0.1443	0.0648	0.1352	0.1711	0.5985	0.2551	0.1864
Random Forest (Combined VLE Clicks)	0.0438	−0.2986	0.1289	0.62	0.0732	0.579	0.2906	0.2329

Table 10: Comparison of analyses including all attributes against those using assessment results only

Analysis Technique	*Prediction accuracy measure*	*EVS3*	*Group presentation*	*Individual report*
K Nearest Neighbour, K = 3 (Combined VLE Clicks)	Relative % Accuracy	74%/67%	74%/82%	72%/73%
	Mean squared error	0.0591/0.0553	0.0334/0.0427	0.0532/0.0498
	Correlation coefficient	0.2973/0.4164	0.146/−0.2093	−0.1536/0.1254

Correlations between assessments

An analysis of the cross-correlation between each of the interim assessments and the overall module result (Table 13) shows moderate, high and very high correlations with the overall module result. Of these five interim assessments, we found high and very high correlations between the two major interim assessments (Group Presentation and Individual Report) and the overall module result. The initial three interim assessments were all moderately correlated with the overall module result.

Graphical analyses to support potential interventions

Example graphical analyses performed at EVS3 and individual report assessment points are discussed and shown below (Figures 2–7). In each figure, the student identification number (1 to 23) is labelled on the x axis. Note that student 10 withdrew from the module prior to assessment commencement.

Machine learning predictions for students 12 and 14 highlighted 62% and 97% negative disparities with their actual and expected progress raising concerns with module leadership. We can see form this table that in both cases their attendance records are very high and therefore not a cause for leadership concern. Student 22 had scored well in EVS1 and EVS2 assessments and given that the best two of the three assessments only are included chose not to take EVS3.

A glance at this chart shows that both student 12 and student 14 are registering average VLE accesses and this could be an area for concern and potential intervention.

As above, using students 12 and 14 as our examples, we can see that their high average EVS1 and EVS2 results indicate why machine learning prediction disparities were evident.

Machine learning predictions for students 19 and 20 highlighted 159% and 179% negative disparities with their actual and expected progress raising concerns with module leadership. We can see from this table that both students are maintaining average attendance.

A consideration of this chart shows that both student 19 and student 20 are registering above average VLE accesses but may still be an area for potential intervention.

As above, using students 19 and 20 as our examples, we can see that both have good average assessment results to date. In this case, module leadership considered intervention unnecessary.

Research question 2

We now consider the value and usefulness of prediction analyses for intervention opportunities. For these analyses to be of value for interventions they must be available to module leadership while sufficient time is left for successful interventions to be made and any consequent positive effects to be achieved by the student. The early and mid-timed assessments in the

Table 11: Module result prediction at each assessment point

Analysis technique	*Prediction Accuracy Measure*	*EVS1*	*EVS2*	*EVS3*	*Group presentation*	*Individual report*	*Average accuracy*
Random Forest	Relative % accuracy	82%	82%	86%	83%	85%	84%
	Mean squared error	0.0325	0.0334	0.0253	0.0323	0.0206	
	Correlation coefficient	−0.0207	0.1114	0.3763	0.1866	0.5483	

Table 12: Range of individual student final result percentage prediction accuracies at assessment points

Prediction accuracy	*EVS1*	*EVS2*	*EVS3*	*Group presentation*	*Individual report*
Lowest	38%	11%	52%	28%	35%
Highest	98%	98%	100%	99%	99%

Table 13: Assessments correlation matrix

	EVS1	*EVS2*	*EVS3*	*Group presentation*	*Individual report*	*Overall module result*
EVS1	1.00	0.53	0.63	0.47	0.44	0.55
EVS2		1.00	0.59	0.60	0.60	0.66
EVS3			1.00	0.42	0.42	0.51
Group Presentation				1.00	0.73	0.90
Individual report					1.00	0.95
Overall module result						1.00

Key	
Very highly correlated	(0.9 to 1.0)
Highly correlated	(0.7 to 0,89)
Moderately correlated	(0.5 to 0.69)
Low correlation	(0.3 to 0.49)

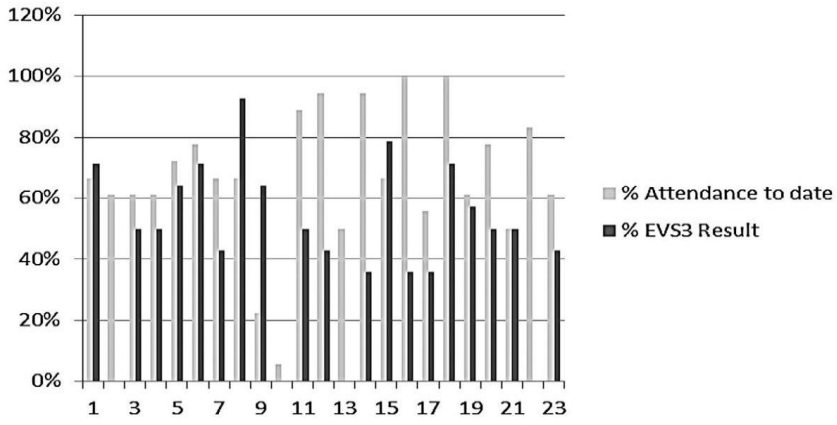


Figure 2: Attendance to date v EVS3 result

selected module provided this opportunity. The progressive prediction analyses conducted may also provide module leadership with useful data in respect of module and assessment design. For example, if our predictions on individual assessments were consistently accurate, it may be that these assessments are adding little value in their current form and require revision. Adaptive learning systems dynamically adjust the number of questions upwards and downwards and dynamically adjust student learning paths depending upon student performance (Wakelam *et al.*, 2015). The graphical analyses (Section "Graphical analyses to support potential interventions") proved useful for module leadership to perform "at a glance" assessments of

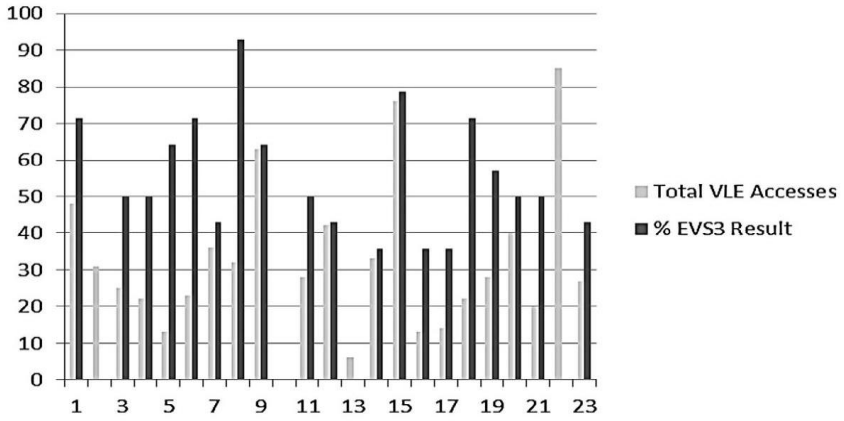


Figure 3: Total VLE accesses v EVS3 result

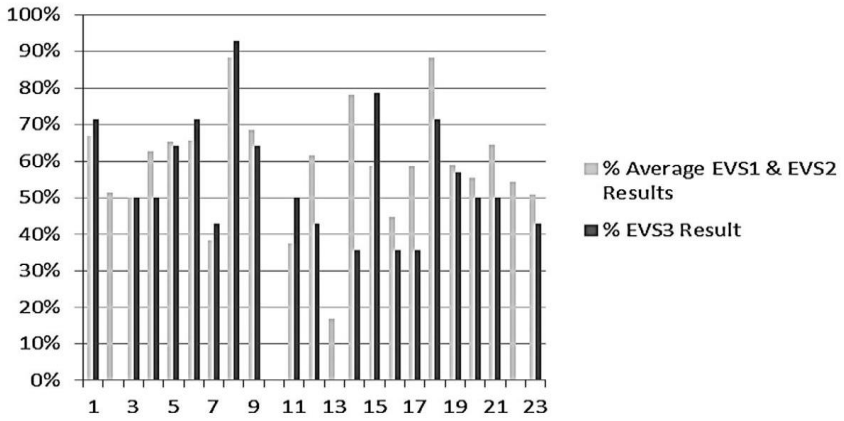


Figure 4: Average of EVS1 and EVS2 results v EVS3 result

student activities. For example, where a student prediction suggests a performance risk, module leadership were able to quickly view their attendance and VLE usage in support of personal experience of the student. This in itself may suggest intervention methods, ranging from encouraging improved attendance or more usage of VLE material. In the case of this module, we were able to review students where machine learning predictions identified potential poor outcomes, supported by "at a glance" comparisons of their attendance, VLE accesses and prior assessment marks. This information coupled by module leadership knowledge of each student through face-to-face lectures and tutorials supported direct interventions, including coaching and the provision of additional teaching material. These interventions may be grouped under the heading of providing additional scaffolding to students. Research conducted by Stubbs *et al.* at Manchester Metropolitan University discusses how a metaframework for assisting the design of learning frameworks to educational designers to support improved learning outcomes (Stubbs *et al.*, 2006).

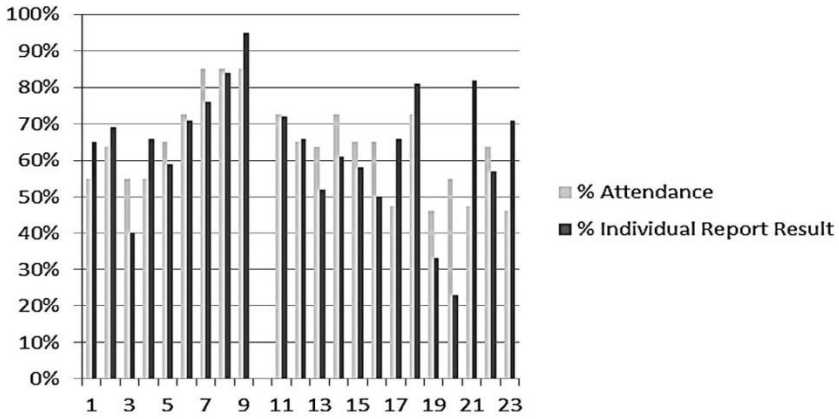


Figure 5: Attendance to date v individual report result

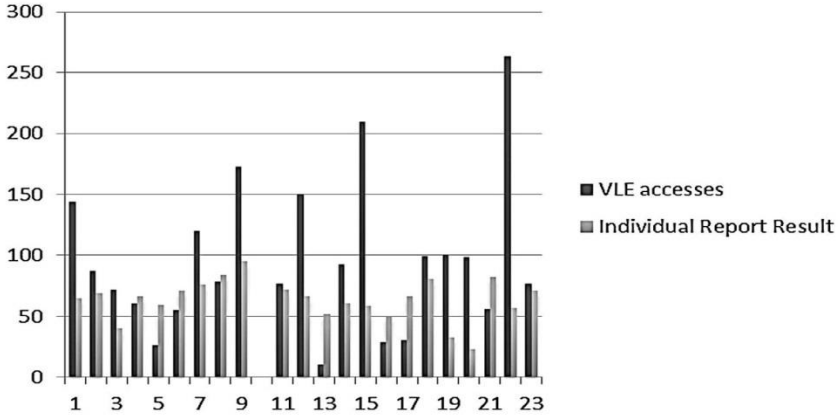


Figure 6: Total VLE accesses v individual report result

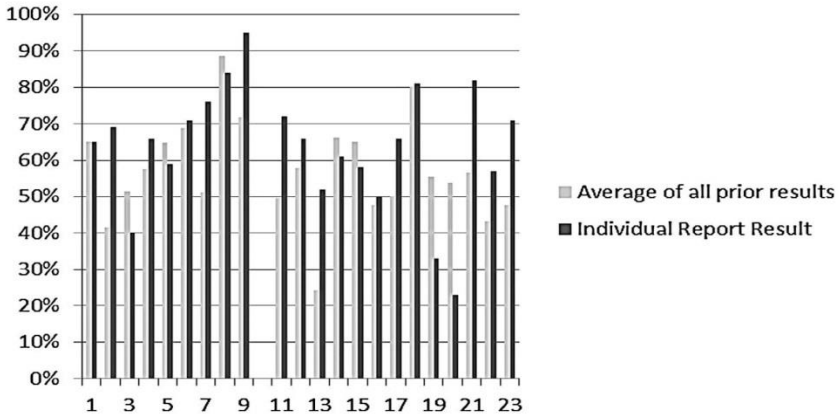


Figure 7: Average of EVS1, EVS2, EVS3 and group presentation results v individual report result

Discussion and conclusions

Research question 1

Is it possible and useful to predict student performance on courses comprising of relatively small student cohorts, where a very limited set of student data is readily available for analysis?

Experimental results show some potential for analysing and predicting student assessment marks on courses comprising relatively small student cohorts, and where only very limited set of student data is readily available for analysis. The average prediction accuracy across all machine learning techniques used was 67%, with KNN and RF prediction accuracy between 66% and 75%. This compares favourably with student prediction accuracy levels achieved across a variety of machine learning techniques applied to large student cohorts with significantly more student attributes (Ashraf *et al.*, 2018). The results in Ashraf and colleagues' study ranged from 50% to 97% (Tables 2 and 3). Importantly for potential intervention opportunities, we obtained some promising results at the point of the third assessment, approximately two thirds of the way through the module, with prediction accuracies of 74% and 70% for KNN and RF Analyses respectively. Reducing the attributes used in our analyses gave us mixed results. Combining VLE News and Teaching accesses into one total had very little effect upon prediction accuracy, in some cases giving a 1% improvement and in others the reverse. Reducing the attributes to only the intermediate assessment results gave us mixed results in comparison with prediction accuracy using all available attributes, hence we could not reliably consider student interventions. Similarly, this provided us with little opportunity to determine the effect of including attendance and VLE accesses on prediction accuracy. We believe that the inclusion of all available attributes may be considered as at least benign to our analyses. There is some evidence (Heuer & Breiter, 2018) that the analysis of VLE accesses alone can be a useful predictor of student performance. Future work accumulating year-on-year module data to investigate the effects on prediction accuracy of multi-year data may provide further insight. As we might expect, the final assessment, the student's Individual Report which is submitted in week 15 of 18, contributing 50% to their overall mark, correlated very highly (CC 0.95) with their overall module result. Additionally, the penultimate assessment, the Group Presentation, submitted in week 11 of 15, correlated highly (CC 0.9) with the overall module result. Usefully, for the potential of earlier intervention opportunities, given their early assessment points of weeks 4, 6, 10 of 15, we found moderate correlations (CCs of 0.55, 0.66 and 0.51 respectively) between EVS1, EVS2 and EVS3 and the overall module result. In particular, student usage of VLE material and correlations between attendance and VLE usage on assessment marks provided valuable insights.

Research question 2

How useful would these analyses be in order to provide course leaders with the opportunity to make timely supportive interventions at appropriate points during the module?

The analyses demonstrated three opportunities for module leadership to identify potentially "at risk" students and to consider appropriate timely interventions. These were machine learning analyses at intermediate assessment points, and the identification, post module completion, of which intermediate assessments provided the likeliest indicators of overall module success. Student performance in their third assessment, week 10 of 15, appears to be a useful measure of individual progress. In this experiment, module leadership were then able to review attendance and VLE access patterns for students whose performance was of concern. Alongside personal experience of the student in question an intervention decision could then be made. In the case of the module, our analyses led to module leadership identifying two specific opportunities for direct,

interventions, both following the third assessment, EVS 3. In each case, a student's predicted performance showed a likelihood of failing their next assessment. In case 1, further analysis showed a reduction in tutorial attendance. In case 2, analysis showed a combination of reduced lecture/tutorial attendance coupled with minimal activity in the VLE. This enabled leadership to engage in positive discussions with each student and provide specific guidance on their future studies. A variety of possible interventions are described in Section "The opportunity to make interventions", but could be as simple as evidence based discussions drawing a student's attention to their attendance, arranging additional individual or group lectures/tutorials or the availability of further and focussed supporting material on the VLE. Graphical analyses allowing the visualisation of relationships between attributes provides module leadership with further opportunities to identify any interesting correlations which could support positive interventions. These graphical presentations compared different combinations of attendance, VLE usage and assessment results providing easily referenceable "at a glance" supporting material to machine learning results for module leadership. In the case of the module in this experiment, we found these representations supported intervention decisions. Given their significant mark contribution to the overall module result this was to be expected. Additionally, promising results at the earlier third assessment point gave module leadership the opportunity to consider interventions in time for their effects to be useful.

Implications to practice and/or policy

University expectations are currently that the application of LA necessitates the availability of so-called "big data," in particular, modules with large student cohorts. Our results show that university practice can now usefully consider smaller scale deployments of LA. Where student attributes for analysis are limited to readily available data such as student attendance, VLE accesses and intermediate assessment results, with no inclusion of demographic/personal data, either none, or very limited modifications are necessary to university policies. It is good practice to provide students with a clear explanation of what data are being collected and how the analysis is being done, allowing them to individually opt in or opt out of LA implementations. In addition, alternative intervention methods should be documented and where possible students given the opportunity to express their preferences. For example, dashboard presentation of predictions, system generated emails, offers of face to face supportive meeting with course tutors.

Future work

We plan to perform DT, KNN and RF analyses using classification (i.e. binary prediction of pass or fail), instead of regression, and compare student marks prediction accuracy with the results of this experiment.

In order to investigate the predictive results of the respective scenarios of a wider range of student attributes and of a large cohort size, we would propose to conduct two related experiments. Firstly, where the student cohort is small, but where a wider selection of student attributes is available, for example, prior student module marks and examination results from previously attended institutions. Secondly, where the student cohort is much larger, but with the same student attributes as with this experiment.

It would be valuable to accumulate year-on-year module data to investigate the effects of the inclusion of multi-year data on prediction accuracy, using the same analyses techniques as in this experiment.

The module in our experiment is comprised of a relatively even spread of formal assessments, with two at an early stage. The effects upon prediction accuracy of applying the same experimental

analyses to a module where either there are fewer intermediate assessments or where they are conducted later in the module may be of value.

Finally, a logical next step would be to design and conduct an experiment which tracks and measures any resulting changes in individual student attendance, VLE accesses and assessment scores resulting from academic staff interventions.

Statements on open data, ethics and conflict of interest

The data in this study can be accessed upon request.

Formal ethical approval for the conduct of the experiment described in this study was granted by the university.

The authors confirm that there is no conflict of interest in this study.

References

Ashraf, A., Anwer, S., & Khan, M. G. (2018). A Comparative study of predicting student's performance by use of data mining techniques. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *44*(1), 122–136.

Australian Government Department of Education and Training. (2016). Attrition, Success and Retention. *Higher Education statistics, Appendix 4*.

Choi, S. P., Lam, S. S., Li, K. C., & Wong, B. T. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Journal of Educational Technology & Society*, *21*(2), 273–290.

Clow, D. (2012, April). The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 134–138). New York: ACM.

Corrin, L., Kennedy, G., French, S., Shum, S. B., Kitto, K., Pardo, A., ... Colvin, C. (2019). *The ethics of learning analytics in Australian higher education*. Melbourne: University of Melbourne.

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920.

de Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., & Arnab, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British Journal of Educational Technology*, *46*(6), 1175–1188.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304–317.

Ferguson, R., & Clow, D. (2017, March). Where is the evidence?: a call to action for learning analytics. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 56-65). Vancouver: ACM.

Gibbons, S., Neumayer, E., & Perkins, R. (2015). Student satisfaction, league tables and university applications: Evidence from Britain. *Economics of Education Review*, *48*, 148–164.

HESA. (2018). *Non-continuation summary: UK Performance Indicators 2016/17*. Retrieved from https://www.hesa.ac.uk/news/08-03-2018/non-continuation-summary

Heuer, H., & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*.

Horning, N. (2013). Introduction to decision trees and random forests. *American Museum of Natural History*.

Huxley, G., Mayo, J., Peacey, M. W., & Richardson, M. (2018). Class size at university. *Fiscal Studies*, *39*(2), 241–264.

Lin, T. C., Yu, W. W. C., & Chen, Y. C. (2012). Determinants and probability prediction of college student retention: New evidence from the Probit model. *International Journal of Education Economics and Development*, *3*(3), 217–236.

Marburger, D. R. (2001). Absenteeism and undergraduate exam performance. *The Journal of Economic Education*, *32*(2), 99–109.

Mearman, A., Pacheco, G., Webber, D., Ivlevs, A., & Rahman, T. (2014). Understanding student attendance in business schools: An exploratory study. *International Review of Economics Education, 17*, 120–136.

National Center for Education Statistics. (2017). Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected years, 2006 to 2015. *Digest of Education Statistics*. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_326.30.asp

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology, 45*(3), 438–450.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. San Mateo, CA: Elsevier.

Rienties, B., Boroowa, A., Cross, S., Farrington-Flint, L., Herodotou, C., Prescott, L., & Woodthorpe, J. (2016). Reviewing three case-studies of learning analytics interventions at the Open University UK. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 534–535). Edinburgh: ACM.

Rienties, B., Boroowa, A., Cross, S., Kubiak, C., Mayles, K., & Murphy, S. (2016). Analytics4Action evaluation framework: A review of evidence-based learning analytics interventions at the open university UK. *Journal of Interactive Media in Education, 2016*(1). 1–11.

Rienties, B., Nguyen, Q., Holmes, W., & Reedy, K. (2017). A review of ten years of implementation and research in aligning learning design with learning analytics at the Open University UK. *Interaction Design and Architecture (s), 33*, 134–154.

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*, 333–341.

Rodríguez-Triana, M. J., Martínez-Monés, A., & Villagrá-Sobrino, S. (2016). Learning analytics in small-scale teacher-led innovations: Ethical and data privacy issues. *Journal of Learning Analytics, 3*(1), 43–65.

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., ... Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1), 30–41.

Sclater, N., & Bailey, P. (2015). Code of practice for learning analytics. *Joint Information Systems Committee (JISC)*.

Sclater, N., Peasgood, A., & Mullan, J. (2016). *Learning analytics in higher education*. London: Jisc.

Simpson, O. (2006). Predicting student success in open and distance learning. *Open Learning: The Journal of Open, Distance and e-Learning, 21*(2), 125–138.

Simpson, O. (2013). Student retention in distance education: Are we failing our students? *Open Learning: The Journal of Open, Distance and e-Learning, 28*(2), 105–119.

Slade, S., & Tait, A. (2019). *Global guidelines: Ethics in learning analytics*. Oslo: International Council for Open and Distance Education (ICDE).

Stubbs, M., Martin, I., & Endlar, L. (2006). The structuration of blended learning: Putting holistic design principles into practice. *British Journal of Educational Technology, 37*(2), 163–175.

UK Government. (2018). *Guide to the General Data Protection Regulation (GDPR)*. Retrieved from https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation

Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2015). The potential for using artificial intelligence techniques to improve e-Learning systems.In *ECEL 2015 Conference Proceedings*. Hatfield.

Wilson, A., Watson, C., Thompson, T. L., Drew, V., & Doyle, S. (2017). Learning analytics: Challenges and limitations. *Teaching in Higher Education, 22*(8), 991–1007.

Wong, B. T., & Li, K. C. (2018, July). Learning analytics intervention: A review of case studies. In *2018 International Symposium on Educational Technology (ISET)* (pp. 178–182). Osaka: IEEE.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine, 4*(11), 218.