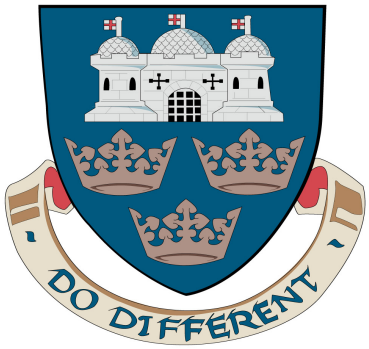
Analytics and Information Management in Higher Education

Zahyah Alharbi

A thesis submitted in fullment
of the requirements for the degree of
Doctor of Philosophy

University of East Anglia
School of Computing Sciences



Abstract

Educational data mining, or the ability to exploit educational data to detect patterns, is an area of increased activity. In this research, we look at the practical application of predictive models in a Higher educational setting in the UK. Firstly, we investigate the use of data mining techniques to highlight performance issues early on and propose remedial actions. We predict good honours outcomes based on data at admission, and some early results from the 1st year.

Secondly, we study more granular predictions at the module level. We compare several data mining techniques in order to build both regression and classification models. One of the difficulties we encounter is that, within our problem, missing data is abundant because students do not always take the same module choices. The problem of missing data is prevalent in many data mining applications and remains challenging. We address this problem in a novel way by using multiple imputation combined with an ensemble setting to produce our models. The results show that all the data mining algorithms that use multiple imputation perform better than those without multiple imputation, both in the cases of classification and regression. The algorithms developed, and in particular Support Vector Machines and Random Forest, give us reasonably accurate predictions that could be used as the basis for a future recommender system to assist with module choice selection.

Lastly, we study how to use the knowledge found in a way acceptable to students and other stakeholders. For this we design a survey questionnaire to understand student views. We also carry out several interviews with students and some key stakeholders to understand any barriers to change and also to identify enablers. We then analyse the collected data and propose recommendations for the final system.

Publications and presentations

A copy of all the following publications is included in Appendix F. The main work in all these publications was conducted by the first author including: proposing solutions, performing the experiments and the interviews, writing the first draft of the papers, retrieving, collecting, and cleaning the datasets used, analysing the results, etc. All the others co-authors efforts have enhanced the quality of the work produced.

Publications:

1. We have published a paper [1] titled " Using data mining techniques to predict students at risk of poor performance" accepted in SAI 2016, the SAI Computing Conference, which took place in London, UK between the 13th and the 15th of July, 2016. The paper has been added to IEEE Xplore on September, 2016. The conference proceedings are indexed by IEEE, Scopus, IET Inspec and Google scholar.
2. A journal paper titled "Multiple imputation and ensembles for student performance prediction in the context of missing data" is under consideration by the Journal of Educational Data mining (JEDM). Note: The paper was submitted to JEDM on 1st May 2017 and the reviewers comments were received on 29th November 2017. The article was reviewed by three experts in the field and we were asked to make some changes and encouraged to consider a resubmission. Some of the changes requested including the addition of public datasets to our experiments and a change of focus on the paper. We have incorporate all the required comments from the reviewers and re-submitted the article on 1st March, 2018.
3. We are also at the moment targeting British Journal of Educational Technology with an article titled "A little knowledge is a dangerous thing? Making data endogenous in higher education."

Presentations:

1. A 20 minutes oral presentation of our paper, "Using data mining techniques to predict students at risk of poor performanc", in SAI, 2016.

2. A 20 minutes oral presentation and discussion on my PhD research "Analytics and Information Management in Higher Education", in British Academy of Management E-business and E-government Special Interest Group (BAM), 28/29 January 2016.

3. A poster representation of the research progress in the CMP Postgraduate day on the 31st of October 2014 with a research abstract published within the event's abstract booklet.

4. A poster representation of the research progress in the CMP Postgraduate day on the 23rd of October 2015 with a research abstract published within the event's abstract booklet.

5. A poster representation of the research progress in the CMP Postgraduate day on the 5th of October 2016 with a research abstract published within the event's abstract booklet.

6. A poster representation of the research progress in the CMP Postgraduate day on the 11th of October 2017 with a research abstract published within the event's abstract booklet.

I would like to dedicate this thesis to my beloved parents, husband, children, sisters and brothers for supporting me all the way!

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Dr Beatriz de la Iglesia, a data mining researcher and Senior Lecturer at the School of Computing Sciences, for her continuous support of my Ph.D. study, and for her patience, motivation and immense knowledge. Her guidance helped me in all of the research and in the writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study. Thank you Beatriz for giving me the opportunity to work with you and for helping me to develop my confidence in being a researcher and a more mature person.

I would also like to thank my second advisor Mr. James Cornford, a Senior Lecturer at Norwich Business School, for the hours he spent discussing and sharing with me a great deal of knowledge with regard to doing this research. I would also like to recognise his helpful feedback and advice in improving the quality of my research, especially the management aspect of it.

Moreover, thanks go to the reviewers of my thesis-related papers for their insightful comments and suggestions, which incentivised me to improve my research from different perspectives. I thank each participant in my survey and qualitative study, for sharing their views with me and trusting me. I learned more than I am able to put on the pages of this thesis. A grateful acknowledgement goes to my sponsor, King Saud University in Riyadh, Saudi Arabia, for the full scholarship that has been awarded to me.

I am grateful for my beloved family; a special thanks go to my beloved parents for all their emotional and financial support. Thank you Mom and Dad for your unconditional love, your encouragement to pursue my dreams and be an independent and hard-working person; you made me the person that I am today and I hope I made you proud. I thank my sister Deya for her unconditional support and for always being there for me when I needed her. I am blessed to have her in my life. I also thank my brother Dr Sami for introducing me to the field of data mining, and for all of his support and being not just my brother but also a second father to me.

Finally, all my love and thanks go to the only man in my life, my soulmate and dearest husband Abdullah. When I started my PhD, this day seemed so far away. Now it's here and I can't believe

that time has passed so quickly and I will be back in your arms and we will reunite as a whole family again. Thank you for your sacrifices and your support for studying abroad and being what I want to be. Your support, endless love and kindness helped me overcome the most difficult times. My two lovely daughters, Shomoukh and Reema, who have grown up watching me studying and juggling between family and work, I love you more than anything and I appreciate all your support and patience during mommy's Ph.D. studies. Thank you for the happiness and joy you have brought to my life. I hope my journey inspires and motivates you to chase after your dreams.

Contents

List of Abbreviations

Abbreviation	Meaning
AIHS	Adaptive and Intelligent Hypermedia System
APL	Accreditation of Prior Learning
BI	Business Intelligence
CART	Classification and Regression Tree
CBE	Computer-Based Education
CBES	Computer-Based Education System
CBLE	Computer-Based Learning Environments
CMP	Computing Sciences School
CMS	Course Management Systems
CPCC	Cophenetic Correlation Coefficient
CRISP-DM	The Cross Industry Standard Process for Data Mining
CSCL	Computer Supported Collaborative Learning
CV	Cross Validation
DM	Data Mining
EDM	Educational Data Mining
EDM	Educational Data Movement
EU	EUropean
EM	Expectation Maximisation
GH	Good Honours
HE	Higher Education
IBM	International Business Machines
ITS	Intelligent Tutoring System
KDD	Knowledge Discovery and Data Mining
KDD	Knowledge Discovery in Databases
KT	Knowledge Tracing
LA	Learning Analytics
LMS	Learning and Management Systems
LRMB	Layered Reference Model of the Brain
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
MI	Multiple Imputation
NBS	Norwich Business School
NGH	Not Good Honours
ODM	Organisational Data Mining
OS	OverSeas
PAM	Partitioning Around Medoids
PRS	Personal Recommender Systems
Ref	reference
RF	Random Forest
RMSE	Root Mean Square Error
RS	Recommender system
SVM	Support Vector Machine
SNA	Social Network Analysis
UCAS	Universities and Colleges Admissions Service
UEA	University of East Anglia
UK	United Kingdom

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Data Mining and Big Data

Big data is a phrase that refers to the growth of the volume of data that is obtainable by an organisation and the potential to find new observations when analysing such data. IBM scientists divide big data into four dimensions: volume (scale of data), variety (different forms of data), velocity (analysis of streaming data), and veracity (uncertainty of data) [3]. There is debate as to what actually constitutes big data, but there is no debate about the current state of affairs: all organisations are gathering increasing amounts of varied and complex data. Organisations therefore have a challenging task in analysing and making sense of this ever growing data, and they require semi-automatic solutions. Data mining (DM) can help organisations to find useful information from large amounts of data in order to improve decision making [4].

Data mining has its origins in computer science, statistics, machine learning and artificial intelligence [5]. There are several different DM tasks, such as classification, clustering and association rule mining. Each of these tasks can be utilised to discover hidden patterns and information by quantitatively analysing a large amount of data. Data mining is an explorative process, but can be employed for confirmative investigations [6]. It is unlike other analysis and search techniques, because it is highly exploratory, while other techniques are usually confirmatory and hypothesis-driven. Data mining tasks can have three objectives [7]: descriptive when DM is applied to increase the understanding of the data; predictive when data is used for forecasting or predicting the future, which might inform the decision-making process; and prescriptive when DM is oriented at automating the decision-making process.

Data mining is also considered as one phase in an overall knowledge discovery (KDD) process

[8]. The difficulties in dealing with large amounts of real-world (messy, uncertain, complex) data have led the data analysis community to build a KDD process for DM activities.

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a life cycle process that describes the approaches for developing DM models [9]. The CRISP-DM process is essential since it provides particular techniques and tips on how to move from the first phase, understanding the business data, to the last phase, the deployment of a DM model. CRISP-DM divides the DM process into six main phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment . The advantages of CRISM-DM are that it is a widely supported and non-proprietary standard, and it offers a solid framework for direction and assistance in DM. The model also contains templates to support analysis. A number of pieces of educational data mining (EDM) research [10, 11, 12] have utilised this process, but it is not always clearly stated.

DM has been employed in several areas of human knowledge, for example in medicine [13, 14], finance and information management [15, 16], banks [13], the retail industry [13, 17], telecommunications [13] and the exploitation of information from the web [18]. However, it is only recently receiving consideration and notice in the educational context [19]. Educational data mining is an area of research that includes the application of DM to resolve educational issues and concerns. Educational data mining has its own challenges due to the nature of the data and the environment in which it is collected.

1.2 Educational Data Mining

Educational data mining can be defined as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" [20, p.1]. In other words, EDM focuses on almost any type of data in educational organisations. It depends on a number of reference disciplines and there will be additional growth in the interdisciplinary nature of EDM [21].

EDM has drawn upon ideas from organisational data mining (ODM). ODM concentrates on helping institutions to sustain a competitive advantage [22]. The main distinction between DM and ODM is that ODM depends on organisational theory as a reference discipline. Organisations that take their data and transform it into valuable knowledge and information in an efficient way should achieve enormous benefits such as improved decision making, enhanced competitiveness and potential financial gains. This is an essential relationship because the focus of study within EDM

can explore and investigate phenomena at various levels of analysis, i.e. at the organisational, societal, unit, or individual level.

EDM gives greater emphasis to quantitative analyses using artificial intelligence, statistics and machine learning algorithms. Qualitative methods, including document analysis and interviews, can also be utilised to support research in EDM. However, the dominant type of study is quantitative, with outcomes presented in the form of clusters, predictions, associations, or classifications. Therefore, EDM comprises a set of techniques such as classification, multivariate statistics, association rule mining and web mining. These are highly exploratory techniques that can be employed for the prediction of learning improvement [23]. The methods can be utilised for modelling individual variances in students, and they offer an approach for responding appropriately to these variances, which will lead to an improvement in student learning [24]. Scheuer and McLaren [25] distinguish EDM through two points. First, its features precisely concentrate on educational data and issues, both practical (e.g., enhancing a learning tool) and theoretical (e.g., examining a learning hypothesis); second, EDM provides a methodological contribution by researching and improving DM methods for educational applications.

It is worth noting that EDM is similar to DM that it requires a strong and consistent data warehousing strategy to be successful. Guan et al. [26] discussed how it is essential for each institution to have meaningful information and good quality data available for future research and decision making. Without a warehouse it is difficult to obtain the information that the decision makers require efficiently and quickly. Some of the main reasons for starting warehouse projects are an increasingly competitive landscape, and increased accountabilities of reporting to exterior stakeholders such as community leaders, legislators, board members and parents [26].

EDM is concerned with areas such as mining module content, improving domain knowledge structure, and analysing educational processes such as module selection, alumni relations and admissions [20]. It is also concerned with the development of learning support systems such as course management and enrolment recommender systems. A recommendation systems in a learning context is a software agent that attempts to “intelligently” recommend activities to a learner based on the activities of previous learners [27] .

1.3 Motivation

Data routinely collected is often not used in decision making to the extent that it could be used. In the context of education, data can be used to improve student outcomes, offer better choices more suited to the students, identify points of failure, improve the competitiveness of institutions, etc.

However, this requires not only good-quality data and the effective application of DM technology, but also a process that ends with the successful deployment of any learned models. The deployment stage of any DM project is often one of the most challenging to materialise, yet without it the whole project may be deemed as irrelevant. Hence in this work we try to investigate the use of analytical DM methods in data mining and in the KDD process, from the initial stages of assessing the data quality, through building relevant models that encompass some of the knowledge embedded, through to performing an analysis of the challenges and barriers in the deployment phase.

Much work has been done on data mining educational data to predict student performance (e.g.[28, 29, 30], and others), as nowadays the achievement of good outcomes in undergraduate degrees has become very important in the context of Higher Education (HE), both for students and for the institutions that host them. However, important methodological issues remain, for example how to develop accurate predictions, especially in the context of large amounts of missing data.

Personal recommender systems (PRS) can be one tool to improve outcomes for students by directing them to the choices they may be more successful at. They are considered as advisable automated solutions for assisting students to make better choices [31] leading to better outcomes. Recommender systems must be modified and adjusted to be employed in an educational context, which is different from a commercial environment. Recommender systems are therefore highly domain dependent [32, 33]. In the context of module selection advice, a recommender system may be based on projected student performance, which requires good predictions of performance. We focus particularly on this aspect.

Another aspect of EDM that has received little attention is how stakeholders react to the utilisation of the extracted knowledge and this is important as we aim to provide systems that stakeholders will accept. To close this loop and provide deployment advice, we investigate acceptance of the utilisation of knowledge extracted from a management perspective.

1.4 Research Aim and Objectives

The main aim of this research is to investigate how data collected routinely by universities can be used in the context of EDM to improve student experiences and outcomes. In particular, we will investigate and experiment with some of the analytical techniques that can be used to predict student outcomes in the context of Higher Education, both at the programme level and at the module level. For this we will focus on DM techniques. We will achieve this aim by addressing the following objectives:

1. We will extract and prepare student data for analysis and report on its quality (Chapter 3).
2. We will perform initial analysis to identify students who are at risk of obtaining poor outcomes using data mining methods (Chapter 5).
3. We will then investigate how to construct a robust predictive model for module performance, which could be deployed as part of a future enrolment support system. Our initial recommendations will be based on potential student performance on a module so for this we will present a comparison of different predictive models that could be used in the context of module outcome prediction. As part of this effort, we will investigate methodological issues that arise in educational data mining models (Chapter 6).
4. We will perform robust experiments by using a number of datasets with different characteristics (Chapter 6).
5. We will investigate qualitatively, by means of interviews, the views of students and staff on deploying the knowledge found, for example as part of an enrolment recommender system or as a programme of remedial action for students at risk of poor outcomes (Chapter 7).

1.5 Research Questions

Our main research question is: How can data routinely collected by a University be used in the context of educational data mining to enhance student outcomes and experiences? In order to answer this question we address the following associated questions:

1. How can we, using DM techniques, develop an effective method for building competitive predictive models for students overall outcomes from regularly collected data, and how to highlight features associated with poor performance?
2. How can we develop a novel method for constructing predictive models for module outcomes?
3. How can we design a management-focused study that investigates the views of both the students and the institutions on how to utilise any knowledge derived from the answers of the previous research questions to improve students performance and implement a future enrolment system?

1.6 Research Limitations and Boundaries

The limitations and boundaries of our study are described below:

- One difficulty is finding good-quality data on anything other than performance. For example, our models could have used data on students' engagement, attendance and employability, but our investigation concluded that no good quality data was available to measure those in our specific setting. One possible outcome of the research will be a recommendation that data on those aspects should be improved to enable more factors to be considered in future studies.
- Although ethical approval was obtained, ethical considerations when handling personal data limit our ability to extend the research beyond the initial remit. We are also constrained to using student data associated with one UK university, as such data from other universities is not readily available. Therefore, our conclusions may not be as generalisable as we would wish. However, as we answer some methodological questions, and add for that some publicly available datasets, those should be generalisable to other settings.
- It is worth mentioning that due to the time constraints of our degree and the complexity of the ethical considerations, our research concentrated on the British education systems. A further study that addresses different educational systems would be advantageous.

1.7 Research Novelty and Contribution

The contributions we expect from our work include:

1. A method for constructing competitive predictive models for student outcomes from routinely collected data, highlighting features associated with poor performance. A conference paper summarising the first part (Chapter 5) of this work has been published [1].
2. A novel method for building predictive models for module outcomes, innovative for its use of multiple imputation combined with an ensemble to handle missing data. A journal paper including this work (Chapter 6) is under consideration by the Journal of Educational Data Mining.
3. A management-focused study of how to utilise any knowledge derived from the exercise in the educational context both from the point of view of the students receiving help and the institutions implementing a future enrolment system. We are working at the moment on producing a journal paper on this work (Chapter 7).

1.8 Preliminary Thesis Outline

- **Chapter 1: Introduction** This chapter discusses the background and the importance of the research. It includes the motivation to conduct the study. It also, summarises the aims and objectives of the research as well as its limitations and boundaries. It explains the research contributions and lastly introduces this thesis outline.
- **Chapter 2: Literature Review** This chapter provides a review of the literature on EDM which includes EDM definitions, objectives, the methods used, the analysed data, the process of applying EDM, Recommender System in EDM, the technological tools used in EDM, and EDM SWOT (Strengths, Weaknesses, Opportunities, Threats). Finally, since in the last few years the number of studies on EDM has grown noticeably in the literature, this chapter will detail some of EDM applications and related works.
- **Chapter 3: Data Description** In this chapter, we explain in detail the three types of data utilised in our research.
- **Chapter 4: Research Methodology** In this chapter, we present how our result chapters are connected and explain our research design. We explain the research methods, including prediction models and how to evaluate them. We explain how we handle missing data. We also included the methodology for the management study. Lastly, we include the ethical considerations of our research.
- **Chapter 5: Predicting the outcomes of Students at risk** In this chapter, we frame the more general problem of performance prediction and apply data mining models to identify groups of students who may be at risk of poor outcomes so that targeted interventions can be proposed to improve their outcomes. We compare results across two schools of study.
- **Chapter 6: Generating module-level performance predictions** This chapter shows the more granular problem of performance prediction through conducting a comparison of module-level predictive models. It also proposes a novel multiple imputation method combined with an ensemble for dealing with missing data which is shown to improve the predictive models.
- **Chapter 7: From data to decisions - a management perspective** This chapter presents a management study of how to use the knowledge derived from the performance predictions experiments in Chapter 5 and Chapter 6 from the perspective of both the students and the institution. This includes an investigation of the acceptability of a future enrolment system based on the results of Chapter 6 and an investigation of how to utilise the derived knowledge from performance prediction in Chapter 5.

- **Chapter 8: Conclusion and Future Research** This chapter provides the conclusions drawn from discussing our results. It will also include recommendations for future studies.

Chapter 2

Literature Review

2.1 Introduction

As mentioned previously in Chapter 1, Data-mining (DM) is recognised for its powerful role in revealing hidden information from massive amounts of data, and it is also referred to as knowledge discovery in databases (KDD) [34]. Its application has delivered benefits in various fields, including bioinformatics, e-commerce, and more recently, educational research, where it is known as Educational Data Mining (EDM) [35]. Educational Data Mining is a novel DM application on raw data from educational systems for the purpose of solving educational problems and answering educational questions [36, 37].

In the last few years, the demand for work in this area has greatly increased the number of research studies [37]. EDM is now an established field and, as such, a number of reviews have been published (e.g. [38, 30, 39, 40, 41, 20]). In particular, Peña-Ayala [30] covers 240 of EDM works. We review some of that work and use it to apply best practices to our own problem. Moreover, it is important to make EDM accessible enough for instructors to perform advanced analytics on data that is relevant to them, for example in the context of online Course Management Systems(CMS). However, one of the shortcoming of the existing research is that outcomes are not always generalisable to other Higher Education institutions. This indicates that the outcomes are highly related to a particular institution at a particular time. Research in EDM should investigate approaches that are more generalisable.

In 2008, EDM reached a high point by becoming an independent research area with the establishment of the Journal of Educational Data Mining and the annual International Conference on Educational Data Mining [42].

The goal of this chapter is to review the literature on EDM which includes EDM definitions and disciplines, EDM objectives, the methods utilised, the analysed data (including problems with missing data), issues about personalisation, decision making, etc, and also to review EDM's SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis. Finally, we cover application of all these methods discussed to problems similar to ours.

2.2 EDM Definitions and Disciplines

The term 'Educational Data Mining' has a number of definitions. We will mention the differences among them. According to the Educational Data Mining Society's website, Baker et al. [20] defined EDM as *"concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in"*. As pointed out by Huebner [21], this definition does not mention data mining specifically, so as a result it allows researchers to develop and explore alternative analytical methods (e.g. ad-hoc reporting, statistical analysis,etc.) that can be implemented on educational data. In contrast, Romero and Ventura [2, p.1] consider data mining in their EDM definition: *"the application of data mining (DM) techniques to specific types of data sets that come from educational environments to address important educational questions"*. Despite the minor difference, both definitions put emphasis on improving educational systems through discovering knowledge based on educationally related data.

EDM can also stand for Education Data Movement [43] at the level of a normative argument about what should be done, whereas Educational Data Mining, focuses more on the technical aspects of how it is done. To distinguish more clearly between these two terms, EDM (movement) presents several themes from the wider 'big data' movement. It also shows a concern with the construction of models and the different levels of complexity, based on large volumes of available data, to make predictions about future outcomes at an individual or collective level. Indeed, an alternative, if rather more restrictive term, that is often applied is 'predictive analytics' [44]. The models may be utilised to evaluate students, courses, curricula, modules or, more controversially, individual instructors. A key application has been the attempt to measure educational 'Value Added,' or to identify 'learning gain.' The EDM (movement) often emphasises that Higher Education has been a fairly 'late adopter' of predictive analytics as a management instrument [44].

Yet another term often used is Learning Analytics (LA), which according to the Learning Analytics and Knowledge website [45] is *"the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimising learning and the environments in which it occurs"*. Both EDM and LA focus on how to exploit "big data" to

enhance education [46]. Despite the fact there is no solid distinction between these two areas of study, they are associated with different research histories. Also, they are currently growing as separate research areas [47]. Siemens and Baker [46] have discussed some differences between these two communities. For example, EDM is concerned with automated methods whereas LA focuses more on human-led methods for analysing educational data. Additionally, EDM studies give more attention to analysing individual components and the connections between them, whereas LA studies emphasise a more holistic approach, aiming to understand the systems as a whole, including their full complexity. More details and comparisons can be found in Romero and Ventura [2], Bienkowski et al.[47], and in Siemens and Baker [46].

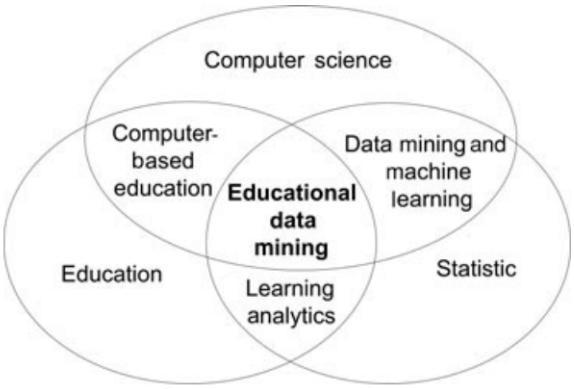


Figure 2.1: Main areas related to EDM. Adapted from [2]

Next, EDM can be considered as the combination of three main fields (see Figure 2.1): Education, Statistics and Computer Science. The intersection of those three fields forms other subfields closely related to EDM such as DM and Machine Learning, Learning Analytics, and Computer-based Education [2].

EDM as an interdisciplinary area applies methods and techniques from, but not limited to, information retrieval, data mining, recommender system, machine learning, cognitive psychology, psycho-pedagogy, statistics, etc. The determination of which technique or methods should be applied depends on the educational concern being addressed [2].

2.3 Objectives of EDM

In the past few years, EDM has been applied to achieve many aims that are all part of the universal objective of improving learning [2]. A list of these aims has been provided in some studies (e.g. Baker et al. [42], Calders et al. [23], Romero et al. [2], Bienkowski et al. [47], Scheuer et al. [25]).

Romero and Ventura [48] presented EDM objectives by classifying them according to the standpoint of the end user (i.e. students, educators, researchers, academics/administrators):

- **Students:** The objective is to propose to students resources, activities and learning tasks that would make their learning performance better, to react positively towards students' needs, to give good feedback, new learning experiences, a simple path to follow, etc. These recommendations usually depend on tasks and activities that have been done by similar students or have been done previously by the same learner.
- **Educators:** Their objective is to develop the teaching methods and the structures of the course content, to assess the efficiency of the course content by understanding the learning processes, to categorise students into groups based on their need for monitoring and guidance, to discover learning patterns of irregular as well as regular learners, to become aware of the most repeatedly made mistakes, to understand behavioural, cognitive and social aspects, etc.
- **Researchers:** Their objective is to compare and improve DM techniques. This will make them qualified to recommend the most efficient one for solving a particular educational problem or for performing a particular educational task, etc.
- **Academics/administrators:** Their objective is to assess the best practice in order to better coordinate the resources (material and human) of the Higher Education institutions, to improve the educational programmes offered, and to regulate the efficacy of the new methods and technology related to mediated instruction and a distance learning approach.

It is sometimes difficult to categorise objectives based on these four actors, especially as some of the objectives are linked to more than one actor. Therefore, another categorisation is based on area of application, according to a number of studies (Baker and RSJD[42], Bienkowski et al.[47], Scheuer et al.[25], Romero et al.[49]). This could be constructed as follows:

- **Learner modelling**. There are several applications of student modelling in the educational context. For instance, the real time identification of student characteristics such as learning progress, satisfaction, motivation, experiences, skills, knowledge, learning styles, meta-cognition, and precise problems that negatively affect a student's learning outcomes (e.g. performing many errors, gaming the system, misuse or inefficient use of the available help or learning resources). The main objective in this category is to utilise usage data to build a student model. The techniques utilised for this type of objective are not limited to classification, clustering, and association analysis, but also include psychometric models, statistical analyses, Bayesian networks (containing Bayesian Knowledge-Tracing) and reinforcement learning.

- **Predicting student learning outcomes and performance**. The objective is to predict any type of learning outcomes such as final grades, retention on a degree course or future aptitude to learn. These predictions are usually based on data from module activities. The most utilised techniques for this type of objective are association, clustering and classification.

- **Generating recommendations**. The objective is to suggest to learners which tasks (or links or content) are the most appropriate for them at a given time. The techniques frequently utilised for this type of goal are sequencing, association rules, clustering, and classification.

- **Communicating to stakeholders**. The objective is to assist the course instructors for example to develop alerting/reporting tools for student engagement. The most frequently used techniques are process mining and data analysis through visualisations, reports, and statistical analysis.

- **Domain structure analysis**. The objective is to identify the domain structure and enhance the domain models through utilising the capability to predict students' performance as a quality assessment of the domain structure model. The performance in exams or within an educational environment is used for this objective. The most commonly used techniques for this type of objective are space-searching algorithms, association rules and clustering.

- **Module improvement and maintenance**. The objective is to assist instructors and administrators in enhancing the modules (contents, activities, etc.) through utilising information regarding students' learning and usage. Clustering, classification, and association are the most commonly used techniques for this type of objective.

- **Studying pedagogical support**. This concerns the investigation of the impact of different types of pedagogical support that can be provided through automated learning tools. Finding the most efficient type of pedagogical support is one of the most interesting areas for EDM. The common used techniques are relationship mining and Bayesian Knowledge Tracing.

  Bousbia et al. [37] noted that the above EDM objectives try to enhance various features of education systems in general and the computer-based learning environments (CBLE) specifically. In this context, student modelling is the most essential point to achieve various aims and tasks (personalisation, adaption, tutoring, etc.). Therefore, the others objectives depends mainly on the first goal of 'learner modelling'.

2.4 The Methods

Most of the long-established DM techniques such as classification, association analysis, clustering, etc. have been successfully applied in the educational setting. However, the educational domain consists of distinctive characteristics that need special treatment such as data hierarchy and non-independence (discussed later in 2.5.4) [49, p 1-5]. For this reason, EDM researchers do not limit themselves to the utilisation of DM techniques. They also apply techniques drawn from other areas related to EDM such as Statistics, Data Modelling, Psychometrics, Web Log Analysis, etc.

Baker [42] proposed a grouping of these methods into clustering, prediction, distillation for human judgement, relationship mining and discovery with models. Then, both Romero and Ventura [2] and Bienkowski et al. [47] expanded on this. Here we introduce a grouping of these methods based on these studies (i.e. [42], [2],[47], [20],[46]):

2.4.1 Prediction

The goal is to infer a target attribute or single characteristic of the data (predicted variable) from other explanatory attributes of the data (predictor variables). Mainly, there are three types of prediction: classification, regression, and density estimation. In classification, the predicted variable has a categorical value. In regression, the predicted variable has a continuous value. Lastly, in density estimation, the predicted variable's value is a probability density function. It is used, for example, to predict student academic performance [50] and behaviour [51].

2.4.2 Clustering

The goal is to divide data into groups with similar characteristics. Clustering is useful when the categorical labels are unknown in advance. It uses a variety of distance measures to determine how similar each data point is to other data points. If a set of clusters has been established, a new data point can be assigned to it by calculating the closest cluster. An example of an EDM application is clustering students based on their interactions or learning patterns, or clustering similar course materials [52].

2.4.3 Relationship mining

This is used to find the relationships between variables in a data set that contains a large number of variables, then to convert these relationships into rules that will be useful later.

There are several types of relationships such as association rule mining (discovering any relationship between variables), causal data mining (finding causal relationships between variables), correlation mining (finding a positive or negative linear correlation between variables), and sequential pattern mining (finding a temporal association between variables). It is used for example to discover students' learning mistakes or difficulties that often happen simultaneously [53].

2.4.4 Distillation of data for human judgement

This places an emphasis on describing the data or patterns in intelligible ways. This methodology uses visualisation, summarisation, and interactive interfaces to support decision making and to underline useful information. It has been utilised to assist instructors with analysing and visualising the students' module activities and usage information [54].

2.4.5 Discovery with model

The goal of this method is to utilise an existing validated model (e.g. developed using clustering) as a component in different analysis (e.g. prediction). This method is beneficial in the educational context such as in the discovery of relationships between learner's behaviour and his/her characteristics [2].

2.4.6 Outlier detection

The goal is to identify data points that are distinctive from the remaining data. An outlier is a measurement (or observation) that does not fit well with the other values in the data set. In EDM, this method can be applied to detect students with learning disability, deviations in the instructor or the student actions or behaviour, and to detect unusual learning processes [55].

2.4.7 Social network analysis (SNA)

The goal of SNA or structural analysis is to understand, examine, and measure the relationships between individuals within a network. It assesses social relationships using network theory composed of nodes (that represent individual entities within the network) and links or connections (that represent relationships between the entities, such as family connections, friendship, etc.). In EDM this method can be used to analyse and interpret the structure

and associations in collaborative functions and interactions with applications or websites [56].

2.4.8 Process mining

The goal of this method is to derive knowledge related to a process from event logs that are automatically documented by an information system to have a visual representation of the entire process. This method consists of three events: model discovery, conformance checking, and model extension. In EDM, process mining can be applied to student behaviour with regards to their traces consisting of a series of modules, timestamp and grades triplets for each student [57].

2.4.9 Text mining

The goal is to extract information (such as rules, patterns, models, trends, and direction) from unstructured text. The main tasks of text mining are concept/entity extraction, text categorisation, text clustering, sentiment analysis, production of granular taxonomies, entity relationship modelling, and document summarisation. Text mining is also known as text analytics or text data mining. In EDM for example, text mining has been applied to analyse and investigate the content of emails, forums, boards, Web pages, documents, chats, etc. [58].

2.4.10 Knowledge Tracing

The goal of Knowledge Tracing (KT) is to estimate a learner's level of knowledge and skills attainment. These skills have been used in cognitive tutor systems [59]. These are computer programs that imitate human tutors by offering individualised instruction to learners [60]. KT utilises, as proof of student knowledge on a precise skill, both logs of students' accurate and inaccurate answers and a cognitive model that associates each problem-solving item with the skills needed. This method traces students' knowledge throughout time and it is parameterised by variables. There is a corresponding formulation of Knowledge Tracing as a Bayesian network.

The selection of an appropriate method is determined by the nature of the learning system, the research goals and the type of available data discussed next.

2.5 The Analysed Data

The type of data analysed in EDM research has characteristics that will help us to differentiate it [37]. They are described below .

2.5.1 Origin of the data

Data used in the EDM environment originates from a variety of sources:

- A massive amount of available data has been stored over the years in the log files of educational software or in the educational institutions' databases.
- Some specific data is produced through experiments within a research study.
- There is also publicly available data obtainable by researchers through benchmark repositories, such as the PSLC DataShop[1], which is a well-known public data repository opened by the Pittsburgh Science of Learning Centre [20].
- Data collected from existing online courses that are utilised by a big number of students worldwide, such as WebCAT[2] and Moodle [3] [20].

Some of these sources may contain private data that belongs to a specific educational institution. This type of data cannot be obtained by all researchers and there may be specific policies and procedures to access the data [61]. By contrast, data from the last two sources is considered public hence there are no restrictions on its usage for analysis and validation. Public data is beneficial in allowing researchers to learn from past experiences, establish comparisons and perform more robust research. This in turn will lead to a science of education that is better validated, progressive, and concrete [20].

2.5.2 Mode of collection

There are two main modes of collection:

- Digital: based on the utilisation of software that stores student activities. The results of this could be information recorded in databases, video or audio recordings, or numerical traces that might be in a log file.
- Manual: carried out by a human observer taking notes on the learning circumstances to assess the participants' undertakings and accomplishments.

[1] `https://pslcdatashop.web.cmu.edu/`
[2] `https://web-cat.org/`
[3] `https://moodle.org/`

2.5.3 Learning environment

Currently, there are several educational environments, both in computer-based and traditional education as shown in Figure 2.2. Each type offers different sources of data that require pre-processing taking into consideration both the nature of the obtainable data and the particular tasks to be resolved by the DM techniques.

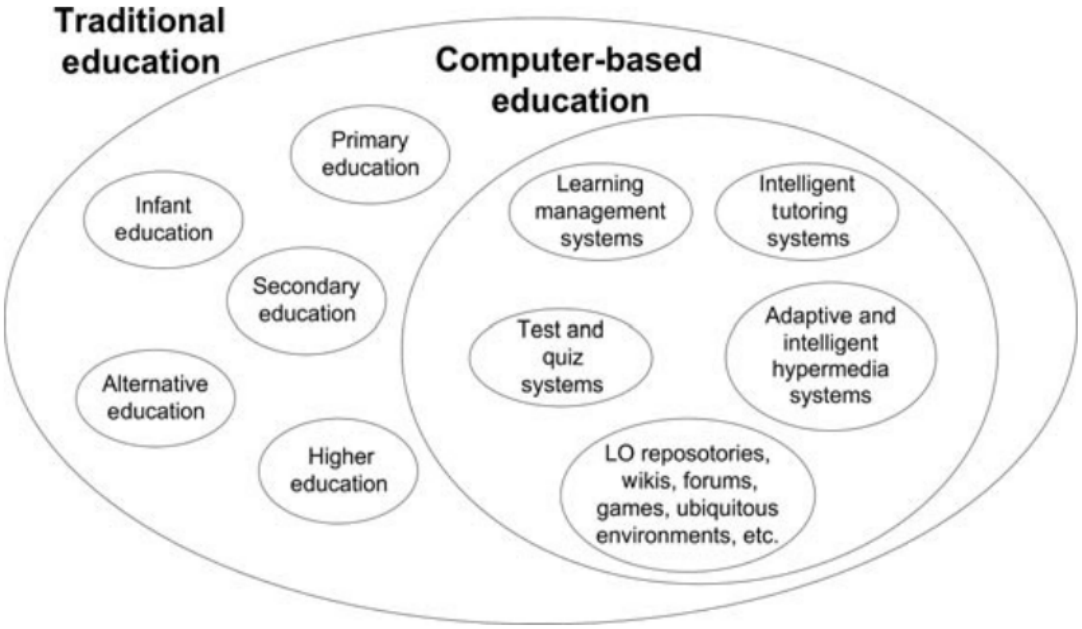


Figure 2.2: Types of traditional and CBE environments and systems. Adapted from [2]

- **Traditional education**. As seen in schools, thus very popular and long-established. It encompasses, for example, infant, primary, secondary, higher, and alternative education. These environments depend mostly on face-to-face communication between instructors and learners structured around class discussion, lectures, individual work, small groups, etc. These systems collect information on learner marks, attendance, personalised plans and curriculum aims. In addition, educational institutions may establish databases for managerial data such as information on the learners, the educators, schedules, etc.). In orthodox classrooms, instructors usually improve teaching by keeping track of learners' activities and analysing their performance through documents and observation [2].
- **Computer-based education (CBE) systems**. This refers to the use of computers in education to offer guidance, to recommend or to manage guidelines given to the learner. Thus, this type of collected data is usually digital, and its size is usually

less than that from traditional education, which consists of enormous databases. At first, CBE systems were operated independently on a local machine without utilising artificial intelligence methods for student personalisation, modelling, etc. The universal use of the internet has created new web-based educational systems, for example e-training systems, e-learning systems, online instruction systems, etc. However, the growing use of artificial intelligence methods has led to the appearance of new adaptive and intelligent educational systems. Some of the current and popular types of CBE systems are Computer Supported Collaborative Learning (CSCL), Learning and Management Systems (LMS), Adaptive and Intelligent Hypermedia System (AIHS), Intelligent Tutoring System (ITS), test and quiz systems, serious games, etc. [2].

2.5.4 The described level

It has become important to take into consideration the hierarchy and non-independence of educational data, as each student contributes enormous amounts of data while proceeding through their course of study, and those students are influenced by their classmates, instructors and other course level effects. Educational data has various levels of meaningful hierarchy, such as the answer level, the keystroke level, the session level, the classroom level, the student level, the school level, and the instructor level. Each level of granularity provides different types of data; thus it is essential in EDM to select the accurate level of granularity in order to only recognise the attributes that can be logged at that particular level of granularity [42, 39, 62]. Utilising and benefiting from these multiple levels of meaningful structure in educational data has often made the methods of EDM different to the methods of the broader Data Mining literature [42].

The non-independence of the data comes into play, for example, when we gather data from education discussions and need to categorise whether the discussion's input are off-topic or on-topic. We have to consider that inputs are not statistically independent of each other because several inputs originate from the same student or discussion [25].

2.5.5 Types of data:

Variables collected may be of different types including:

- Administrative, personal and/or demographic data (gender, age, etc.).
- Exams marks and/or answers to questions.
- Responses to psychological questionnaires for evaluating user skills, cognitive characteristics, motivation, satisfaction, etc.

- User interactions with the educational system: from low-level actions such as mouse clicks, to high-level ones which include the browsing pattern, number of attempts, etc.
- Facial and visual interactions.
- Social interaction (forum participation, chat, instant messages, etc.).

Bousbia et al. [37] noticed that the type of data recorded will vary to a large degree, depending on the type of learning environment. However, there are some studies that have incorporated different types of data to give a comprehensive representation of student performance and behaviour. For example, Romero et al. [63] have attempted to predict student success in the final test depending on the level of their participation in online forums (social interaction), assignments, quizzes and demographic data. The incorporation of these dissimilar kinds of data requires a number of steps in the implementation of the EDM process.

2.6 Process of Applying EDM

Romero and Ventura [2] and Romero et al. [49] explained that the process of applying DM to educational environments can be interpreted from two different perspectives.

The first perspective is from an experimental and an educational standpoint; it can be identified as a repetitive cycle of hypothesis development, testing, and modification as represented in Figure 2.3. The aim of this process, in addition to transforming the data into knowledge, is to utilise the resulting knowledge to enhance the learner's experience. This is a kind of formative assessment of a learning program during its improvement process, and with the purpose of continually enhancing the program.

The second perspective is from a DM standpoint, it can be observed almost identical to the general KDD (Knowledge Discovery and Data Mining) process (2.3), albeit there are particular features or essential differences in each phase, as illustrated in the next subsections [64].

2.6.1 Educational environment

The type of educational environment (such as computer-based, or traditional classroom) and its supportive information system (such as adaptive hypermedia, intelligent tutoring or a learning management system) cause different types of data to be gathered to resolve several educational issues. All these data are associated with difference sources (such as motivational questionnaires, field observations, administrative data, final marks etc.). Integrating and

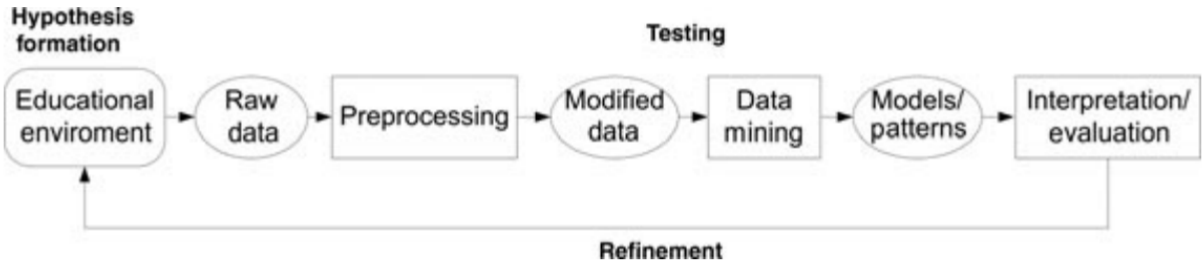


Figure 2.3: Process of Applying EDM. Adapted from [2].

collecting this raw data for mining is a significant task, therefore a pre-processing phase is needed [2].

2.6.2 Pre-processing

In the educational setting this phase can account for more than half of the total time spent on the DM process. First, the existing (original) data is not in an applicable form initially. Second, educational data have a hierarchical and heterogeneous nature (as we explained previously in 2.5.4) that make choosing data formats and structures for a particular event a crucial task. The optimal data structure is also determined by the type of educational problem. Therefore data require numerous transformations for solving a particular educational problem. In addition, determining the suitable granularity level for the data integration process is important. For example, data at different levels of granularity may be required (school level, classroom level, department level, session level, answer level, and keystroke level) (see Figure 2.4).

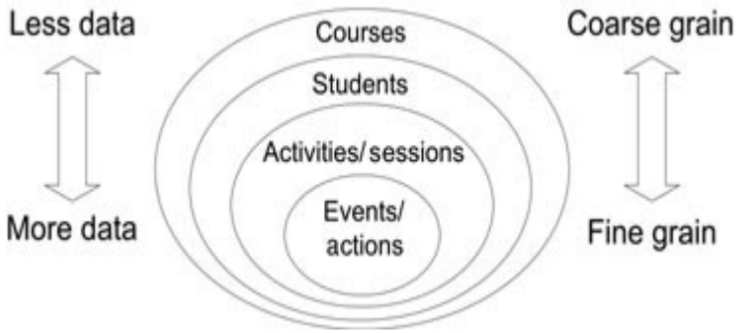


Figure 2.4: Different level of granularity and their Relationship to the amount of data. Adapted from [2].

Moreover, educational data may include missing and/or incorrect data. Each type of missing

data may require different treatment. The risk of bias as a result of missing data analyses is related to the mechanisms that generated the missing data. There are three mechanisms described in the literature [65]:

- Missing Completely At Random (MCAR), which means the missing values of a specific attribute X are not related to other attributes in the data sets, in addition to the underlying values of X itself. For example, consider a participant has missing BMI in a their doctor's records because traffic delayed him/her and he missed his appointment with the nurse that would take the measurement. Alternatively, consider that a patient has a missing blood test because it was accidentally damaged in the lab.
- Missing At Random (MAR), which means that missing values of an attribute X could be related to other observed variables, yet still must not be related to the underlying values of X itself. For example, consider that one gender is less likely to disclose their weight in their medical records so the probability of the weight being missing depends on the gender but not on the weight itself. A second example of MAR, consider that a school district applies a math aptitude test, and students that score above a certain cut-off join in an advanced math module. The math module marks are MAR because absence is completely determined by scores on the aptitude exam (such as, students that score below the cut-off do not have a mark for the advanced math module).
- Missing Not At Random (MNAR), where the probability of missing values of an attribute X is related to the underlying values of X. MCAR is the safest scenario whereas in MAR and particularly MNAR missing values may introduce biases. For instance, consider that obese or heavy people are less likely to disclose their weight. Then weight is MNAR as the probability of missing depends on the value of the weight itself. Another example of MNAR, suppose we are investigating mental health and individuals who have been diagnosed as depressed are less likely than others to reveal their mental status, the data are MNAR. Obviously the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have got with complete data. In the same way happens when individuals with low income are less likely to record their income on a data collection form.

We will discuss how to deal with missing data, particularly in our context, in Chapter 4.

Some researchers have simply used only complete records. However, it may not be feasible to find complete real data for a large number of students in the higher educational context due to the frequent change of the modules and/or courses; or there may be students who complete their first year in one institution, and then transfer to a different second institution, meaning that this institution may not have complete records of their first year, etc.

Often, there are a large number of attributes associated with each student that can be condensed into a summary table for effective analysis. In addition, continuous fields are usually discretised into categorical fields to enhance their clarity and make them self-explanatory. Issues of order, and context are also essential during the study of educational data. Order is important in determining how tutoring and practice should be arranged, presenting a sequence of materials. On the other hand, context is essential in discussing whether any resulting model will work. Lastly, it is very important to protect and maintain the privacy of student records when collecting and integrating data by implementing a data management plan, such as the one we have in place for our thesis (see Appendix E). This may involve the removal of sensitive information and other forms of anonymisation [2].

It should be noted that due to the complexity of this phase some studies attempt to eliminate this step, for example Kruger et al. [66] offer a data model to structure the data stored by a LMS. They have also implemented a tool that performs the actual export/structure functionality for the Moodle LMS. However, these studies may fail to take into full account the quality of their data. Huebner in [67] and Brown and Kros in [68] explained that in the educational context good DM models depend on the quality of the core data.

2.6.3 Data Mining

In this phase, suitable DM techniques are applied. Many traditional DM techniques have been applied successfully in the educational field. Also, more specific methods for longitudinal and hierarchical data may have to be utilised in EDM. Some discussion of suitable methods has already been covered in section 2.4.

2.6.4 Interpretation of results

The final phase is essential to enhance the educational domain. The interpretability of the model may be an important consideration for this phase. For instance, decision trees may be preferred over neural network models because they are more comprehensible, even if they prove less accurate. Also, visualisation techniques may enhance interpretability, For instance, it is more effective to present only a part of the resulted association rules in a graphic form, rather than to present hundreds/thousands of association rules in a text format. Lastly, recommender systems may provide an avenue to present decisions to a non-data mining expert audience such as students or teachers [2].

2.7 Recommender System in EDM

In the educational discipline, a recommender system is an agent that recommends, in an intelligent way, actions to learners based on preceding decisions of other learners with similar demographics, academic or personal characteristics [27], individual's activities, the next problem or task to be done, links to visits (e-learning), and so on. The system should also be capable of adjusting contents, sequences, and interfaces to each individual student [39].

Schafer [69] explained that DM algorithms are, and will remain, a very crucial part of the recommendation process, because the have helped a number of promising applications to improve the accuracy of their recommendations. Moreover, using DM has also improved the type of recommender systems available. For example systems can take into consideration changes over time and offer a suggestion about when the user should use an item or when the recommendation should be made. Those are different to the traditional recommender systems, which were built using collaborative filtering [70, 71] and content based methods [72]. Traditional recommender systems are focused only on which item the user should consume, for example Netflix (recommends movies and TV-shows), YouTube (recommends videos), Amazon (recommends items), etc. [73].

There are many uses for recommender systems in education such as recommending the most suitable future e-links that learner should visit, learning materials in e-learning system, applicable discussions to the students, etc. [39]. However, to the best of our knowledge the usage in module selection has been limited.

2.8 Some Technological Tools Used in EDM

Currently, there are various commercial and free tools for EDM that help users to engage in DM on a smaller scale. These applications are not specifically designed for educational and/or pedagogical domains, for instance R[4], Weka [5], SPSS Modeller[6], MatLab [7], etc. However, educators may find these types of tools complicated to use.

In the current decade, a growing number of DM tools have been developed that focus on solving various educational issues. Romero and Ventura [2] mentioned some of the best tools. Nevertheless, Bousbia and Belamri [37] have analysed these tools and found that they are often designed for CBEs. They also have found that some of the tools, aside

[4]https://www.r-project.org/
[5]http://www.cs.waikato.ac.nz/ml/weka/
[6]http://www-01.ibm.com/software/analytics/spss/products/modeler/
[7]http://uk.mathworks.com/products/matlab/

from the benchmark repositories (e.g. Datashop), have never been re-used in the EDM environment.

2.9 EDM (SWOT)

Huebner in [67] and Papamitsiou and Economides in [74] have analysed articles on real case studies from educational domains and discussed EDM strengths, weaknesses, opportunities and threats. Some of those are summarised in Tables 2.1 and 2.2.

Table 2.1: Summary of EDM strengths and weaknesses.

Strengths	Weaknesses
– The large improvement in the accuracy of experimental outcomes. – The availability of different and comprehensible visualisations tools that can support students/instructors. – The discovery of different and very important patterns of learning. – The increase in the awareness of different learning behaviours and strategies.	– The misinterpretation of the results because of the human judgement aspects and a focus on reporting rather than on taking decisions. – Most of the statistical significant results are based on quantitative research methods, because the qualitative methods have not yet shown statistical significant outcomes. – The overload of information may cause over-complexity of the systems. – Up to now, only expert instructors and researchers can understand and describe the outcomes correctly, which may lead to human resource limitations for some educational institutions that wish to implement DM [67].

[8]http://www.cs.waikato.ac.nz/ml/weka/datasets.html

Table 2.2: Summary of EDM opportunities and threats.

Opportunities	Threats
– The ability to use Open Linked Data for data compatibility and standardisation among many applications and tools, leading to the development of a generalised platform. – The ability to deliver multimodal learning that helps achieve effective learning opportunities based on complex metrics. – Self-learning and self-awareness of autonomous, intelligent and large systems. – The ability to reach technology acceptance by the users. – The ability to modify the open source DM software to meet the users needs [67]. – The availability of the help documentation that is associated with open source DM software, which eases the user learning process. There are also sample data sets that come with these software packages [8], which will help the users to learn the software before applying DM to their own data. – Online forums, discussion boards, and FAQs relating to these software packages give users the ability to discuss their problems[67].	– Ethical issues, such as issues related to private or sensitive data about students. – Over-analysis: the deeper the analysis become, the less the possible generalisation of the results. – Patterns can be misclassifed. – There are conflicting findings during the implementations process which may as a result affect the trust of these findings.

2.10 Student Decision Making, Choice and Data

We will discuss briefly the literature on several themes related to student data, student decision making, and module choice, since they are essential to understanding our study.

2.10.1 Student data and decision making in the UK

In the last few decades, the UK Higher Education system has shifted towards a more market-based approach with undergraduate students shouldering the majority of the costs of their tuition. The relationship between universities and students has come to be increasingly governed by market rules and by market regulators. Governance by the Higher Education Funding Councils has been increasingly supplemented by governance by the Competition and Markets Authority (see e.g., [75]). However, the lack of readily available and comparable data to support the choices made by prospective students has long been an aspect of this market (see e.g., [76], [77], and [78]).

There have been great government efforts to make information available, so comparisons can be made, at the institutional and course level, for example through the "Key Information Sets" provided under the UNISTATS brand in the UK (see e.g.,[79], [80]). This standardised data is collected and presented with the intention of aiding students to make systematic comparison of institutions and courses. There is a noteworthy body of literature on the processes and information that prospective students (and their schools and families) use in choosing institutions (see [81] for a review and [82] for criteria for evaluating league tables). There is also evidence that different groups of prospective students may choose differently [83]. There does not, however, appear to be much evidence that students are making use of these sources of 'cool,' rational information, preferring instead to rely on 'hot' information from their direct social networks, supplemented by 'warm' information from visit days and other forms of direct and indirect contact [84].

In comparison with research on institutional and course level choice, there is much less research into the choice of modules or electives within courses. As Hedges et al. [85] point out, what research there is tends to focus on supply side issues:

> *The existing literature on student module choice whilst in tertiary study emphasises supply side issues, such as curricula design and enhanced learning opportunities, but rarely examines why students demand particular modules [85, p.52].*

This is unsurprising as, in many subjects and institutions, there is limited (if any) choice available to the student to shape their degree through electives, and where such choice exists it can be quite tightly circumscribed. Nevertheless, module choice is often presented as an important way for students to shape part of their degree programme to their interests and aspirations. Because the maximum number of options are often offered in the final years of study, the chosen modules often colour students' evaluations of the course as whole.

2.10.2 Personalisation, decision making and choice

While much of the information that is presented to support the choice of university or the choice of degree programme is generic, predictive analytics promises the personalisation of information to support student decisions, which some have seen as a step to a wider 'personalisation of Higher Education' (e.g., [86]). For example, rather than informing the student that, on average, students were satisfied with a particular course or module, predictive analytics provides the opportunity to say that students *with the same or similar characteristics*, however defined, were satisfied with the course. Equally, rather than being told what the average mark for a module was, they can be given information about what the predicted mark would be for a student who more or less accurately matched their specific characteristics, such as gender or the pattern of marks in previous modules. Such personalisation might be familiar from e-commerce contexts ('recommended for you' 'people who bought/liked this product also bought/liked the following products').

We next review the theory of decision making and choice, as our study in Chapter 7 is concerned with examining how data can be analysed and presented to inform the student module choice process. Decision making is a vital part of human activity, which involves how people choose a suitable choice or set of intended actions from among various alternatives, according to given criteria or strategies [87] and [88]. Decision making is considered one of the 37 fundamental cognitive processes modelled in the layered reference model of the brain (LRMB) [87] and [89]. Decision making is a study topic that draws on different disciplines from computer science, psychology, management science and economics to political science, cognitive neurology and philosophy [90]. Each of those fields of study has highlighted a different aspect of decision making. It is known that there is a demand to find a rigorous and axiomatic model of the cognitive decision-making process in the brain, which may act as the foundation of varied decision-making theories in the literature [90]. However, some of the most fruitful work has come from the intersection of economics and psychology. This work has focused on how individuals make decisions within a given 'choice architecture' and the range of 'biases' that affect them —in effect how the context in which choice takes place shapes that choice [91, 92] and [93].

A choice architect is the individual or institution that is accountable for managing and coordinating the context in which individuals make decisions [91]. Thaler et al. in [91, p.430] state, *"If you indirectly influence the choices other people make, you have earned the title [of a choice architect]"*. A few examples of choice architects are parents describing the available educational choices to their 16-year-old child, doctors explaining the available treatments to patients, and in our context, the enrolment recommender system describing

the available module choices to the student. Thaler et al. have shown that the way choices are presented and the amount of informative details they contain can affect human decisions. Therefore, choice architects can have significant power to shape, or as Thaler has put it 'nudge,' choices. Obviously, choice architects will not always have the best interests of the influenced individuals at heart. For example, they may push a more profitable item to consumers, in the place of a more suitable option.

Thaler et al. [91] have also explained that a useful system of choice architecture is one that is capable of assisting individuals to enhance their ability to map from the available choices to what economists call the 'welfare choice.' A 'welfare' choice can be interpreted in various ways such as an unhidden preference, optimisation of life span, income, or another measure of happiness [94]. In short, a good choice architecture helps individuals to make choices that will make them better off. For example, the system would present the information about varied choices in a more transparent and comprehensible way. This can be done by converting numerical information, which many individuals find hard to interpret, into units or other representations that interpret more readily into actual use.

In the different disciplines described, there are many ways in which people, during their decision-making process, adopt various strategies in addition in considering the size and the complexity of the available choices [91]. However, we are not able to explore them all here. Rather, we will focus on three relevant approaches to decision making to see how they can give insight into how students make their module choice.

- The first choosing approach is called a reflective system (the conscious thought). This is what Kahneman has called System 2 thinking [93]. It is a calculative and self-conscious thought process by which students use reasoning and logic to support them in making their decisions. This is considered a typical economist approach [91]. Additionally, Browning et al. [95] called this approach rational choice theory. The reason that differentiates this theory from other types of choice theory is that it denies the presence of any form of action other than purely rational and calculative ones. This means that people must predict the outcomes of alternative possibilities of action and calculate which option will provide them with the greatest satisfaction [96, p.3].
- The second approach to choice is sometimes called the automatic system (also known as gut instinct). This is akin to what Kahneman has called System 1 thinking [93]. It is an intuitive and rapid process that is not related with what we usually consider thinking [91]. The enrolment recommender system can use the instinct or the default thinking of the students who use this approach, to nudge them in a better direction.
- The third choosing approach is called discursive practice. It is developed from 'under-

standings' within a discursive context of debate, instead of a psychological frame, and signifies that the attachment of an individual to a specific viewpoint is a process of the acceptance, whether deliberate or unreflective, of an articulation of that perspective [97]. A discursive practice can also be defined as speech that includes paraverbal and verbal patterns, which might uncover the meaning of actions and experiences that form structured organisations [98]. For example, this approach has been beneficial for medical physicians. In some critical cases, physicians need to discuss the treatment choices with other physicians, patients and/or the patient's family to decide on the most effective treatment for a certain patient [99]. This approach differs from the two previous approaches in that it sees preferences as emergent from the choice process, rather than being established prior to, and outside of, the choice process. In short, from this perspective choice is a process of discovery.

2.11 Applications of EDM in the literature

There are many applications of EDM (e.g. [100, 49]). Predicting student performance is the most popular and oldest task. Nevertheless, in the last few years a number of different and new educational issues have been addressed using EDM applications such as: evaluating learning materials to provide students with better learning guidance; establishing knowledgeable understanding of educational phenomena; identifying unusual problems and learning behaviour, and offering feedback based on the learning behaviours of students [58]. Romero and Ventura [2] have presented some examples of EDM application/tasks.

Here we survey some of the applications of DM techniques in the educational context. We have a special focus on module performance prediction, as it is an important part of our study. For each group of studies, we describe the solved issues/objective, the size and type of data, the main DM methods used, and the reported quality of results.

2.11.1 Predicting academic success

The first group of studies looks at the academic success of students in Higher Education. The objectives were to predict dropouts at the start of the studies [101, 102, 103], successful completion of the studies on time [104, 102], overall performance [105], or the requirement for remedial classes [106]. The used data sets were considered quite large (between 500 and 20,000 records, on average 7,200 records). The data was collected from several institutions, or from the entire university or, for a number of years. The number of obtainable fields was also large (between 40 and 375), and only the most significant were utilised. The data

not only consisted of demographic data and module marks, but also frequently included questionnaire data on student experiences, perceptions, and their financial situation.

All of these studies compared multiple classification techniques. Decision trees were the most popular, but also neural networks and Bayesian networks were common. The accomplished accuracy was 79% on average, which was considered a good outcome by the studies. In the bigger data sets (greater than 15,000 records), 93%-94% accuracy was obtained. It should be noted that the classification accuracy for these studies and the following studies is measured by classification rate (which is the proportions of accurately classified rows in the dataset [49]).

There are other studies on student performance prediction. For example, Norwawi et al. [107], Kabakchieva[108], and Sivakumar and Selvaraj [109] used several well-known DM techniques to present performance prediction studies, taking into account grade point average (GPA), cumulative GPA, degree classification, student marks or grades as dependent/predictive variables for predicting student performance in specific modules or subjects. Kotsiantis et al. [110] applied an ensemble to predict student performance in a few written assignments in a distance learning environment. Pardos et al. [111] and Baker et al. [112] applied a range of ensemble methods to track student knowledge within intelligent tutoring systems.

2.11.2 Predicting module outcomes

Here we examine a group of studies looking to classify the success of students in a given module. The objectives were to predict failing/passing a module [113, 114, 115, 116], the actual mark [117], or dropout [50, 118, 119]. In most studies, the module was a distance learning module, where dropout and failure are very important issues.

The data sets used were considered small (between 50 and 350 records, on average 200 records). This is because the collection was limited by the number of students who took the same module. Normally, the data involved just one set of students, however, if the module had remained the same without any alteration, it was feasible to collect data from a number of runs of the module.

The attributes that were taken into consideration were demographic data, questionnaire data, students' activity in a particular module, and exercise tasks. It was possible at the beginning for the number of attributes to be larger than 50, but they were then reduced to 3-10 prior to the learning of the model. A large selection of classification techniques were applied and compared in these studies. The most popular techniques were decisions trees, K-Nearest Neighbour classifiers, neural networks, Bayesian networks and regression-based

methods. Accuracy of the models obtained was considered good or adequate. The most significant factors that affected the accuracy of the classification were the number of class values used (the best was for binary problems) and at which stage the predictions were obtained (the best time was at the end of the module, when all the fields were accessible).

2.11.3 Succeeding in the next task

In this group of classification studies, the objective was to predict the success of the students in the next task, given his/her answers to the preceding tasks. This is an issue mainly in automated adaptive testing where, based on the student's present knowledge level, the next question will be selected. In [120, 121], and [122], only the correctness (content) of the student's response was predicted, whereas Liu [123] predicted the student's mark in the next task. The data sets used were considered small (between 40 and 360 records, on average 130 records). The data included the students' answers to the preceding tasks such as the accomplished mark and measured skill, and perhaps other attributes related to the students' activities within the learning system. All of the studies employed probabilistic classification techniques (Hidden Markov models or Bayesian networks).

2.11.4 Motivation, metacognitive skills, and habits

A group of studies was concerned with metacognitive skills and other aspects which have an impact on learning. The objectives were to predict level or motivation [124, 125], skill in using the learning system [126], "gaming" the system [127], cognitive style [128], or recommended intervention strategy [129]. The first five studies used real log data. The data sets were varied (between 30 and 950 records, on average 160 records). This is because some studies pooled together all the data that belonged to one student's activities, whereas other studies employed short sequences of sessions. The attributes that were taken into consideration were number of pages read, number of attempts for each task, navigation habits, and time devoted to different activities. The number of attributes employed to learn the models varied between 4 and 7 attributes, which is considered a small number. Hurley and Weibelzahl [129] simulated a large collection of artificial data. They used four attributes to illustrate the students' metacognitive skills: locus of control, goal orientation, perceived task difficulty, and self-efficacy. The notion was that at a later time these attributes may possibly be derived from log data. The most commonly used classification techniques were Bayesian networks, decision trees, regression-based techniques, and K-Nearest Neighbour classifiers. The classification accuracy was stated only in four studies and was in the range

of 88% - 98%. One reason for the high accuracy is that class values were usually determined by experts, using similar rules and fields to those the classifier used.

2.11.5 Applications of clustering

Additionally, examples of EDM using clustering techniques are given by Parack et al. [51]. They presented a study that applied k-means clustering and Apriori algorithms based on students' academic records such as attendance, test grades, term work grades and practice tests. The aim of this study is to simply group the students, discover the hidden patterns that are relevant to the students' learning style, detect abnormal student behaviour, and implement student profiling. Maull et al. [48] conducted a study that applied clustering algorithms to model and discover the online curriculum planning patterns for middle and high school teachers.

2.11.6 Summary of applications

The 24 reviewed studies in (section 2.11.1, section 2.11.2, section 2.11.3, section 2.11.4) described as a group at the beginning of each of the four main EDM sections, provide a decent overview of typical educational data and the most common classification methods applied. In most studies the class attributes were associated with a student, and there was only one record of data representing each student. The size of the data set was large in studies at university level, while it was small (varying between 50 and 350 rows) in the studies at module level. Larger data sets were accessible for some tasks, such as sequencing of log data which was classified individually. The original data consisted of both numerical and categorical attributes. Usually, the data was discretised before modelling, but occasionally both categorical and numeric versions of the data were modelled and compared. In some cases the data set contained only pure numerical data. This was when all the fields were task marks (such as test marks or assignment marks) or statistics on log data (frequencies of activities, time spent on activities). Nevertheless, the task marks usually had only a few values, and the data was discrete. This is an essential characteristic, because different classification techniques may be appropriate for continuous and/or discrete data. The most popular classification techniques were Bayesian networks (13 studies), decision trees (16), K-Nearest Neighbour classifiers (6), neural networks (6), different kinds of regression-based techniques (10), and support vector machines (SVMs) (3).

2.11.7 Performance prediction in the context of missing data

Vialardi et al. [36] employed a C4.5 decision tree to predict student academic performance, and in turn, to develop an enrolment system that would help the students to make the optimal decision regarding both the number and choices of modules on which they should enrol. They tackled this as a *Pass/Fail* classification problem, regardless of whether or not they were predicted to obtain a high grade in the module. The researchers used data from 100,000 enrolment records associated with only one school of study. The data included demographics, module of enrolment, module marks, the number of modules for each academic term, and the cumulative GPA of each academic term. There was more than one record associated with each student based on the number of modules; for example if a student took C modules, he/she would have C number of records in the database. In the evaluation phase, the system was able to accomplish 80% accuracy using the final year in the data set as the test data; the remaining data was used in the learning phase of the system. They evaluated their system using only a classification accuracy metric on real student data, although as we mentioned previously there is much more than accuracy to determine whether an item should be suggested. As only students that took a particular module were included in the predictive system for that module, their work was based on complete analysis.

On a related paper, Vialardi et al.[11] compared a number of data mining algorithms such as Naive Bayes, K-Nearest Neighbour, C4.5, Bagging and Boosting for module performance prediction, and used two attributes to improve the significance of the recommendations made: the difficulty of each module (taken as the average of previous student grades), and the level of a student's knowledge before taking the module (computed from previous obtained grades in related modules). They also employed other attributes such as demographic data, number of modules per academic term, grades obtained, enrolment on modules, average grade and the cumulative average grade per academic term. The data related to a period from its formation in 1991 to the first term of 2009. The results showed that Bagging was the best technique for accurate predictions, by predicting 85.36% of accuracy. This study also evaluate only the accuracy of their system using the classification accuracy metric. Again, the authors used only complete records from students that took the same module for the prediction.

Bydovska et al. [130] presented a study to recommend passable elective modules to students. They performed the prediction using data mining and social network analysis with real data from the Information System Faculty of Masaryk University. The data consisted of several attributes that related to student demographics, modules and course profile. They also

used attributes related to social behaviour data, such as email-communication, publication co-authoring, discussion forum messages, etc. The data was mined using Naive Bayes (NB), Support Vector Machine (SVM), Instant Based Learning (IBl), Classification Rules (PART), One Rule (OneR) and a decision tree (J48). They used several ensemble learning methods to improve their outcomes, such as the Vote, AdaBoost and Bagging techniques. SVM was the most accurate DM algorithm, and in some cases the results were improved using AdaBoost ensemble learning. In this study they showed that if enough social data is available, its use is significant and can influence the prediction model. The same authors published a previous study [131] to this one that showed that the social ties did not influence the model. They believe this was because the social data was incomplete (lack of data) at that time and there were hidden social relations that could not be discovered yet. An example of a social tie is given by students who have intelligent friends (who have higher marks) as they have a higher probability of passing the module than others. However, they found that when a module requires additional specific skills a friend's help will be less essential than a student him/herself mastering this skill. They also found that by increasing the difficulty and the specialisation of the modules the effect of the social ties will decrease. In terms of missing data, the study was based on students that took the investigated modules from 2010 to 2012. Each student was represented by one row, regardless of his or her study profile, which may have varied. Therefore, missing values were probably part of the data but the study did not explain how they were addressed.

Strecht et al. [132] conducted a study that applied classification techniques (such as K-Nearest Neighbour, Random Forest, AdaBoost, CART, Naive Bayes and Support Vector Machines) and regression techniques (such as Ordinary Least Squares, Support Vector Machines, CART, Random Forest, AdaBoost.R2) to predict student success/failure and marks in a module, respectively. The aim of the study was to compare the predictive accuracy of both the classification and regression methods, taking into consideration the performance metrics being different for the classification and regression methods. The research was based on the academic year 2012/2013, so was restricted in this sense. The researchers used only students' general characteristics associated with 5779 modules, which in turn were related to 391 programmes of study. They did not use previous module attributes. In terms of missing data, drop-out student marks were replaced with the value '0' in the final mark (which is the target variable), as regression did not accept non-numerical values but this does not make a distinction between drop-out and failure.

Other studies address similar module-level prediction problems using standard recommender system techniques. For example, Thai-Nghe et al. [73] applied a Matrix Factorisation technique to predict student performance in a given set of exercises from a tutoring system.

For validation purposes, they compared the developed system with logistic/linear regression methods which model the relationship between a dependent variable either continuous (linear) or categorical (logistic) and one or more predictors variables [133]. These methods have been used in predicting student performance in several studies[134, 135]. The results show that the Matrix Factorisation technique performs better than the other methods (logistic/linear regression), and logistic regression shows similar results to linear regression. The study used two educational data sets from the KDD (Knowledge Discovery and Data Mining) Challenge 2010. The first data set consists of 23 attribute and more than 9 million instances and the second data set consists of 21 attributes and more than 20 million instances. We believe that this work was based on complete datasets, as the target of the prediction is a correct first attempt ('1' or '0') for solving the given exercise. That is, '1' indicates a student successfully completed the exercise on the first attempt, and '0' indicates otherwise.

O'Mahony and Smyth [136] recommended previous choices from students with similar interests, whereas Unelsrød [137] recommended modules that were preferred by others in the student's social network. In terms of missing data, the study by Unelsrød [137] mentions that the used dataset includes many missing values, and the applied algorithms can handle missing data. Furthermore, Cho and Kang [138] proposed a system focused on suggesting a module that matches the student's preferences, even if the student might fail the module. The system was implemented using a hybrid filtering technique, a technique that combines the outcomes of the Collaborative Filtering and Content Based methods. They used real data associated with undergraduate students at their university. The study did not mention the size of the used data. The system used all the basic module and student attributes, in addition to the quantity of the required credits to be completed for a particular career path. They evaluated their system using the classification accuracy metric.

Regarding missing data, this is often not well addressed. For example, Mohsin et al. [139] and Schalk et al. [140] used a simple imputation mechanism for missing data (median for continuous variables, mode for categorical variables). Wook et al. [141] and Kabakchieva [108] stated that they had managed missing data, without mentioning any further details. Other studies such as that of O'Mahony and Smyth[136] and Cho and Kang[138] did not mention missing data.

In a recent study, Chau et al. [142] attempted to predict students at risk of graduating with poor overall performance. They used real data belonging to 1334 undergraduate students from one school of study, associated with one university. They applied their experiment on three datasets for 2nd, 3rd and 4th year students. Each student was associated with 43 attributes that represent the number of subjects in the programme. The first dataset (2nd

year) had the highest percentage of incomplete data, 50.34%, while the latter two datasets (3rd year and 4th year) had 31.77% and 21.14% incomplete data, respectively. The K-Nearest Neighbour method was used for the missing data imputation, although it is worth noting that Waljee et al. [143] showed that Random Forest imputation outperformed the K-Nearest Neighbour method. For the performance prediction, Chau et al. [142] applied semi-supervised versions of Random Forest, using self-training, a Support Vector Machine and C4.5. For measuring performance, the authors used classification accuracy (%) and One-Way ANOVA to check statistical differences. The results showed that a Random Forest self-training algorithm outperformed the others.

In terms of managing missing data in the wider context, Burgette et al. [144] and Arnold and Kronmal [145] presented epidemiological studies using multiple imputation. Burgette et al. [144] applied CART regression trees to the multiple imputed data. The purpose of Arnold and Kronmal's [145] study was to compare the analysis results between the imputed dataset and the complete data set, which was largely consistent. Additionally, Sambo et al. [146] proposed a novel Bayesian network tool to impute missing values among type 2 diabetes risk factors. There are other medical studies [147, 148, 149] which have applied data imputation and data mining techniques.

Table 2.3 summarises the most relevant performance prediction studies.

2.12 Summary

The primary objective of this chapter is to present a comprehensive review of the literature of EDM and any relevant management themes such as personalisation, decision making, choice, student decision making and data in the UK.

EDM is very similar to DM but it focuses on any type of educational data. As a result, it may have different objectives, the process of applying EDM can vary, and the utilised technological tools and applications can also be different. Second, EDM as an interdisciplinary area can apply methods and techniques not just from the DM field, but also from recommender systems, cognitive psychology, etc.

In this chapter we introduced some of the related works by grouping them into different types of studies, illustrating the four main educational issues that have been most frequently discussed in previous research. We found here that researchers applied several DM techniques to solve their educational problems or improve their educational setting. There was no overall winner as different techniques showed good results in different settings.

Lastly, we discussed in detail the most relevant works for what we want to achieve in this

Table 2.3: Summary of related works

	Input data		**Goal**		**Technique**		**Google Scholar**
Study citation	Using typical data (demographic, preferences, etc)	Using other types of data	address missing data	address management aspect	DM technique	Another technique	Number of citations in Google Scholar
[73]	✓				✓	✓	124
[138]	✓	✓				✓	4
[142]	✓		✓		✓		0
[36]	✓				✓		113
[11]	✓	✓			✓		37
[132]	✓		✓		✓		19
[137]	✓	✓	✓			✓	4 (MSc thesis)
[136]	✓	✓				✓	52
[130]	✓	✓			✓		17

thesis. We observed that the related works included different sizes of data, between 50 and 20 million records, depending on the availability and on the level of the studies (e.g. university level, department level, module level, etc.). They included different data types such as social interactions, exams, demographics, and so on. Some of the studies collected their data from a computer-based environment while others did so from a traditional education environment.

Our thesis, following the reviewed research articles, looks at predicting performance as both a regression and classification problem in order to understand whether one approach can provide better results than the other. Many of the studies have been conducted on a single dataset; we attempt to make our study more robust by using different datasets. We will examine the best DM algorithms to accurately predict student module performance in the context of extensive missing data, and to study the effect of the missing data on performance through a novel approach (Chapter 6).

An important aspect of our work, according to our review of the literature, is that it will address the management aspect of the module choice problem (Chapter 7). To the best of our knowledge, none of the related published studies have addressed this; instead they have solely focused on the technical aspect of the module enrolment system.

Chapter 3

Data Description

We were given access to the data stored in the university's data warehouse which related to students, their characteristics, performance, attendance and engagement. Our initial mandate was to understand how such data could be used to improve student performance and experiences and also to assess the quality of the data for such purposes and produce recommendations for its improvement. In this chapter we describe in some detail the data which we had access to, its characteristics, the pre-processing of it for the purpose of further analysis and we also introduce other datasets which we obtained to validate and consolidate our study.

3.1 The University's Data Warehouse

Our main data sets are extracted from a university in the United Kingdom. Most bachelor degrees in the UK take three years to complete. Some courses might extend to a fourth year because they include a work placement year or a year abroad. The majority of bachelor degrees are honours degrees. These, typically, are classified into one of four classes of honours, based on the marks achieved in examinations and assessments: first class honours (1st), when the weighted average mark is +70%, upper-second class (2:1), lower-second class (2:2), and third class honours (3rd), when the student achieves average marks between 60−69%, 50−59%, and 40−49%, respectively. Students usually attempt to achieve "good" honours degrees, which are considered to be those classified as first or upper-second class (i.e.$\geq$ 60%). This is to open employment opportunities, as employers often use good honours as a threshold for applications.

Each university tends to have many schools of study for different subjects. Each school

of study has a number of degree courses, including a number of modules that the student can take, some of them optional and some compulsory for a given degree. The first year of study does not count towards the degree classification, although students are required to pass it. The second and third years count towards the final degree, sometimes with different weightings. The third year usually contains the highest number of optional (*elective*) modules.

The used data was retrieved from the University of East Anglia's data warehouse, where information of the students and their performance is collected. Additionally, the data warehouse contains other important data that is required for external agencies, e.g. those collating league tables.

3.1.1 Data selection and pre-processing

Initially, when we started the experiments in Chapter 5, 19,811 records were provided, which corresponded to 984 undergraduate students that obtained their academic award throughout the years 2005 to 2013 and were enrolled to a particular school of study (School of Computing Sciences). In our experiments we analyse data for particular schools separately since results may only be meaningful for students undertaking the same programmes and taking similar module choices. For example, if an enrolment system is build in the future, predictions for each group of students will be made based on students of the same school of study.

After cleaning and filtering the data for the purpose of removing irrelevant items, the remaining data was associated with 898 students. For example, we removed the data for 25 students, because their first year data was missing due to either exemptions or transferring from a different school. These students have accreditation of prior learning (APL) recorded in the university's data warehouse to indicate that the equivalent work has been done elsewhere. This elimination was important because some of the focus of Chapter 5 is to identify student at risk of poor outcomes at the end of Year 1 but using their Year 1 performance for the prediction.

Additionally, for quality purpose we removed data that corresponded to 55 students, because for some reason that would require further investigation they appeared to have taken the investigated first-year modules in their second or third year of studies. We also removed 6 students on discontinued courses.

However, In Chapter 6, 66 additional student records became available to us, which were associated with students who completed their academic degree in 2014 and 2015.

Next, we were provided with 38,608 records that corresponded to 2,214 undergraduate students that obtained their academic award throughout the years 2005 to 2013 but were

enrolled to a different school of study (Norwich Business School) associated with a different discipline. We also cleaned and filtered the data by removing data for 416 students because their first-year data was missing as with the other dataset. We also removed data corresponding to 9 students either because of data linking errors or because they did not take the investigated first-year modules. The remaining data was associated with 1,789 students. We should note that 637 students have missing library loan records which may also signify less engagement with the university. 876 students have missing attendance records due to data quality issues. Again, for the experiments in Chapter 6 an additional 180 student records became available to us. These records corresponded to students that completed their academic degree in 2014 and 2015.

For both schools, we did not include data prior to 2005, because of problems with data migration which compromised some of the quality of the earlier data.

In both datasets, each student was represented by one row of data regardless of their study profile. Tables 3.1 and 3.2 represent the attributes that we choose to made up each initial student record. Table 3.5 summarises the cleaning process of our data.

Our outcome variable (the award class in Table 3.1) for overall performance prediction was whether the students obtained Good Honours (GH) or Not Good Honours (NGH). Those in the GH class were individuals who were awarded a CLASS I*, CLASS I, or CLASS II, DIV 1 degrees. Those that achieved any other degree classifications were labelled as NGH. The grading scheme in this thesis is based on the British Higher Education system. An explanation of undergraduate grading system in the UK can be found in [150]. To perform the classification task, we transform the final overall weighted scored for a student into a Good Honours/Not Good Honours categorical label, as shown in Table 3.3

One of the attributes in Table 3.1 is the English Entry Qualification. This indicates if a student took any of the following subjects before his/her degree:

- 'English'
- 'English Language'
- 'English Language & Literature'
- 'English Literature'

If the student took such subjects, then the field value is 'Yes', otherwise the value is 'No'. We have precisely chosen the above four English subjects, because they were the only English subjects available in the students' entry qualification records. It should be noted that the reason we did not take into consideration the IELTS[1] or TOEFL[2] test scores for OS and

[1] `https://www.ielts.org/`
[2] `https://www.ets.org/toefl`

Table 3.1: This table shows attributes relating to student demographics and general performance that represent each student.

Attribute	Type	Description of values
gender	categorical	female / male
age band at entry	ordinal	16-20, 21-24, 25-34, 35-44, and so on. As stored in the University data warehouse
disability	binary	Yes/No
level of widening participation in Higher Education	ordinal	very low, low, medium, high, very high, Non-UK
nationality	categorical	nationality1 / nationality2 / .. etc.
overall score in year 1	decimal	0-100
overall score in year 2	decimal	0-100
overall score in year 3	decimal	0-100
the award class	binary	(G)ood(H)onour / (N)ot (G)ood (H)onour
fee status	categorical	(H)ome / (O)verseas / (EU)ropean
foundation year	binary	Yes/No
English entry qualification	binary	Yes/No
Maths entry qualification	binary	Yes/No
name of the course of enrolment	categorical	course1 / course2 /.. etc. The first and the second school include 4 and 14 courses, respectively.
library loans in year 1	integer	1-15 (no. of items)
library loans in year 2	integer	1-15 (no. of items)
library loans in year 3	integer	1-15 (no. of items)
the year they obtained their academic award	number	2005-2015
UCAS tariff points	real number	As stored in the University data warehouse.

EU students was because they have not been collected in the university's data warehouse at present.

Another of the attributes in Table 3.1 is the Maths Entry Qualification. This indicates if a student took any of the following subjects before his/her degree:

- 'a mathematical subject'
- 'Additional Mathematics'
- 'Further Mathematics'

Table 3.2: This table shows attributes relating to modules that represent each student.

Attribute	Type	Description of values
name of module	string	module1 / module2 /.. etc
module code	string	code1 / code2 / .. etc
number of students enrolled on the module	integer	0-999. As stored in the University data warehouse
average mark for module computed for students registered at the same time as the current student as this was not stored in the data warehouse	decimal	0-100
individual mark for module	decimal	0-100
percentage of sessions attended for a given module	decimal	0% - 100%
performance of student compared to his/her peers	categorical	Fair, Average, Poor as described in Table 3.4

Table 3.3: Categorical value transformation for students' numeric scores.

Continuous Values	**Categorical Values**
student's score ≥ 60	(G)ood (H)onours
student's score < 60	(Not)(G)ood (H)onours
missing value	Not Taken

- 'Mathematics & Statistics'
- 'Mathematics (I)'
- 'Mathematics'
- 'MEI Further Mathematics'
- 'MEI Mathematics'
- 'Pure Mathematics'
- 'Pure Maths. & Statistics'
- 'Statistics'
- 'Use of Mathematics'

- 'Using and Applying Statistics'
- 'Working with Algebra'
- 'Using Numbers'

If the student took such subjects, the attribute value is 'Yes', otherwise the value is 'No'. As with the previous field, we have precisely chosen the above Maths-related subjects, because they were the only subjects available in the students' entry qualification records.

We initially included data that was suggested to us as representing a measure of engagement. That is loan records from the library (in Table 3.1) and attendance records for module sessions (in Table 3.2). However, it is worth nothing that the quality of this data is questionable at present. For example, 492 students have missing library loan records which could be interpreted as showing less engagement with the university. They have never borrowed an item from the university library but this could be for a number of reasons that may not be related to engagement. 534 students have missing attendance records. Again, it could be that such missing data represents data quality issues and in the case of attendance monitoring in particular data quality issues are known to exist.

In Table 3.1 we present a modified version of 'name of the course of enrolment' attribute as shown in Figure 3.1, as some of the course names have changed over the years. The unmodified version resulted in having very low number of students enrolled in some of the courses. For example, during the past nine years only one student has graduated with the degree of 'Computing for Business within a Year in Industry'. The very low number of records contrasts with the general concept of DM application because it will not produce meaningful results. Therefore, we unified the names of the courses to be consistent for the past nine years. There were low numbers of students enrolled in 'Software Engineering' and 'Computing for Artificial Intelligence' courses, but we left them because there were no equivalent/similar courses that they could be merged with. Additionally, some courses such as 'Master of Computing in Computing Science' and 'Master of Computing in Computing Graphics' that refer to extensions to Undergraduate Degrees where merged with the equivalent UG degree.

In Table 3.1, the UCAS (Universities and Colleges Admissions Service) *tariff points* attribute indicates the points that a student is allocate according to the different qualifications at entry. We discounted this attribute in the analysis phase since it is only associated with Home students except for very low number of students who either have an accredited equivalent to British A-Levels or have completed their A-Levels in the UK.

In Table 3.2, as part of the module attributes used for each student we use performance of each student on a given module compared to peers. This is calculated according to the

boundaries given in Table 3.4. However, those boundaries are somehow arbitrary and were obtained from the traditional 5-points grading scale. We intend to experiment with different boundaries in the future.

Table 3.4: Description of Performance Field

Value	**Description**
Fair	student mark > (module' s average mark + 5%)
Average	(student mark >= module‘s average mark - 5%) and (student mark <= module‘s average mark + 5%)
Poor	student mark < (module‘s average mark - 5%)

In Chapter 5 we conduct experiments to predict overall performance. This is to determine whether there are any statistically significant patterns that could be exploited from students that completed their degree without obtaining good honours. Those could subsequently be used to suggest staged interventions for other students with similar characteristics to improve performance when possible. Therefore, we use the following two sets of attributes: attributes that relate to student demographics and general performance as described in Table 3.1, and attributes that related specifically to student modules as described in Table 3.2.

In Chapter 6, we attempt to predict elective (or optional) module performance. For this we select five optional modules with the highest number of students from each school of study, and handle each of the modules as an individual dataset with one row per student. For each row, we have general characteristics of students (as per Table 3.1) and we use a list of modules they took and their overall mark for the module (as per Table 3.2), including the module of interest which becomes the prediction target. We ignore other attributes from Table 3.2. To perform the prediction of module performance as a classification task, we transform the final mark for a module into a Good Honours/Not Good Honours categorical label, as shown in Table 3.3 and add it to the datasets. We therefore have for each student a list of general characteristics along with a number of modules and their marks on those. The target will be the selected optional module mark (0-100 for the regression task); or GH/NGH for the classification task.

Table 3.5: This table summarises the cleaning process of our data.

CMP Data	NBS Data
– we did not used data prior to 2005 due to a data migration issue that affected its quality. – we removed 25 students due to the absence of their first year data. – we removed 55 students due to their quality. They appeared to have taken the examined first-year modules in their second/ third year of studies. – we removed 6 students on discontinued courses. – each student was represented by one row of data. – Table 3.1 and Table 3.2 show the selected attributes.	– we did not used data prior to 2005 due to a data migration issue that affected its quality. – we removed 416 students due to the absence of their first year data. – we removed 9 students due to data linking errors. – each student was represented by one row of data. – Table 3.1 and Table 3.2 show the selected attributes.

For the first school of study, the 5 modules chosen, along with the number of students enrolled on them is shown in Table 3.6. The module offering can change from year to year, and different programmes are associated with different first-year choices. This lack of homogeneity results in missing values appearing in the datasets for some first-year module attributes, since, when looking at optional modules, the students that took a particular module and are included in the dataset may have different associated first-year modules. Thus, the missing data are unobserved values missing at random (MAR). This mechanism for generating missing data does not preclude the use of imputation procedures. Table 3.7 provides an indication, for each dataset, of the number of first-year modules associated with the selected optional module, the number of those that contain missing scores and the average (standard deviation in brackets), minimum and maximum proportions of missing year-1 values. Missing data is only associated with first-year module marks, as other attributes are complete.

Next, for the second datasets, Table 3.8 shows the optional modules selected on the basis of the highest number of students enrolled on them. Table 3.9 indicates, for each dataset,

Table 3.6: The first school's selected optional modules with the number of students that took them. The module's acronym is in brackets.

Module Name	Number of students
DATABASE SYSTEMS (DS)	368
NETWORKS (NW)	351
INTERNET TECHNOLOGIES (IT)	317
SYSTEMS ANALYSIS (SA)	260
SYSTEMS ENGINEERING (SE)	239

Table 3.7: Missing data information for each dataset in the first school of study, including mean, minimum and maximum proportion of records for each Year1 module with missing values.

Module Name Acronym	Number of Year1 Modules	Number of incomplete Year1 Modules	Mean (Standard Deviation)	Min	Max
DS	25	25	.511(0.499)	0.484	0.997
NW	26	26	0.525(0.499)	0.488	0.997
IT	17	17	0.377(0.485)	0.102	0.989
SA	21	21	0.445(0.497)	0.425	0.996
SE	21	21	0.447(0.497)	.500	0.979

the number of first-year modules associated with the optional module, the number of those that contain missing scores, and the average (standard deviation in brackets), minimum and maximum proportions of missing values in year-1 modules. Again, missing data is only associated with first-year module marks, as other attributes are complete. This second dataset has fewer missing values since the year-1 modules have been more stable over the years.

We want to assist students in choosing options for year-3, and this is often done before the year-2 module marks are known. Hence, the prediction of module outcomes only takes into consideration year-1 module performance.

Table 3.8: Selected optional modules with the number of students that took them for the second school. The module name's acronym is in brackets.

Module Name	Number of students
ENTREPRENEURSHIP AND SMALL BUSINESS (EB)	731
INTERNATIONAL FINANCIAL SERVICES (IS)	440
PERSONAL AND CORPORATE TAXATION (PT)	468
STRATEGIC BRAND MANAGEMENT (SM)	585
FINANCIAL MODELLING (FM)	336

Table 3.9: Missing data information for each dataset in the second school of study.

Module Name Acronym	Number of Year1 Modules	Number of incomplete Year1 Modules	Mean (Standard Deviation)	Min	Max
EB	13	13	0.191(0.393)	0.003	0.998
IS	11	6	0.130(0.336)	0	0.993
PT	11	9	0.125(0.331)	0	0.888
SM	11	10	0.123(0.329)	0	0.838
FM	8	6	.069(0.254)	0	0.264

3.1.2 Publicly available datasets

For additional validation of our methodology in Chapter 6, we used two public student performance data sets from the Machine Learning Repository website, which are associated with a study by Cortez et al. [151]. To further validate imputation of missing data, we included the two complete datasets, but we also created two other versions of the same data with 25% and 45% missing values, as we randomly removed some of the available data. In this scenario, the data is removed by an MCAR mechanism. Therefore, in total we include six datasets from this source in the comparison.

The first of these datasets corresponds to secondary school Maths subject (*module*). The second dataset is associated with secondary school Portuguese (Por) language subject. The data attributes have been explained by [151]. Appendix A includes a copy of the explanation

of the attributes that represent each student in the public dataset. As a summary, every student has three assessment grade variables in addition to the other variables: G3 is the mark of the final evaluation for the module (the target output); G1 and G2 are the marks of the first and second assessments. For consistency, we did the following:

- We attempted to predict the G3 variable without G2, since for our data we tried to predict a particular year-3 module grade without using year-2 grades.
- For the classification task, we transformed G3 marks to two levels: GH when the student achieved +60%; otherwise, NotGH.

3.2 Data Collection

Qualitative study can be performed by using different data collection methods or by selecting one method in particular. Marshall and Rossman in [152] argued that data collection approaches in qualitative research could be classified into four types: in-depth interviews, direct observation, participation in the setting, and document analysis. To perform our study in Chapter 7, we utilised in-depth, individual interviews as the main approach to collecting our data. We collected additional data by conducting a questionnaire survey and from documents provided through the staff interviews.

Coffey and Atkinson [153] have explained that data collection and analysis are better conducted at the same time in qualitative research to allow for essential flexibility. Hence, data collection and analysis followed a cyclical process until themes and concepts became redundant and detailed, and new information ceased to appear (Miles and Huberman [154]; Strauss and Corbin [155]). Appendix B includes our final coding themes that have arisen from the transcription analysis.

Our sample includes two different schools in two different faculties with dissimilar admission strategies, both for validation purposes and in order to be consistent with our previous work in Chapter 5 and Chapter 6. The collected data is associated with one university due to the ethical considerations involved in collecting data from students and staff members of different UK universities.

We believe that using data associated with students who enrol in the same programmes and experience similar module choice, throughout the thesis work, will add more meaning in connecting our PhD study. Therefore, the main group of research subjects were:

- Year 2 and Year 3 undergraduate students (59 students participated in our questionnaire survey, and 28 students were interviewed). Students were drawn from the Computing Science School (CMP) and the Norwich Business School (NBS) in the

University of East Anglia (UEA). We targeted Year 2 and final year students, as they have experienced choosing and studying elective modules.

- Seven staff members associated with different roles (*levels*) at the university. These roles included a pro-vice chancellor, university-level academic directors and undergraduate programme leaders in the two schools.

To ensure data confidentiality and anonymity, we did not collect personal identifiable information and we anonymised the results of Chapter 7. In addition, in order to avoid identifying individuals, we have omitted the transcripts of the interviews from the appendices. However, in Appendix C, we include the dates and the duration of each conducted interview.

In the following chapter, particularly section 4.6, we explain thoroughly the methodology of collecting our data.

3.3 Summary

In this chapter, we presented the data used throughout our research, including the input attributes and targets for the predictive models, the missing values, how we filtered and managed the datasets, and the derived datasets for each experiment. In summary, we completed our research by using three types of data:

1. Data that has been retrieved from the university's data warehouse. This type of data corresponds to undergraduate students from two schools of study and has been used in building the predictive models in Chapter 5 and Chapter 6.
2. Data retrieved from the Machine Learning Repository website, which is publicly available. This type of data has been used to validate our methodology in Chapter 6
3. Data that we have collected ourselves by conducting both a questionnaire survey, which included 59 responses from students, and interviews with 28 students and seven staff members. The collected data corresponds to participants from UEA, particularly the students were associated with the same two schools of study used in the experiments of Chapter 5 and Chapter 6.

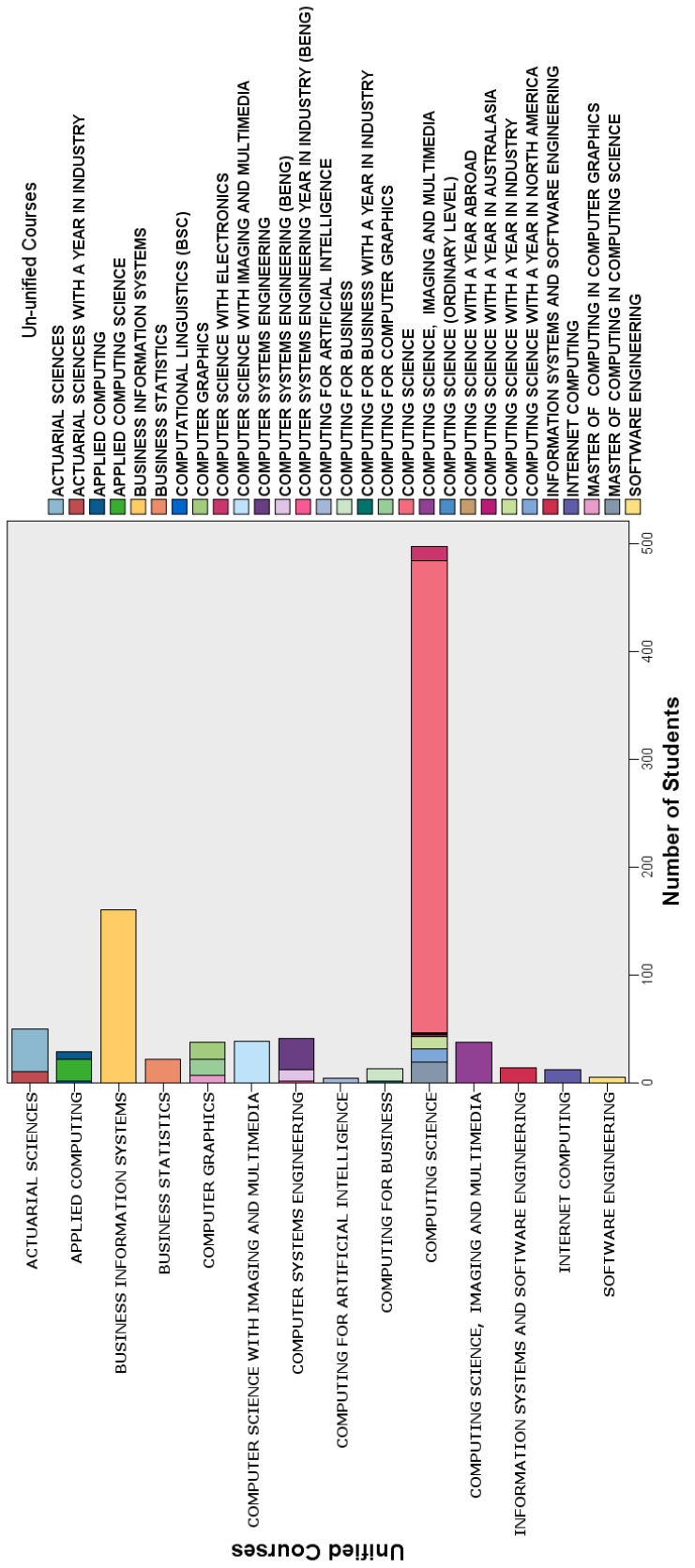


Figure 3.1: The unification of the courses' name. The y-axis shows each bar overlays with the courses that have been merged together, while the x-axis show the number of students.

Chapter 4

Research Methodology

This chapter describes our research methods. Section 4.1 introduces the experimental design and presents how the different strands of research come together as part of the thesis. Section 4.2 discusses the prediction methods applied. Section 4.3 discusses the evaluation techniques for the prediction experiments. Section 4.4. explains the methods that can be applied for dealing with missing data, including our own proposal of multiple imputation and an ensemble. Section 4.5 introduces the used software tools. Section 4.6 describes the methodology of our management study. Section 4.7 explains the ethical considerations of our research. Finally, section 4.8 summarise this chapter.

4.1 Performance Prediction from Student Data

In this thesis, we focus on the performance prediction of undergraduate students but we also want to extend the work by looking at the management aspect and how the university could utilise the derived knowledge from the predictive models. Nowadays, university managers are required to translate the derived knowledge into improved decision making and performance [156]. Also, model deployment, the last phase of the CRISP-DM process [157], is the 'gold standard' for a successful project [158]. Due to time constraints and the complexity of ethics requirements, we were not able to build a prototype of an enrolment system and test it on students. Instead we perform an initial qualitative study to understand user (students and staff members) perceptions of module choices and their effect on performance, and the assistance they could receive from an enrolment system.

Therefore, we started our thesis investigating the prediction of overall outcomes. Then we produced more granular performance predictions at the module level. Lastly, we studied

the management aspect of utilising the derived knowledge. Figure 4.1 summarises our thesis result chapters and their research design.

4.2 Prediction Methods

In this thesis, we focus on DM techniques, because we believe that accurate performance predictions may be useful to students and educators alike. Predictions can provide students with additional information at the time of making module choices, for example, or they could be used by educators to offer remedial sessions or other similar actions that may improve performance. At this point, we focus on predicting academic performance based on general characteristics and previous performance, as for this we have data of sufficient quality. However, we envisage accommodating other priorities, including employability, engagement indicators and others, as those become available with better quality. In what follows, we present our methods.

4.2.1 Feature Selection Ranking algorithm

Feature selection[159] is one of the important and frequently utilised techniques in data preprocessing to maintain useful features through eliminating irrelevant and redundant features, resolving the dimensionality problem, improving classification performance and speeding up the DM algorithm. According to the IBM Knowledge Centre [160] feature selection involves three steps:

1. Screening which is eliminating problematic and statistically insignificant inputs and cases, or records, for example, removing input variables with too many missing values or with too much or too little variation to be beneficial.
2. Ranking which is sorting the remaining inputs and allocating ranks based on importance.
3. Selecting which is defining the set of features to be utilise in subsequent models, for instance, by keeping the most important inputs, and excluding or filtering all others.

We use feature selection when trying to predict overall performance to understand the predictive capabilities of the attributes available to us.

4.2.2 Regression versus classification

Regression is a predictive modelling method that maps each attribute set into a continuous-valued output, which means the response variable to be estimated is continuous [161]. The

regression aims to discover a mapping technique that can fit the input data with the least possible error. [161]. In regression, the error function can be computed in terms of squared error or sum of absolute differences:

$$SquaredError = \sum_{i=1}^{n}(Y_i - f(X_i))^2 \tag{4.1}$$

$$AbsoluteError = \sum_{i=1}^{n}|(Y_i - f(X_i))| \tag{4.2}$$

where

n is the number of objects.

Y_i is the response variable.

$f(X_i)$ is the function that maps the input data (or predictor variables X_i).

In our context the response variables for each module range from 0 to 100. In contrast, classification is a predictive modelling technique that maps each attribute set into one class-label output, which means the response variable is one of several predefined categories [161]. The response variable's type is the key feature that distinguishes classification from regression. In our context, the class labels for each module are "Good Honours (GH)/ Not Good Honours (NGH)" (defined in Table 3.3). The categorisation of GH/NGH is important in the UK Higher Education setting, as universities must report the percentage of GH degrees awarded, and so they are interested in strategies that may improve GH achievement.

4.2.3 Algorithms applicable to classification/regression methods

The selection of algorithms for the overall performance prediction, in Chapter 5, are based on all the applicable classification techniques that are included in IBM SPSS Modeler tool. These algorithms are Logistic regression, Neural Networks, Decision Lists, Bayesian Networks, Discrimination analysis, and four decision trees (Quest, CHAID, C&R Tree, and C5.0). However, for the module prediction, in Chapter 6, the algorithms were chosen as those most used in the research community [162]. In addition they showed better results in the context of dealing with missing data based on the literature studied in Chapter 2. These algorithms are Rpart, C5.0, Random Forests and SVM. For all of the selection techniques, we consider their suitability for our data characteristics.

Rpart [163] is used for regression and classification analysis. It uses the *Classification and Regression Tree (CART)* algorithm. This means the tree model is going to be built through,

first finding a single variable that will "best" divide the data into two groups. Then this process is applied individually to each subgroup, and so on, recursively, until a termination criterion is met. In this context, "best" is computed using the *Gini impurity measure* both for selecting variables and best split value. Rpart accepts missing values using what is called the *surrogate variable method* (more detail is given in [164]) and that makes it suitable for our data characteristics.

C5.0 is a supervised learning algorithm. It is an improved version of C4.5 introduced by Quinlan [165]. C5.0 generates decision trees from root to leaves using the *entropy and information gain* measure (more explanation is provided in [166]). In this work we use a particular implementation [167] for classification problems.

Random Forests [168] is a class of ensemble learning techniques for regression and classification. A Random Forest is the combination of multiple decision tree predictions, where each tree depends on the values of an independent set of random vectors. A Random Forest is a very effective prediction method because it tends not to overfit the training set [168].

SVM (Support Vector Machine) [169] is a supervised learning algorithm that is applicable in classification and regression analysis. Basically, SVM generates nonlinear boundaries by building a linear boundary in a large, transformed version of the feature space [170].

Logistic regression [171] is a statistical technique for classification, similar to linear regression technique, but it accepts a categorical target variable instead of a numeric value. Basically, logistic regression builds a group of equations that link the values of the input variables to the probabilities corresponded with each of the output categories. After the model is built, it can be utilised to compute probabilities for new data. For each row, a probability of membership is calculated for each attainable output category. Then, the target category with the maximum probability is allocated as the predicted output value for that record.

Neural Network [172] is a model that tries to emulate the way the human brain processes information. It consists of three parts, an input layer, with units signifying the input variables; one or more hidden layers; and an output layer, with a unit or units signifying the target variables. The units are linked with various linking strengths (weights). It functions by simulating a big number of interconnected simple processing units that look like abstract versions of neurones. Typically, the model trains by studying the records individually, producing a prediction for each record, and making alterations to the weights whenever it makes an inaccurate prediction. This process is repeated several times, and the model continues to enhance its predictions until it meets one or more of the terminating criteria.

Decision List [173], basically, works by identifying segments or subgroups that present a

lower or higher likelihood of a provided binary outcome respective to the overall population. Decision List models involve a list of rules in which every rule has an outcome and a condition. Rules are employed in sequence, and the first rule that applies determines the outcome.

Bayesian Network [174] is a visual model that shows variables (nodes) in a dataset and the conditional, or probabilistic, dependencies between them. Causal relationships [175] between nodes may be denoted by a Bayesian network; nevertheless, the links in the network (which are also known as arcs) do not usually reflect direct effect and cause. It can be built by combining recorded and observed evidence with real-world knowledge to form the likelihood of occurrences. The SPSS Modeler implementation is based on Tree Augmented Naive Bayes and Markov Blanket networks that are mainly used for classification.

Discrimination Analysis [176] works by building a predictive model for group membership. The model consists of a discriminant function (or, for more than two groups, a collection of discriminant functions), which depends on linear combinations of the predictor fields that deliver the greatest discrimination between the groups. The functions are created from a sample of cases for which group membership is acknowledged. Then, the functions can be employed to new cases that have measurements for the predictor fields, but have unidentified group membership. Discrimination analysis can create more stringent assumptions compared to logistic regression models. However, it can be a valuable supplement or alternative to a logistic regression model when those assumptions are fulfilled.

Quest [177] is a binary classification technique for creating decision trees. This method tries to minimise the processing time required for large CART analyses, while also minimising the tendency found in classification tree techniques to favour inputs that result in more splits. The input variables can be continuous, however, the target variable must be categorical. All the tree splits are binary. It utilises an order of rules, depending on significance tests, to assess the input variables at a node. For the purpose of speeding the analysis, QUEST approach is unlike C&R Tree technique because all splits are not investigated, and unlike C&R Tree and CHAID approaches, category combinations are not verified when assessing an input variable for selection.

CHAID [177] creates decision trees using the chi-square statistic to determine ideal splits. Unlike the QUEST and C&R Tree approaches, CHAID can produce non-binary trees, so some splits have more than two branches. Both target and input variables can be continuous or categorical.

C& R Tree [164] stands for Classfication and Regresstion Tree. It is in fact the original implementation of Rpart technique (mentioned earlier) [177]. Hence, it functions in the

same way as Rpart.

4.2.4 Clustering method

Our first attempt at prediction is achieved using a clustering approach which we devised to take into account the similarity of students for the prediction of performance. As this does not rely on using a standard algorithm, we explain here how the method works. The basic idea is to cluster similar individuals and then obtain a predicted grade as the average for regression (or most commonly occurring class for the classification) of all individuals in the cluster to which the student belongs. To establish our clustering, we first need a measure of object dissimilarity.

The Gower dissimilarity measure [178, 179] is used to handle numeric, ordinal, nominal, and asymmetric binary data, and can also deal with missing values, so we apply it to our problem. It standardises each numeric attribute $[i]$ to a range [0,1] through dividing each record by the range, R_i, (i.e. the difference between the lowest and highest values) of the same attribute [180]. It calculates the final dissimilarity between the xth and yth object as a weighted sum of dissimilarities for each attribute:

$$d(O_x, O_y) = \frac{\sum_{i=1}^{n} \delta_{o_x o_y i}\ d_{o_x o_y i}}{\sum_{i=1}^{n} \delta_{o_x o_y i}} \tag{4.3}$$

where

n is the number of attributes in each object.

$\delta_{o_x o_y i}$ is the weight of the attribute $[i]$ and that is:

- 0 when the column is asymmetric binary and both objects (o_x, o_y) have a value of 0, or when one or both objects (o_x, o_y) have a missing value for the ith attribute.
- 1 otherwise.

$d_{o_x o_y i}$ is the distance between xth and yth object, taking into account the ith attribute. It is determined by the nature of the attribute. For nominal or binary attributes the value of $d_{o_x o_y i}$ is 0 if both (o_x, i) and (o_y, i) are equal, 1 otherwise. For numeric-scaled attributes, the value of $d_{o_x o_y i}$ is the absolute difference of both objects' values, divided by the total range of that attribute. For ordinal attributes, their values first are replaced with the matching position index in the factor level ($r_{o_{xi}}$). Then, they standardised through the following formula :

$$z_{o_{xi}} = \frac{(r_{o_{xi}} - 1)}{Max(r_{o_{xi}}) - 1} \tag{4.4}$$

Lastly , the new values $z_{o_{xi}}$ will be handled as the numeric-scaled attributes.

Moreover, the dissimilarity $d(o_x, o_y)$ will remain in the range $[0, 1]$ since the value of $d_{o_x o_y i}$ falls in this same interval. The dissimilarity will be set to NULL, if all weights $\delta_{o_x o_y i}$ are zero.

Hierarchical clustering includes two basic approaches: divisive methods [179] and agglomerative methods [181]. Divisive techniques are a "top down" approach as they begin with one all-inclusive cluster and, at each step in the algorithm, split a cluster into new smaller clusters until only clusters of individual objects remain. This approach requires a method that helps in deciding which clusters should be divided at each step and a method for the division. Agglomerative techniques are "bottom up" techniques, as they work in the opposite way. They start with the objects as individual clusters, then merge the closest pair of clusters as the algorithm moves up the hierarchy. This approach requires identifying the notion of cluster proximity. Agglomerative techniques are more widely used and so we selected them for our experiments.

Partitioning Around Medoids (PAM) is a common implementation of the k-Medoids algorithm [182]. It is very similar to k-means, except for the computation of medoids in PAM rather than centroids. The PAM algorithm is presented as algorithm 1.

Require: The number of cluster k, dataset with — objects
Ensure: A set of k clusters
Randomly select k objects as the initial medoids
repeat
  Assign each remaining object in the dataset to the cluster with the nearest medoid
  for *each medoid* **do**
    Randomly select non-medoids object
    Compute the swapping cost function to replace medoid with non-medoid object
    if *the replacement can decrease the value of the cost function* **then**
      swap is confirmed
    else
      the medoid is not replaced
    end
  end
until *no more change is possible*

Algorithm 1: PAM algorithm

The clustering method will work as follows:

- First, we will apply the general dissimilarity coefficient of Gower to our data frame

since we have mixed (categorical and continuous) data types. We can then visualise and explore distances using a heatmap [183].

- We will then apply two different clustering algorithms on the dissimilarity matrix: hierarchical clustering (with average linkage) and PAM. We chose those methods among other clustering methods because they outperform others, such as CLARA and DIANA [179], and secondly, they accept a dissimilarity matrix as an input [184, 185].
- Next, the clustering technique with the best performance will be chosen after investigating three well-known internal validation methods. These methods are the Dunn index, which measures both compactness and clusters separation [186]; Silhouette, which measures how well objects fit within their clusters [187]; and Connectivity, which evaluates the degree to which neighbouring objects were placed in the same cluster by computing penalties for each object [188]. After that, we will apply the Cophenetic Correlation Coefficient (CPCC), which is the correlation of the dissimilarity matrix and the agglomerative hierarchical clustering techniques, and is a standard evaluation of how well hierarchical clustering of a specific linkage type fits the data [189].
- Lastly, once the cluster solution is constructed, we will use it to predict optional module marks for new students by assigning the student to the closest cluster, then compute the cluster average marks for the selected optional module, eventually using this computed average as a predictor.

4.2.5 Ensemble methods

Ensemble methods are also recognised as model combiners or committee methods [190]. They are machine learning techniques that leverage the ability of several classifiers to attain better accuracy in comparison to what any of the individual models may attain [191]. The ensemble methods obtain the predictions of their multiple models and then combine them in a suitable approach, such as averaging or voting. Studies in ensemble methods have largely focused on including models that are competent yet also complementary, i.e. diverse. Ensemble methods are as prone to over-fitting as any other model; therefore, it is necessary to apply cross-validation for ensemble evaluation [191]. We use ensembles in the prediction of overall performance, and we also use them to combine multiply imputed datasets to counteract the problem of missing data in our module prediction problem.

4.3 Evaluation of Models

4.3.1 Metrics of performance for predictive models

Root Mean Square Error (RMSE) [192] measures the difference between the predicted values and the actual values for regression problem. RMSE is computed through taking the square root of the average value of the square of the residual (actual value - predicted value) as shown in Equation 4.5. Smaller RMSE values indicate better model performance. The RMSE value should be positive to be valid.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{4.5}$$

where

n is the number of objects.

Y_i is the actual value.

$\hat{Y}_i$ is the predicted value.

Accuracy [193] is a statistical measure of how well a classification algorithm correctly predicts the classes. It measures the ratio of correct predictions to the total number of cases examined.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{4.6}$$

where

TP is the number of true positive cases;

FN is the number of false negative cases;

FP is the number of false positive cases;

TN is the number of true negative cases.

F1-score [194] (sometimes called F-measure) is another measure used in classification problems and represents the harmonic mean of recall and precision (or sensitivity and positive predictive value).

$$F1\text{-}score = \frac{2 \times Precision \times recall}{Recall + Precision} = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{4.7}$$

Gain Charts are visual aids for measuring model performance. They show a visual summary of the efficacy of the classification model calculated as the ratio between the classification results obtained with and without model[195],[196, p.212].

4.3.2 Measuring generalisation in predictive models

All those metrics can be calculated and compared to assess the goodness of fit of a particular model. However, the training data should not be used as the main vehicle to measure the performance of the learned classifier. Instead a test set is required for the purpose of evaluating the performance of the learned model. The learned model (e.g. classifier) will be run to predict a label for each record in the test set, then performance is evaluated by comparing the prediction with the actual labels of such a dataset. In this way, the test set acts as an approximation to new data and can measured the generalisation capabilities of the learned model.

The train/test partition is a common way to measure generalisation performance. However, when not sufficient data for train/test partitions is available, another common method to measure performance on new data is *k-fold Cross-validation*. Cross-validation also gives an indication of how the machine learning model will generalise to independent data and reduces the over-fitting problem. It consists of splitting the dataset into k disjoint splits (folds) of equal size. A model is run k times and each time one fold acts as the test set with the rest being used as the train set. The performance is estimated as the mean of the computed k scores across all folds. Each fold should preserve the distribution of the class variable from the whole original dataset [197]. We have used 10-fold cross-validation in our thesis instead of other known validation techniques such as leave-one out cross-validation [198], because the latter could be highly variable according to Ambrosie and McLachlan in [199].

4.3.3 Statistical tests

Often it is necessary to check whether differences, say in the performance of two classifiers, or from a number of classifiers, are statistically significant for any of the metrics captured. The following are the statistical tests that have been applied in our set of experiments.

F-test (one-way ANOVA) [200] is a test that used to compare groups within a field. It is depend on the ratio of the variance between the groups and the variance within each group. The F ratio is expected to be nearly to 1 if the means are the same for all groups. This is because both are calculated from the same population variance. The greater the ratio, the larger the variation between groups and the higher chance that a statistically significant difference exists.

Pearson chi-square testing is a statistical test that proposed by Karl Pearson according to [201]. It reflects the probability that the two variables are unrelated, in that case any

dissimilarities between observed and expected frequencies are the outcome of chance alone. If this probability is very small, mostly less than 5%, then the relationship between the two variables is considered to be significant. The chi-square statistic should be interpreted cautiously if any of the expected cell has less than 5 values. A one-way chi-square test is when there is only one row / one column, in this case the degree of freedom is the number of cells minus one. The degree of freedom for two-way chi-square is the number of rows minus one times the number of column minus one [202].

Wilcoxon signed-rank test[203] is basically a non-parametric substitute to the paired t-test, which ranks the variances in performances of two models for each data set, discounting the signs, and compares the ranks for the negative and the positive differences. It assumes commensurability of differences [204].

The Friedman rank test [205, 206] [1] is a non-parametric alternative to the repeated-measures ANOVA test. It is considered safer than ANOVA, because it does not require normal distributions or homogeneity of variance. It is the recommended choice to compare a number of algorithms over a number of datasets [204]. We present our comparisons through critical difference diagrams using the post-hoc tests (Nemenyi test in our case) after the Friedman test as recommended by Demšar [204]).

In Chapter 5, we use Pearson chi square to test the relationship between the student attributes and their overall outcomes. We use the F-test to compare the mean mark of students in the GH and NGH group for each module. We also use gain charts to visualise differences in performance.

In Chapter 6, we compute the RMSE for each regression prediction system. We also compute the accuracy and F1-score for each classification prediction system. We use the F-1 score as some studies [207, 208] consider it better than accuracy. We used Friedman rank test and critical difference diagrams to compare the mean of (RMSE / Accuracy/ F1-score) for the various datasets across multiple prediction systems. We also use the Wilcoxon signed-rank test to compare the mean of RMSE/Accuracy between the complete and imputed public datasets.

4.4 Dealing with Missing Data

Often classification/regression models need to be developed in the context of extensive missing data. Although there are some techniques to handle missing values, none guarantee

[1] We used the code available at `http://theoval.cmp.uea.ac.uk/` and adjusted it to work with both accuracy and RMSE.

best performance. Different situations need different solutions but as Allison [209] stated, "the only really good solution to the missing data problem is not to have any". In this section, we discuss our proposed approach for deal with missing data as this is a problem we encountered trying to predict module performance.

In the clustering approach we introduced earlier, both the hierarchical clustering and PAM methods can accept a dissimilarity matrix as input. If the Gower dissimilarity measure is used, as we have proposed, it will automatically handle the missing values in the dissimilarity calculation as explained earlier in section 4.2.4 hence in proposing that method we have already taken account of missing data handling.

For classification/regression, one possibility is to use an algorithm that can construct a model in the presence of missing data. Two examples of regression or classification algorithms that can deal with missing data are Recursive Partitioning (or Rpart) [163] and C5.0 [165]. Rpart uses what is called a surrogate split [164], which is basically an estimate of the missing values using other independent variables. Rpart utilises a surrogate variable (or a number of surrogates in order) within a node if the variable for the next split includes missing values (for an explanation of the procedure see [163]). C5.0 handles missing values in the construction stage in the following way: basically, instances with missing values are discounted while calculating the entropy or the information gain for a particular attribute x. Information gain is then multiplied by the fraction of instances for which the value of x is missing. Accordingly, if x is missing for a large fraction of instances, the information gained by testing x at a node will be fairly small. Quinlan [165] provided a detailed example of how missing values might affect the process of tree construction, and also how data with missing values may obtain a classification.

Other algorithms such as Random Forest and SVM may require pre-processing of the data to deal with missing values. One pre-processing approach is to use imputation of missing data. An imputation is the process of filling in missing data with substituted values by ascribing them to other available data. Hair et al. [210] defined imputation as "the process of estimating missing data of an observation based on valid values of other variables." Dempster and Rubin in [211] noted, "imputation is a general and flexible method for handling missing data problems, but is not without its pitfalls. Caution should be used when employing imputation methods as they can generate substantial biases between real and imputed data." However, imputation techniques tend to receive some praise for handling missing data. Several case studies have been published regarding the practice of imputation in survey research [212] and medicine [213, 214].

There are a number of imputation methods [68]. One of the best known is multiple imputation, first proposed by Rubin [215]. It uses a suitable model that includes random variations

to impute multiple accepted values for each missing data point, rather than a single value, which as a result will take into consideration the uncertainty caused by the imputation. Multiple imputation incorporates a number of imputation techniques into a single procedure [68]. This technique is time consuming, as the researcher must generate the multiple datasets, test the models for each dataset individually, and then pool the models into one summary model. However, some researchers [216] argue that sometimes those efforts are worthwhile, and even required, to prevent biased results. The use of multiple imputation has increased greatly in the last decade, and the techniques are now implemented in different freeware and commercial software packages. Buuren and Oudshoorn [217] noted that there is a universal misunderstanding regarding multiple imputation, since many researchers think it is limited to data missing at random (MCAR, MAR). Although it is indeed true that imputation techniques usually assume the data is missing at random, the theory of multiple imputation is general and can be applied to data not missing at random (MNAR). Many sources give additional details on multiple imputation (Allison, [209]; Enders, [218]; Rubin, [215, 219]; Schafer & Olsen, [220]; Schafer, [221]; Sinharry, Stern, & Russell, [222]).

In our study, we use three imputation methods. The first is a single imputation method based on Random Forests and implemented in the package mice in R [217]. This uses Breiman's Random Forest algorithm [168] to produce a non-parametric imputation of values. It works by constructing a Random Forest model for each attribute. Then it utilises the model to predict missing values in the attributes with the support of observed values. For binary or nominal predictors, the imputed value is the category with the greatest average proximity. For numeric-scale predictors, the imputed value is the weighted average of the non-missing observations, where the weights are the proximities. It also produces an estimation of Out Of Bag (OBB) imputation error. OBB is a measure that provides unbiased estimation of the classification error while trees are added to the forest and provides an estimation of variable significance. This happens internally during running of the methods. Random Forest imputation can offer a high level of control on the imputation process. It can return OOB individually (for each attribute) instead of combining the whole data matrix [168]. Random Forest is an ensemble method, thus its imputation will result in one imputed dataset [168].

The second and third imputations follow on from the multiple imputation method suggested by Rubin [223]. For the second imputation, we consider Predictive Mean Matching (PMM). PMM is a semi-parametric imputation approach which fills each missing value with a value randomly "borrowed" from among the observed values with real data. The implementation we used is part of the mice package in R and follows the chained equations approach [217]. In this method imputed attributes accept the value of one of a set of closest observed values in

the dataset, where the closeness is assessed by regression. PMM is suitable for different types of real data, and also suitable for imputing quantitative attributes that are not normally distributed [224].

For the third imputation method, we used R package Amelia II [225]. This package obtains the imputation of missing values by utilising the well-known expectation maximisation technique on several bootstrapped samples of the initial incomplete data to estimate parameters. Then, the algorithm obtains imputed values from the bootstrapped parameters.

We note that there are other imputation methods in the literature, [226] but either they were not applicable to our datasets, or other studies have shown that the methods we selected outperformed them.

4.4.1 A novel method for multiple imputation with an ensemble of classification/regression algorithms

To provide good classification/regression results in the context of large amounts of missing data, as described in the previous chapter for the problem of module performance prediction, we propose an ensemble to combine the results of multiply imputed datasets, as shown in Figure 4.2. For the multiple imputation method, we created five individual imputed datasets, as this is considered sufficient to provide adequate results [223]. Then, we applied the DM algorithms (classification/regression) to each imputed dataset. Next, we used an ensemble of the five training models obtained from the imputed datasets to predict the student's mark. For each cross-validation, we predicted five test sets using the ensemble model, and compute the average RMSE for regression experiments and the average accuracy for classification experiments. Each test set was associated with one of the five imputed datasets and was not used during training the model. The ensemble model used is based on majority voting (for classification methods) or averaging (for regression methods) from the five base classifiers. We believe that this novel combination of multiple imputation and ensembles to produce the final prediction will result in improved models, even when missing data is extensive.

4.5 Software/Other Tools

In this section, we outline the eight main software tools that we have utilised to complete our research:

1. We used **Microsoft SQL Server Management Studio** [2] software to clean and prepare the data provided by the UEA's Business Intelligence Unit. We also utilise it to select the required records for Chapter 5 and Chapter 6.
2. We used **IBM SPSS Modeler 15.0** [3] to complete our work in Chapter 5.
3. We used **RStudio**[4] to perform Chapter 6's experiments.
4. We used **MATLAB**[5] to perform the Friedman statistical test and produce the critical difference diagrams.
5. We used **iTalk Recorder**[6] to record all the conducted interviews.
6. We used **NVivo 11**[7] to complete our work in Chapter 7.
7. We also used **SurveyMonkey Website's tool**[8] to conduct the questionnaire survey in Chapter 7.
8. We used **G* Power software**[9] to calculate the minimum sample size of the questionnaire survey in Chapter 7.

4.6 Management study

Chapter 7 is an explorative, qualitative study of the acceptability of, and issues involved with, providing individual-level information on predicted academic performance to university students choosing modules for the final year of their degree. The main purpose of the study was to explore the thoughts and feelings of students towards providing a predicted outcome for elective modules. In particular, the study questions how this knowledge may alter students' module choice decisions. The study also seeks to understand the attitudes of academic staff and university managers at various levels, towards the implementation of such a personalised enrolment recommender system. The study is associated with one UK university due to the complexity of the ethical considerations of using student data in the UK (see e.g., the Data Protection Act [227]).

The research is primarily qualitative. Qualitative research has been defined in several ways. Strauss and Corbin [155, p.10-11], for example, have defined qualitative research as:

[2] https://www.microsoft.com/en-gb/
[3] https://www.ibm.com/uk-en/marketplace/spss-modeler
[4] https://www.rstudio.com/
[5] https://uk.mathworks.com/products/matlab.html
[6] https://griffintechnology.com/us/italk-premium
[7] http://www.qsrinternational.com/nvivo/nvivo-products
[8] https://www.surveymonkey.com/
[9] http://www.gpower.hhu.de/

> *Any type of research that produces findings not arrived at by statistical procedures or other means of quantification. It can refer to research about persons' lives, lived experiences, behaviours, emotions, and feelings as well as about organisational functioning, social movements, and cultural phenomena, and interaction between nations.*

Further, they argue that the best utilisation of qualitative research is when the methods are: supportive to the personal experiences and preferences of the scholar, compatible with the nature of the research problem, and applied to investigate areas about which little is known. Lapan et al. [228] argue that qualitative studies concentrate on giving voice to those whose lived experiences cannot be acknowledge directly by others, and asking research questions that motivate insight and reflection rather than quantitative measures such as measuring test performance. Miles and Huberman [154] have shown that qualitative research has many roles including verifying previous research on a topic, giving in-depth detail about known information or a topic, gaining a different perception or a different way of viewing something, and enhancing the scope of existing research. Because this study investigates a seldom researched area – student perceptions to enhance and extend our understanding of module choice – qualitative methods were suitable for this study.

Many researchers (e.g., Strauss and Corbin [155]; Patton [229]) have argued that quantitative and qualitative research can be effectively combined in the same study. For instance, Russek and Weinberg [230] stated that by using both qualitative and quantitative data, their study of technology-based materials for the elementary classroom provided perceptions that could not be reached by using only one of the methods. Quantitative and qualitative methods can be combined in a number of ways, operating sequentially or in parallel. In this study, we have used both methods to collect our study's data, using a quantitative method (questionnaire survey) of students to establish some initial parameters and to recruit subjects for the qualitative research before undertaking qualitative interviews with a subset of respondents. Multiple methods of data collection, analyses, or theories aid as an approach to ensure the validity of the qualitative data and show trustworthiness. The process of checking is called triangulation [231]. Triangulation [232] is the confirmation of results with alternative sources of data.

We should note that there are two well-known qualitative methods: focus groups and interview methods [233]. We chose the interview method over the focus group approach for several relevant reasons that have been discussed in [234]. For example, if the participants are uncomfortable with each other, they will not discuss their views and opinions openly. Also, listening to contributors' perceptions creates expectations for the outcome of the research that cannot be achieved. Therefore, we believe the interview method is more suitable for our study.

There are three main types of research interviews: structured, semi-structured and unstructured [235].

- A structured interview is a series of predetermined questions, with minimum or no variation and with no scope for follow-up questions that warrant additional elaboration. This approach is easy and quick, however it is insufficient if "depth" is required.
- An unstructured interview, the opposite of the structured interview, does not reflect any predefined ideas or theories, and is done with little or no organisation. For example, it starts with an open question, then progresses based on the initial response. This approach is very time consuming and it is difficult to manage. Also, it could be confusing and unhelpful for the participants, due to the lack of predefined questions. It is useful when there is nothing known about the subject area or 'significant' depth is required.
- A semi-structured interview contains several main questions that help to define the areas to be investigated; however, interviewees are also allowed to diverge in order to pursue their response in more detail. This approach is more flexible compared to the structured approach, but still includes some guidance on what to talk about. Therefore, we utilised this approach in our study.

The following are **the sampling procedures:**

1. At first, we started by conducting a questionnaire survey. This is a quantitative approach defined as a research tool that includes a sequence of questions for the purpose of collecting information from respondents [236]). We built the questionnaire survey using the SurveyMonkey website's tools. We used the survey to grasp an initial understanding of the undergraduate students' views on having a future enrolment recommender system. We also used the surveys to recruit students for the following interviews (qualitative approach).

   The questionnaire survey was piloted on three students and minor revisions were made. A link to the survey was sent via an email to the current year 2 and year 3 undergraduate students from the two schools. Participation was optional. The questionnaire survey did not collect any identifiable information and, therefore, the responses were anonymised. The survey began with a consent statement, then the survey questions, which were divided into three parts: six demographic questions; four questions about how students currently choose their optional modules; and seven questions about their thoughts of having individual-level information while choosing their optional modules, as shown in Appendix F. Finally, we asked if students would be willing to be interviewed face-to-face and if so, could they provide us with their

preferred email address, as shown in the last page of Appendix F. The survey included a gift incentive (a £50 Amazon voucher Prize Draw). In terms of the sample size of the questionnaire survey, we consider two types of error when calculating the minimum acceptable sample size [237]. Type1 or α errors appear when rejecting a true null hypothesis and type2 or β errors appear when a false null hypothesis is not rejected. The probability of these errors occurring can be diminished by increasing the sample size. By convention, $(1 - \beta)$ is set to 0.9 or 10% for missing an association and α is set to 0.05 for a 95% confidence level [237]. The effect size indicates the importance of the relation between the predictor and outcome variables. Cohen [238] explains three different effect sizes: small (d=0.2), medium (d=0.5) and large (d=0.8). In exploratory research, effect size is normally set as large. In this work G* Power software was utilised to calculate the lowest sample size required which was 45. The calculation was achieved for a t-test to discover the difference in mean from constant.

2. Drawing on the respondents to the questionnaire, we conducted 28 face-to-face semi-structured interviews with students to gain an in-depth understanding of individuals' perceptions of module choice and the anticipated impact of personalised mark predictions on such choices. Qualitative methodology is often concerned with the multiple interpretations and meanings that participants give to a situation and it emphasises using the participants' own words [239, 240]. The focus of the analysis in qualitative research is the utilisation of the participants' voices. Most importantly, it is the voice of the participants that allows the scholar to study the phenomenon of interest [241]. Hence, theoretical or purposive sampling is often utilised in qualitative research to concentrate on the views of those who are known to experience the phenomenon of interest and may have something interesting to say about it.

   Instead of being focused on the ability to generalise the individual's experiences to a larger population, an in-depth exploration of their experience is the goal [232]. Therefore, the number of desired interviewees was determined in an effort to have participants with a variety of experiences (a variation sample). We continued to interview students until we reached the point where we were no longer learning new things from the interviews, as suggested by [155] and [154]. Each student interview lasted from 15 to 20 minutes. The participants varied as shown in Table 4.1. Each student signed a consent form before starting the interview. We piloted the interview with two students (it should be noted that all our piloted candidates were from the same study sample). Participants were given a £10 Amazon voucher gift incentive in an effort to thank them for their participation.

Table 4.1: The variation of the interviewed students by gender and school. CMP stands for Computing Sciences and NBS stands for Norwich Business School.

	CMP School	NBS School
Male	12	2
Female	7	7

3. We also interviewed selected members of staff from several management and teaching roles in the university concerned with undergraduate education. The interviews lasted from 35 minutes to an hour. We piloted the interview with one staff member. The staff's specific management roles have been anonymised in Chapter 7 due to ethical considerations; however, they ranged from course and programme leaders to a member of the university's senior management team. Staff also signed a consent form before the interview.
4. We should note all the piloted interviewees were from the same sample of our research's participants.
5. To make sure that the interviews proceeded without disruption, we secured and verified with the interviewees a meeting place that had acceptable space and the required equipment. For example, we installed an iTalk recorder application on both the laptop and smart phone of the researcher, and we had laptop and phone charger and microphone sets in place. Also, we had a 'Do not disturb' note on the door. We confirmed all these arrangements the day prior to the interview.
6. Once interview data was collected, audiotapes were transcribed to a written form to enable the qualitative analysis process. The written transcriptions were moved to the Nvivo software tool, for the purpose of thematically coding and analysing them. A thematic coding is a type of qualitative analysis that includes noting or identifying segments of text or images that are connected by a mutual theme or idea, permitting you to index the text into categories and hence develop a 'framework of thematic ideas about it' [242, p.38]. Using the computer software Nvivo:
   (a) We carefully read all the transcriptions and noted down ideas and themes that occurred to us and were relevant to our research question.
   (b) Then, we went through the documented ideas, and considered the underlying implications rather than the substance for the purpose of creating a list of categories (topics).

(c) We grouped together the related categories and arranged these groups into major categories, unique categories or leftovers. We also associated each set of categories with the appropriate theme.

(d) Next, we used this list to code our transcriptions by assigning passages of text to the appropriate category and theme. Through the coding process, we organised our categories by adding new, more appropriate categories or by merging existing ones that were related to each other.

(e) We recoded our existing data, when needed.

Weber [243, p.12] notes that, in order to 'make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way.' We increased the reliability of our analysis, by having two researchers fully code four interview transcripts. Coding was then compared and disagreements were discussed. Two researchers also discussed the derived categories and the number of categories, using questions such as the following: 'Do we agree on the same categories/theme or should we add new more appropriate categories or emerge existing ones that are related to each other?' or 'Do we agree to the same number of categories/ themes?' Subsequent modifications were made to the categories/themes and re-coding was done for the four transcriptions. The remaining transcripts were coded by the first author.

A final source of data was the written guidance given to students when choosing modules. These written guidance documents were provided during the staff interviews.

Our role as researchers, specifically ones who use a qualitative methodology, was complex and challenging. It started with the determination of a meaningful topic, putting together an appropriate research question, and creating a comprehensive research plan. We were accountable for promoting objectivity in our study, taking responsibility for diminishing any personal biases that we might have. For instance, we did not (consciously) try to influence or force interviewees' responses towards our opinions. In an effort to explain our ideas and preferences, we included our personal beliefs in the discussion as they were related to the overall subject of interest, but we made sure that the interviewees felt able to disagree.

4.7 Ethical Considerations

Ethics corresponds to the correctness of the scholar's behaviour towards the study participants. There are ethical and moral principles that direct our manner and rapport with

others. These are associated with how we carry out our research from the start to the end and incorporate research design, data collection, data analysis and data interpretations in morally acceptable ways [244].

In this research study, we have analysed student data available from the data warehouse of the UEA, in addition to the data that has been collected from the conducted interviews and questionnaire survey. If students sense their privacy being breached, they might be hesitant to allow their data to be utilised for analysis and research. In some situations, it is not obvious who owns the data. The data may belong to an individual, an educational institution, or even to an outside vendor who has possession of a data collection tool. Greller and Drachsler in [245] and Ferguson [246] explained that in contrast to the traditional approaches of obtaining data in research settings, there is no definitive framework for scholars to follow when obtaining consent to utilise data, nor are there any widely accepted guidelines for the anonymity of data. Jacqueleen in [247] explained that ethical guidelines should ensure a clear definition of the ownership and stewardship of data and should take privacy concerns into consideration to prevent data abuse.

Each British educational institution has a clear code of ethical practice for research. Therefore, to address the ethical considerations for our research study, we followed the required ethical procedures of the UEA. We obtained three ethical approvals from the School of Computing Sciences Ethics Committee during different phases of our research study. The first ethical approval was related to utilising the student data at the university data warehouse. The second and third approvals were associated with conducting the questionnaire survey and the interviews. Each time the Ethics Committee required the ethic checklist form as shown in Appendix D.

In addition, for the first approval, they requested a data management plan and the project synopsis as shown in Appendix E for using the student data.

For the management study, they required the interviews questions, questionnaire survey questions, a copy of the prize draw question, and a document that consisted of:

- The project synopsis.
- The research protocol including appropriateness of methods, sample size, gaining informed consent, informing participants of their option to withdraw and the risk or benefits of participating.
- Information on the data, including ensuring data confidentiality and anonymity, restricting data access, restricting data use, and informing subjects of ethical issues.

The documents for the ethical approval of the questionnaire surveys are shown in Appendix F. Those for the interviews are shown in Appendix G.

4.8 Summary

In this chapter we explain the connection between the different parts of the thesis, including the management aspect, which in our context relates to how to utilise the derived knowledge from the predictive models. We present a summary of our research design. We describe thoroughly the methods used. Finally, we discuss the ethical considerations of our research study.

	Research Focus and Contributions	Research Design	Chapter
A Technical Aspect	Competitive predictive models for student overall outcomes from routinely collected data, highlighting features associated with poor performance.	✓ Initial exploration of the available data to present a clear understanding of the utilised attributes. ✓ Three experiments were conducted using different types of classification models (logistic regression, Neural Network, Decision List, Bayesian Network, Discriminant analysis, C5, C&R Tree, Quest and CHAID) and three feature sets: o General student demographics and general performance attributes (information available at registration). o Adding Year1 performance attributes. o Adding Year 2 and Year 3 performance attributes. ✓ The set of experiments used real data associated with two different schools of study.	Chapter 5
	Predictive models for module outcomes, innovative for their use of multiple imputation combined with an ensemble to handle missing data.	✓ A regression and classification set of experiments were conducted, using o leading data mining methods, (clustering, C5, SVM, Random Forest, Rpart, in addition to Simple Average prediction system), with ensemble approach; o Three applicable imputation methods: Random Forest; multiple imputation with expectation maximisation (EM); and multiple imputation with chained equations; in addition to without imputation. ✓ The set of experiments used real data from two different schools of study associated with one university, and two public data sets associated with different education institution.	Chapter 6
A Management Aspect	Management study of how to utilise any knowledge derived from the exercise in the educational context both from the point of view of the students and the institutions.	✓ A questionnaire survey was conducted o to grasp understanding of undergraduate student views; o to recruit student for following interviews. ✓ Interviews were conducted with students to gain an in-depth understanding of views of module choice and the anticipated impact of personalised mark predictions on such choices. ✓ Interviews were conducted with selected staff members to also understand perceptions of o implementing a future module enrolment system to improve student outcomes; o how to utilise the knowledge derived from Chapter 5 to improve student outcomes.	Chapter 7

Figure 4.1: Summary of Thesis Chapters.

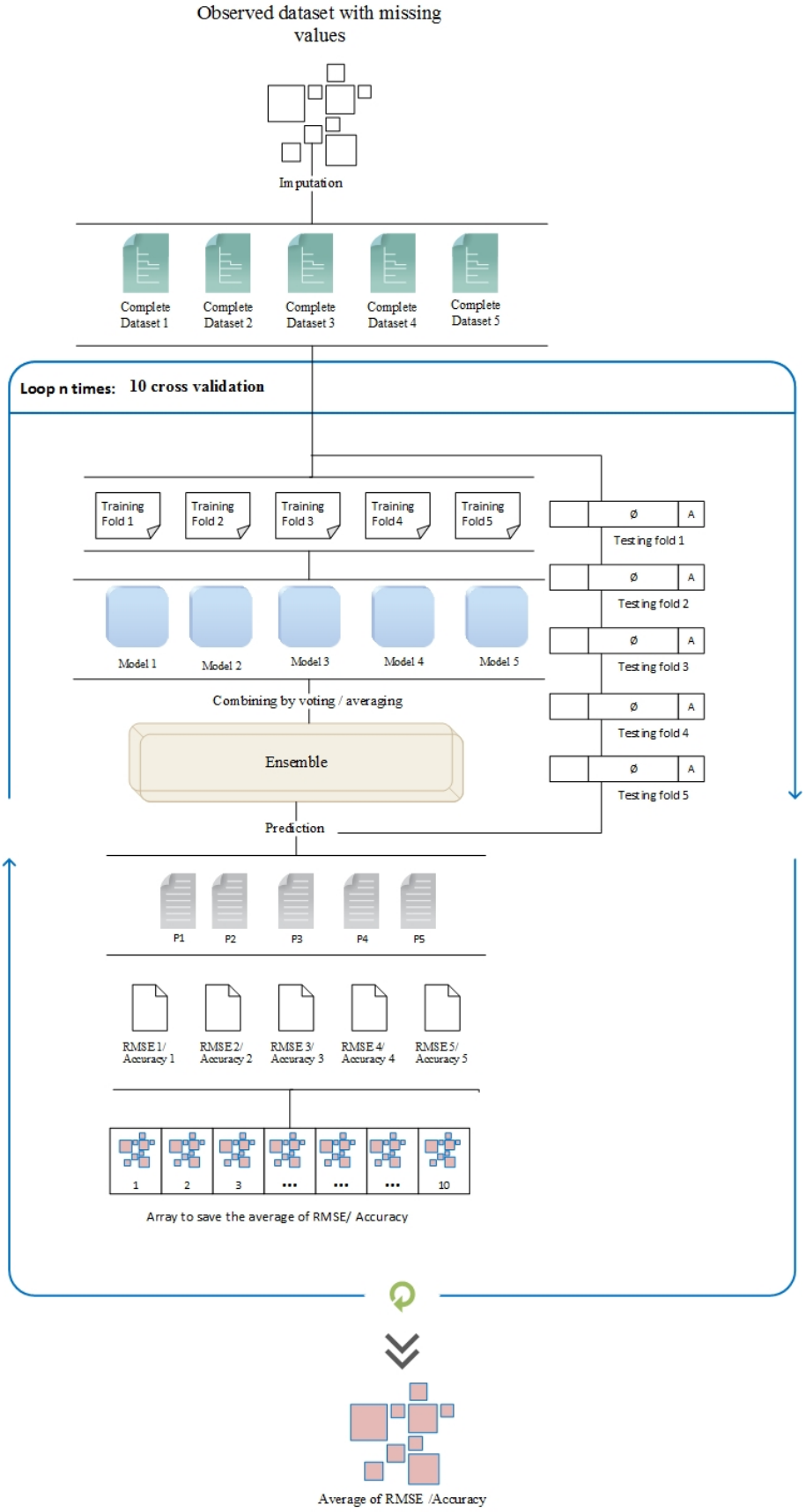


Figure 4.2: A novel approach for multiple imputation with an ensemble of classification/regression algorithms.

Chapter 5

A general model to predict students at risk of obtaining poor outcomes

5.1 Introduction

Nowadays, many UK universities have a specific targets for students achieving good honour degrees. Achievement in terms of good honours is often reported in league tables. For example, the Complete University guide [248] reports good honours as "the percentage of graduates achieving a first or upper second class honours degree". On the other hand, the Guardian League Tables utilises a value-added score that compares students individual degree results with their entry qualifications, to show how effective the teaching is [249]. It is also important for students to achieve a good degree as this can impact on their employment prospects [250]. It is therefore in the interest of both students and Universities to identify students at risk of not obtaining a good honours degree so that early intervention may improve their outcome.

In this chapter we attempt to use data mining techniques to predict student outcomes based on early module performance and other student characteristics. If our methods are successful for predicting the more general problem of student good honours performance, we can then produce more granular predictions at the module level. We hope to uncover early indicators of poor performance that may be used to target remedial action for the concerned students. We aim to investigate the available features that may be used for prediction, as

well as the type of classifiers that may produce the best results.

As we mentioned previously in Chapter 2, EDM is now an established field and, as such, a number of reviews have been published, e.g. [38, 30, 39]. In particular, Peña-Ayala [30] cover a number of work on students performance using data-mining which would be important to this exercise. We reviewed some of that work earlier in Chapter 2 and apply best practice to our own problem.

The rest of this chapter is organized as follows: Section 5.2 describes the purpose of this chapter. Section 5.3 describes aspects that will help in improving students performance. Section 5.4 explains briefly the sequence of experiments and the produced results. The discussion of the result is contained in section 5.5. Lastly, section 5.6 summarises this chapter.

5.2 Purpose of this chapter

The aim of the work we present in this chapter, as with some of the studies reviewed (earlier in Chapter 2), is to identify weak students as early as possible, i.e. those that would end up with poor outcomes. We define good performance in terms of "Good Honours" versus "Not Good Honours" (binary) outcomes because this is currently a measure generally used in the UK and universities are interested in improving their good honour rates. The main aim is to highlight as early as possible (i.e. in year 1) groups of students that may be at risk so that targeted interventions can be proposed to improve their outcomes. Given the variety of models used in the literature with varying degrees of success and the fact that no model has emerged as the overall best, we use a number of classifiers and combine them using ensembles to establish the best possible model. Given also the literature's variation on the features to be included, we include a number of feature sets: first we attempt classification with a feature set which uses only information available at registration, then we add performance on year 1. Furthermore, we take into consideration the difficulty of each module by comparing the performance of each student with their peer group, as some studies suggested (the used data explained previously in Section 3.1). We also initially included attributes on engagement as others studies have suggested, that are only now becoming available (e.g. engagement with library services and attendance monitoring information). However, as we explained earlier in Section 3.1 that due to data quality issue we did not utilise the engagement attributes. After that, we look at combinations of module choices in years 2 and 3 in relation to outcomes, and investigate if this is will improve our results. Our approach aims to provide further evidence of best feature sets and models for classification.

5.3 Long term study objectives

In this section, we consider how to provide students with an appropriate intervention that may improve their overall performance. This would be the ultimate aim of this preliminary exercise.

The most significant aspect is to identify weak students that may be at risk of graduating with a lower class or abandoning their studies. Students at a high risk need particular attention and support with managing their studies if they are to graduate with higher grades. In this sense, it is important to select the attributes that closely represent the chief characteristics of the students at risk; this may include achievement in specific modules as well as personal characteristics. Some personal characteristics may suggest specific strategies. For example, if non-native or overseas students are more often associated with poor outcomes, an intervention based on additional language support may prove fruitful.

We may also identify modules that are associated with good outcomes and bad outcomes given a student profile, so that when module choices are available those modules can be suggested or discouraged respectively for students with similar characteristics and academic achievement records. The intervention in this case may be a future enrolment recommender system which takes account of similar students' trajectories and achievements to recommend what may be best choices for a particular student.

We can also examine the measure of the dependencies or associations between modules. This may alert us to potential problems on related modules once a particular module is associated with a bad outcome. For example, some remedial sessions on a failed module may help students conquer related modules more successfully.

Hence, in this chapter we begin our work by predicting overall good honours outcomes based on generic students' characteristics, on first year performance and on non-mandatory second and third year module performances to inform strategies for intervention. The next step, included in the next chapter (chapter 6), is to explore further the association between individual modules and outcomes that might assist in creating algorithms for a fully fledged future system that leads towards an improvement in good honours rates and perhaps also increased student satisfaction.

5.4 Experiments and Results

The initial analysis of all the data for the 9 years span in both the first dataset and second dataset showed an overall Good Honours rate of 56% and 63.3% respectively. The overall

GH rates for the first dataset are given in Table 5.1 and are divided by fee status. The trend of GH over the years is shown in figure 5.1 and is also divided by fee status into H, EU and OS students. It shows that attainment is worse for OS students with some narrowing of the gap over the years. The number of OS students has grown steadily as a proportion of the total. The number of EU students is low and hence their attainment level cannot be meaningfully assessed but is closer to that of the H students than to the OS students.

Table 5.1: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the first dataset

Status	GH	NGH	Total	GH Rate
Home	433	302	735	58.9%
European	26	18	44	59.1%
Overseas	44	75	119	36.9%
Total	503	395	898	56%

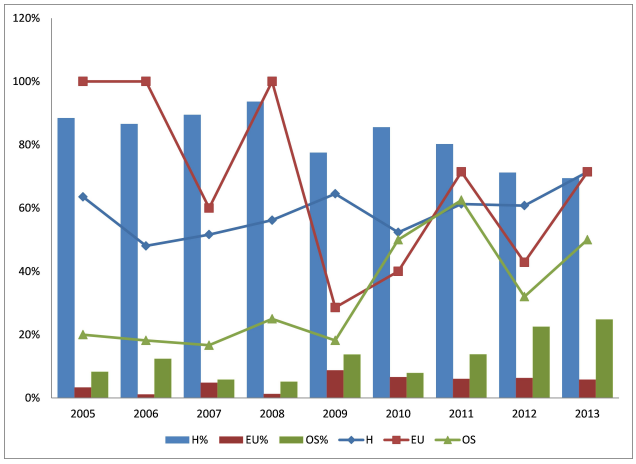


Figure 5.1: GH Rate for the first dataset. The bars show the percentage of students in each fee category (Home, EU and Overseas) whereas the corresponding line charts show the percentage of GH degrees obtained for each category of students.

We explored the relation between the students' Maths or English entry qualifications and the GH outcome. As shown in Table 5.2 and Table 5.3, we found that students with Maths or English entry qualifications are associated with higher GH rates than students without. We tested for statistically significant differences (p-value < 0.05) between the two proportions in each table using a Chi-square test and found there was a statistically significant difference ($x^2 = 18.2929$, p-value $= .000019$) between students with/without maths entry qualifications and GH outcomes, but there was no statistically significant difference ($x^2 = 0.94$, p-value $= 0.33$) between students with/without English qualifications and GH outcome.

Table 5.2: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the first dataset for those with/without Maths entry qualifications

Maths' Qualification	GH	NGH	Total	GH Rate
Yes	156	73	229	68.12%
No	347	322	669	51.86%

Table 5.3: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the first dataset for those with/without English entry qualifications

English Qualification	GH	NGH	Total	GH Rate
Yes	46	29	75	61.33%
No	457	366	823	55.52%

The outcome of the initial exploration for the second dataset was very similar to the outcome of the first dataset and is presented in Tables 5.4, 5.5, 5.6 and figure 5.2. The attainment levels are also better in this second school for H than OS students. There is insufficient data for EU students to consider the trends in the same way. The percentage of OS students has also increased over time in this second school and the performance of both H and OS students has improved over time, although the gap remains large between both groups. The students with Maths or English entry qualifications are also associated with higher GH rates than students without Maths or English qualifications. The results of Chi-square test show that there is a statistically significant difference between students with/without maths entry qualifications and GH outcome (x^2 = 32.4144, p-value = (< 0.00001), and also between Students with/without English qualifications and GH outcome ($x^2 = 27.9143$, p-value = (< 0.00001).

Table 5.4: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the second dataset

Status	GH	NGH	Total	GH Rate
Home	931	370	1301	71.6%
European	39	32	71	55%
Overseas	163	254	417	39.1%
Total	1133	656	1789	63.3%

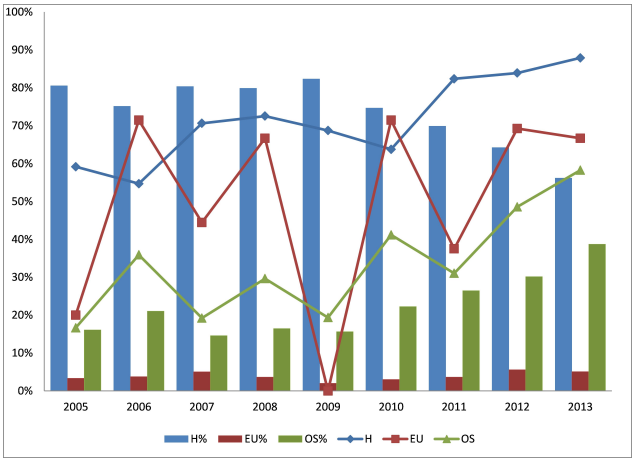


Figure 5.2: GH Rate for the second dataset. The bars show the percentage of students in each fee category (Home, EU and Overseas) whereas the corresponding line charts show the percentage of GH degrees obtained for each category of students.

Table 5.5: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the second dataset for those with/without Maths entry qualifications

Maths' Qualification	GH	NGH	Total	GH Rate
Yes	301	98	399	75.43%
No	832	558	1390	59.86%

Table 5.6: Percentage of Good Honour(GH)/ Not Good Honour(NGH) students in the second dataset for those with/without English entry qualifications

English Qualification	GH	NGH	Total	GH Rate
Yes	188	51	239	78.66%
No	945	605	1550	60.96%

5.4.1 First Experiment: student demographics feature set

In the first phase of the experiment we used the attributes that related to student demographics and general performance, i.e. the first group of attributes, but not the attributes for specific modules (as shown in Table 3.1). We also discounted the year averages as they will be clearly related to the outcome.

After initial exploration, we discounted the attributes that related to the library loans

after taking into consideration that some students prefer to use the required books during the hours they spent in the library without borrowing them, and also most of the students nowadays can access most of the library resources on-line and this information is not recorded in their records. Therefore it was considered that these attributes do not provide us with a real value of engagement.

Assessing each attribute independently using a Feature Selection ranking algorithm based on a Pearson chi-square test with significance level of 0.05 found that fee status, course, nationality, widening participation indicator, maths entry qualifications and gender were all statistically significant ($p<0.05$). According to this preliminary analysis OS students have statistically significant worse outcomes. Some specific courses offered by the first school also had statistically significant worse outcomes than others. The results for nationalities which presents higher granularity than fee status cannot be taken into consideration since some countries have very low numbers of students attending which invalidate the results of the chi-square test. However, there were specific countries with substantial number of students that did have statistically significant lower levels of attainment and may be of concern. The widening participation attribute relates to the participation in higher education of students in a postcode relative to the HE population as a whole. The students are classed as belonging to groups 1 to 5, Non-UK or not known. A classification into the lower groups implies that the student lives in a postcode of low participation. The lowest group is associated with lower attainment. Since widening participation also reflects OS students as a rather large group (over 17% of the students belong to it) they also show lower attainment. There is little difference in the other groups in terms of attainment. In terms of gender, females have statistically significant higher levels of attainment than males. Students with maths entry qualifications also have statistically significant higher level of attainment than students without maths entry qualifications.

Next, we used a combination of classification models to predict GH/NGH. For this we used the software IBM SPSS Modeler v 15, a well known data mining tool-kit. We used an autoclassifier which engages 9 different types of classification models and automatically selects those that perform best on the training data. The models that were tried were: logistic regression; Neural Network; Decision List; Bayesian Network; Discriminant analysis and four decision tree algorithms: C5, C&R Tree, Quest and CHAID. All algorithms used default parameters. Those selected for the first data set were a Bayesian Network, a C5 decision tree algorithm and Logistic Regression. They were combined using an ensemble with confidence-weighted voting. Our ensemble model had an accuracy of over 66.16% on training data (over 60% on a test sample containing 20% of the original data). The gain chart for the ensemble model versus the selected independent models is shown in

Figure 5.3a. The accuracy of individual models was similar to the accuracy of the ensemble. It is possible using the ensemble, to chose those records which are predicted to correspond to NGH students with high prediction probability. This strategy would enable us to select the students most likely to gain NGH, so that interventions could be put in place to help them early on. Using a threshold probability of 0.5 as given by the ensemble model, we were able to select 227 students with a GH rate of 32.6%, considerably lower than the overall population. That group captured 153 or 38.7% of the NGH students. More precisely, 114 of the NGH group were predicted as a 2:2 class degree and the other 39 students were predicted as a lower class degree, for example 3rd or PASS class degree. Lowering the probability of the ensemble prediction to 0.3 captured a group of 280 students representing a GH rate of 35.7% still substantially lower than that of the overall population. The later threshold captured 180 or 45.6% of the NGH students (precisely, 135 of NGH group as a 2:2 class degree and 45 students as a 3rd/PASS class degree). If an intervention could change the outcome for a majority of those students from NGH to GH, it could substantially improve the overall GH rate. Note that students who obtained a 2:2 class degree, the larger group, should require less effort to help them achieve GH degree than students who obtained a lower class degree. However, in the interest of fairness the intervention should be directed to all students at risk of poor outcomes. It is plausible to think that an intervention may also be beneficial for the students that may be captured by this approach but who would have got GH degrees in the first place, i.e. the false positives (32.6 or 35.7 % of students in each scenario) as it would enable them to achieve even better outcomes. The four attributes used in all models were course, gender, fee status and math entry qualification. Two of the models used an additional attributes: widening participation, age band, English entry qualification, foundation year, and disability.

Then for validation purposes, we applied the above series of steps on the second dataset. We began by using only the attributes that related to student demographics. The assessment of each independent attribute using a Feature Selection ranking algorithm based on a Person chi-square test with statistically significant level of 0.05, showed that nationality, widening participation indicator, fee status, gender, maths entry qualification, and English entry qualification were all statistically significant ($p<0.05$). OS students in this dataset also had statistically significant worse outcomes. The results for nationalities will not be taken into consideration for the same reason mentioned in the first dataset. However, the same specific countries as for the first dataset had statistically significant lower levels of attainment. The assessment of the widening participation attribute has shown that the lowest group of students is associated with lower attainment. The OS group within the widening participation attribute included over 26% of students, and was also associated with lower

attainment. There is little difference in other groups in terms of attainment. Students with maths and/or English entry qualifications also have statistically significant higher level of attainment than students without maths entry qualifications. In term of gender, females also have statistically significant higher level of attainment than males. However, the difference in the attribute assessment in this dataset compared to the first data set was that the course attribute was not as relevant. The reason for this is that 63% of the undergraduate students enrolled on same course, hence one of the courses that this school offers has a much higher number of students compared to the other courses.

Next, we used a combination of classification models to predict GH/NGH for the second dataset again using the software IBM SPSS Modeler v 15. The autoclassfier selected C & R Tree, Neural Network and logistic regression classification models for the second data set. They were combined using an ensemble with confidence-weighted voting. Our ensemble model had an accuracy of over 70.78% on training data (over 65.05% on a test sample containing 20% of the original data). The gain chart is shown in Figure 5.4a. It is also possible using this ensemble, to chose those records which are predicted to correspond to NGH students using high prediction probability. Using a threshold probability of 0.5 as given by the ensemble model, we were able to select 206 students with a GH rate of 29.13%, considerably lower than the overall population. That group captured 146 or 22.26% of the NGH students(113 obtained a 2:2 class degree and 33 had a 3rd/PASS class degree). Lowering the probability of the ensemble prediction to 0.3 captured a group of 320 students representing a GH rate of 32.5% still substantially lower than that of the overall population. The later threshold captured 216 or 32.9% of the NGH students(164 of that group had a 2:2 class degree and 52 students had a 3rd/PASS class degree). The three classifiers used 6 attributes: widening participation, fee status, gender, English entry qualifications, maths entry qualifications, and age-band at entry. Two classifiers the Logistic Regression and the Neural Networks used two additional attributes course name and disability. Table 5.7 summarises the results of the First Experiment.

5.4.2 Second Experiment: adding Year 1 performance

After identifying the students that were at high risk of failing to earn a GH award class using only the first group of attributes, a second experiment considered the influence of performance on the year 1 modules on the classification. Our first dataset contained information on students enrolled on 12 different courses. Although most year 1 modules are compulsory and many of them are shared between different courses, there were 12 different modules that we needed to consider to account for all the variations. For each of those 12 modules, we

Table 5.7: This table summarises the results of the First Experiment.

	first dataset	second dataset
The selected models	Bayesian Network, C5, and Logistic Regression	C&R Tree, Neural Network, and Logistic Regression
Accuracy of the ensemble model	60% on test sample (20% of the original data)	65.05% on test sample (20% of the original data)
Using a threshold probability of 0.5	We were able to select 227 students with a GH rate of 32.6%. That group captured 153 or 38.7% of the NGH students (114 of 2:2 class degree and 39 students of 3rd or PASS class degree).	We were able to select 206 students with a GH rate of 29.13%. That group captured 146 or 22.26% of the NGH students (113 obtained a 2:2 class degree and 33 had a 3rd/PASS class degree).
Using a threshold probability of 0.3	We captured a group of 280 students representing a GH rate of 35.7%, 180 or 45.6% of the NGH students (135 of NGH group as a 2:2 class degree and 45 students as a 3rd/PASS class degree).	We captured a group of 320 students representing a GH rate of 32.5%, 216 or 32.9% of the NGH students (164 of a 2:2 class degree and 52 students had a 3rd/PASS class degree).
The used attributes	Course, gender, fee status and math entry qualification. Two of the models used an additional attributes: widening participation, age band, English entry qualification, foundation year, and disability.	Widening participation, fee status, gender, English entry qualifications, maths entry qualifications, and age-band at entry. Two classifiers the Logistic Regression and the Neural Networks used two additional attributes course name and disability.

considered the performance of the students with respect to their peers as defined in Table 3.4 as this could be more indicative than an absolute mark value.

Feature Selection ranking using a chi-square algorithm showed that all of the module performances were important in the classification. Furthermore, an F-test to compare the mean mark of students in the GH and NGH group for each module showed statistically significant differences in the means with students that achieve NGH obtaining statistically significant lower marks on the year 1 modules. Hence poor outcomes seem to be already visible on module performance in year 1. This is an important finding since the year 1 module marks do not contribute to the overall degree classification, but are nevertheless indicative of the expected outcome.

A classification ensemble was built as in the previous experiment, but this time using the year 1 module performance attributes as well as the previous demographic attributes identified by feature selection. The autoclassifier chose a Logistic Regression, C& R Tree and a Decision List as the classifiers and combined them to produce an accuracy over 80.4% on the training data (72.2% on the test sample). This represents a substantial improvement from the previous model. The gain chart in Figure 5.3b shows the evaluation of the model accuracy.

Selecting those that are predicted as NGHs with a probability greater than 0.5, as in the

previous experiment, isolated a group of 355 students with a GH rate of 21.97%. There were 277 or 70.12% NGH students in the group. Specifically the group captures 193 who obtained a 2:2 class degree and 84 who obtained a 3rd class degree. An intervention for this group could be quite effective on the overall GH rate and quite targeted. A final assessment of those in the group showed that they had substantially lower averages for year 1, 2 ,3 and 4, as well as substantially lower averages for all year 1 modules. An F-test showed statistically significance differences ($p<0.05$) for all pairs of averages (in the selected group and all others). The mean values for years 1-3 and for all first year modules are shown for both groups in table 5.8.

Table 5.8: Comparison of means for poor performers as selected by ensemble versus all other students in the first dataset (Second Experiment). Note: * represents statistically significant results.

Attribute	Mean(Poor Performers)	Mean(others)	F-test
COMPUTING FUNDAMENTALS 1	49.38	70.13%	< 0.00001 statistically significant
COMPUTING SYSTEMS 1	52.35	66.91%	.00013*
COMPUTING SYSTEMS 2	41.16	65.53%	< 0.00001*
PROGRAMMING 1	49.02	69.93%	< 0.000013*
PROGRAMMING FOR APPLICATIONS	46.08	62.21%	< 0.000013*
INTRODUCTION TO BUSINESS	49.19	60.06%	< 0.00001*
THE COMPUTING REVOLUTION	65.64	73.83%	.00013*
INTRODUCTION TO FINANCIAL REPORTING	46.09	65.94%	< 0.00001*
ACCOUNTING FOR MANAGEMENT DECISIONS	41.41	57.33%	.00013*
FUNDAMENTALS OF INFORMATION SYSTEMS	47.39	56.32%	.00013*
INFORMATION SYSTEMS AND BUSINESS RESEARCH	45.89	56.74%	< 0.00001*
INTRODUCTION TO ORGANISATIONAL BEHAVIOUR	47.21	60.50%	< 0.00001*
Year1	47.88	65.53%	< 0.000013*
Year2	50.59	62.49%	.00003*
Year3	55.78	65.52%	.00003*

Moreover, we applied the second phase of the experiment on the second data set. The second dataset contained information on students enrolled on 4 different courses. Again for this school, most year 1 modules are compulsory and many of them are shared between different courses. There were 10 different modules that we needed to consider to account for all the variations. For each of those 10 modules, we also considered the performance of the students with respect to their peers.

We found using the Feature Selection based on chi-square algorithm, that all of the module performances of the validated data set were important to the classification except for one module *(FUNDAMENTALS OF INFORMATION SYSTEMS)*. The module that was not statistically significant was taken by a very low number of students (i.e. 4 students) and was therefore discounted from the rest of the analysis. In addition, the F-test showed statistically significant differences in the means between students that achieve GH/NGH. Those that obtained NGH had statistically significant lower marks on year 1 modules, even

though as before, year 1 marks do not count towards degree classification.

Next, a classification ensemble was built using the year 1 module performance attributes as well as the previous demographic attributes identified by feature selection. The autoclassifier chose a Logistic Regression, a Neural Net and CHAID as the classifiers and combined them to produce an accuracy over 77.77% on the training data (76.08% on the test sample). This also represents a substantial improvement from the previous model for the second data set. Figure 5.4b shows substantial gain for the model including year 1 performance attributes with respect to the previous model and to the baseline.

Again by selecting those that are predicted as NGHs with a probability greater than 0.5, we captured a group of 387 students with a GH rate of 21.45%. There were 304 or 46.34% of the NGH students. The NGH captured group included 227 students that obtained 2:2 class degrees; the remaining students in the group obtained a lower class degree. A final assessment of those in the group showed that they had substantially lower averages for year 1, 2 and 3, as well as substantially lower averages for all year 1 modules. Note that in this data set, all students completed their degree within three years, but in the first data set, students may take four years to complete their degree due to year in industry variants. Nevertheless we did not take it into consideration in the mean comparison table 5.8. An F-test showed statistically significant differences ($p<0.05$) for all pairs of averages (in the selected group and all others). The mean values for years 1-3 and for all first year modules are shown for both groups in table 5.9.

Table 5.9: Comparison of means for poor performers as selected by ensemble versus all other students in the Second Dataset (Second Experiment). Note: * represents statistically significant results.

Attribute	Mean(Poor Performers)	Mean(others)	F-test
ACCOUNTING FOR MANAGEMENT DECISIONS	45.58	58.56 %	.00003 statistically significant
DEVELOPING BUSINESS SKILLS	58.08	66.77%	.00003*
ECONOMICS FOR BUSINESS	47.21	56.09%	< 0.00001*
INFORMATION SYSTEMS AND BUSINESS RESEARCH	44.98	54.66%	< 0.00001*
INTRODUCTION TO BUSINESS	48.30	60.53%	< 0.00001*
INTRODUCTION TO FINANCIAL AND MANAGEMENT ACCOUNTING	53.32	59.45%	.003*
INTRODUCTION TO FINANCIAL REPORTING	51.94	64.39%	.000031*
INTRODUCTION TO ORGANISATIONAL BEHAVIOUR	42.87	54.74%	< 0.00001*
PROGRAMMING FOR APPLICATIONS	55.44	59.64%	.005*
Year1	49.03	59.16%	.00003*
Year2	51.42	59.88%	.00003*
Year3	56.16	64.39%	< 0.00001*

5.4.3 Third Experiment: adding Year 2 and Year 3 performance

As further research, we perform a third experiment that evaluates the performance on year 2 and 3 modules on the classification. We considered only the optional modules because students need to make appropriate choices between them. Obviously, year 2 and year 3 modules will have a strong association with outcomes as their marks do count towards overall degree classification. However, the point of this exercise was to assess if specific optional modules were more highly associated with poor outcomes than others as some knowledge of this could be helpful in implementing a future enrolment system.

For the first dataset, there were 39 different optional modules that we need to consider to account for all the variations. Also, for each of those 39 modules, we considered the performance of the students with respect to their peers as defined in Table 3.4. However, there are other different optional modules that we did not take into consideration because a very low number of students enrolled in them. In this phase of the experiment, the first dataset contained information that associated with 878 students instead of 898 students. We observed that those 20 students that were excluded took different optional modules and that all of them were enrolled in one specific course (*i.e. ACTUARIAL SCIENCES*).

Feature Selection ranking using a chi-square algorithm showed that all of the module performances were important in the classification except for three modules (*DIGITAL SYSTEMS DESIGN, CREATIVE MUSIC TECHNOLOGY A and FURTHER MATHEMATICS*). The modules that were not statistically significant were taken by a low number of students and have similar performance for both NGH and GH groups and were therefore discounted from the rest of the analysis.

The F-test applied to Table 5.10 showed statistically significant differences in the means of year 2 and year 3 optional modules between students that achieve GH/NGH for all except 3 modules. The modules for which the difference of the means were not statistically significant were (*MANAGEMENT ACCOUNTING, BUSINESS FINANCE, PROFESSIONAL PRACTICE AND PROJECT*).

A classification ensemble was built as in the previous experiments, but this time using the year 2 and 3 module performance attributes as well as the previous year 1 module performance attributes and demographic attributes identified by feature selection. We also discounted the year 2 and 3 average as they clearly count toward the degree classification. The autoclassifier chose a Neural Network, Logistic Regression and CHAID as the classifiers and combined them to produce an accuracy over 92.61% on the training data (83.91% on the test sample). This represents a substantial improvement from the previous model. The gain

Table 5.10: Comparison of means for poor performers as selected by ensemble versus all other students in the first Dataset (Third Experiment). Note: * represents statistically significant results.

Attribute	Mean(Poor Performers)	Mean(others)	F-test
INTERNET TECHNOLOGIES	48.99	59.79 %	.000031 statistically significant
DATABASE SYSTEMS	46.79	65.26%	< 0.00001*
GRAPHICS I	46.98	61.88%	< 0.00001*
ARTIFICIAL INTELLIGENCE	45.81	60.28%	< 0.00001*
PRINCIPLES OF MARKETING	55.47	63.07%	.000013*
INTRODUCTORY COMPUTER GRAPHICS	40.14	57.06%	< 0.00001*
SOUND AND IMAGE I	49.10	67.37%	< 0.00001*
ARCHITECTURES AND OPERATING SYSTEMS	47.87	65.26%	< 0.00001*
SYSTEMS ANALYSIS	56.87	64.34 %	.000031*
SOFTWARE DEVELOPMENT TECHNIQUES	51.29	72.48%	< 0.00001*
OPERATING SYSTEMS KERNELS& ARCHITECTURE	45.57	63.79%	< 0.00001*
OPERATIONS STRATEGY AND MANAGEMENT	50.65	62.87%	.000031*
THEORETICAL FOUNDATIONS II	47.89	63.26%	< 0.00001*
CIRCUITS AND SYSTEMS	49.78	65.11%	.000013*
LEGAL ISSUES IN BUSINESS	51.94	62.03%	*.00003
DATA STRUCTURES AND ALGORITHMS	44.55	64.42%	< 0.00001*
FINANCIAL ACCOUNTING	48.23	56.70%	.00003*
SYSTEMS ENGINEERING	56.66	65.41%	.00003*
NETWORKS	49.03	66.44%	< 0.00001*
INFORMATION RETRIEVAL	42.39	61.66%	< 0.00001*
COMPUTER NETWORKS	54.71	67.04%	.000031*
GRAPHICS II	47.11	64.61%	< 0.00001*
ANIMATION; VIRTUAL ENVIRONMENTS AND GAMES DEVELOPMENT	54.69	67.21%	.00003*
SOFTWARE ENGINEERING II	51.54	72.74%	< 0.00001*
MACHINE LEARNING	43.19	62.90%	< 0.00001*
INFORMATION RETRIEVAL AND NATURAL LANGUAGE PROCESSING	42.34	59.63%	.000013*
SOUND AND IMAGE II	48.52	67.23%	< 0.00001*
BEHAVIOURAL ASPECTS OF MARKETING	57.69	66.01%	.000031*
ENTREPRENEURSHIP AND SMALL BUSINESS	57.18	63.53%	.00003*
EMBEDDED SYSTEMS	50.74	70.08%	< 0.00001*
SOFTWARE ENGINEERING FOR THE INTERNET	45.78	59.45%	.00003*
COMPUTER VISION (FOR DIGITAL PHOTOGRAPHY)	38.17	61.62%	< 0.00001*
ADVANCED GRAPHICS	45.69	65.08%	< 0.00001*
MANAGEMENT ACCOUNTING	48.85	49.20%	.72 not statistically significant
BUSINESS FINANCE	61.71	66.41%	.3 not statistically significant
PROFESSIONAL PRACTICE AND PROJECT	64.26	67.00%	.4 not statistically significant

chart in Figure 5.3 c shows the evaluation of the model accuracy. In addition, the gain chart in Figure 5.5 a shows the improvement of the ensemble model of this experiment compare to the previous two experiments' ensemble models. Table 5.11 presents the most important Year2 and Year3 modules in building the classifiers. The three modules that were used in all models were *GRAPHICS II, NETWORKS, INTRODUCTORY COMPUTER GRAPHICS*.

We applied the third phase of the experiment on the second dataset. Again, we considered only the optional modules of year 2 and 3. There were 36 different modules that we needed to consider to account for all the variations. For each of those 36 modules, we also considered the performance of the students with respect to their peers. Again, there were other different optional modules that we did not take into consideration due to the very low number of students enrolled in them. The second dataset contained information that associated with 1,775 students instead of 1789 students. We observed that those 14 students took different

Table 5.11: Year 2 and Year 3 most important modules in building each classifier (The first dataset).

Ensemble	Neural Net	CHAID	Logistic Regression
1. DATABASE SYSTEMS. 2. PRINCIPLES OF MARKETING. 3. SOUND AND IMAGE I. 4. INTRODUCTORY COMPUTER GRAPHICS. 5. SOFTWARE DEVELOPMENT TECHNIQUES. 6. NETWORKS. 7. GRAPHICS II. 8. SOFTWARE ENGINEERING II.	1. INTERNET TECHNOLOGIES. 2. GRAPHICS I. 3. INTRODUCTORY COMPUTER GRAPHICS. 4. FINANCIAL ACCOUNTING. 5. NETWORKS. 6. INFORMATION RETRIEVAL. 7. COMPUTER NETWORKS. 8. GRAPHICS II. 9. EMBEDDED SYSTEMS.	1. INFORMATION RETRIEVAL. 2. DATABASE SYSTEMS. 3. ARTIFICIAL INTELLIGENCE. 4. INTRODUCTORY COMPUTER GRAPHICS. 5. ARCHITECTURES AND OPERATING SYSTEMS. 6. SYSTEMS ANALYSIS. 7. SYSTEMS ENGINEERING. 8. NETWORKS. 9. GRAPHICS II.	1. All Year 2 and Year 3 modules.

optional modules and that all of them enrolled in one specific course (*ACCOUNTING AND FINANCE*).

We found using the Feature Selection based on chi-square algorithm, that all of the module performances of the second dataset were important to the classification except for two modules: *(COMPANY LAW) and (THE ECONOMICS OF FILM AND TV (CW))*. Again, the modules that were not statistically significant were taken by a low number of students and were therefore discounted from the rest of the analysis. In addition, the F-test showed statistically significant differences in the means reported in Table 5.12 between students that achieve GH/NGH. Those that obtained NGH had statistically significant lower marks on year 2 and 3 optional modules.

Moreover, a classification ensemble was built using the year 2 and 3 module performance attributes as well as the previous year 1 module performance and demographic attributes identified by feature selection. The auto-classifier chose a Logistic Regression, a Neural Net and a Decision List as the classifiers and combined them to produce an accuracy over 88.18% on the training data (83.83% on the test sample). This also represents an improvement from the previous model for the second dataset. Figure 5.4 c shows substantial gain for the model including year 2 and 3 performance attributes. In addition, the gain chart in Figure 5.5 b shows the improvement of the ensemble model of this experiment with respect to the previous two experiments' ensemble models and to the baseline. Table 5.13 presents the most important Year2 and Year3 modules in building the classifiers. The two modules that were used in all models were *STRATEGIC BRAND MANAGEMENT, INTERNATIONAL FINANCIAL SERVICES*.

Table 5.12: Comparison of means for poor performers as selected by ensemble versus all other students in the Second Dataset(Third Experiment). Note: * represents statistically significant results.

Attribute	Mean(Poor Performers)	Mean(others)	F-Test
BUSINESS SKILLS FOR MANAGERS	58.78	68.96 %	.00003 statistically significant
BUSINESS FINANCE	50.54	58.66%	.000031*
PRINCIPLES FOR CORPORATE STRATEGY	52.07	61.93%	< 0.00001*
FINANCIAL ACCOUNTING	46.47	59.91%	.000031*
MANAGEMENT ACCOUNTING	47.49	64.61%	< 0.00001*
BEGINNERS SPANISH I	48.79	62.29%	.00003*
STRATEGIC BRAND MANAGEMENT	55.30	65.03%	.00003*
PERSONAL AND CORPORATE TAXATION	52.28	69.28%	< 0.00001*
INTERNATIONAL FINANCIAL SERVICES	52.81	66.44%	< 0.00001*
AUDIT AND ACCOUNTABILITY	57.18	69.08%	.00003*
LEGAL ISSUES IN BUSINESS	52.49	64.84%	.00003*
FINANCIAL MODELLING	67.89	76.55%	.000031*
MANAGEMENT CONSULTING AND DEVELOPMENT	58.99	68.27 %	.000031*
MARKETING: SOCIAL RESPONSIBILITY AND THE LAW	53.45	64.68%	.00003*
FURTHER MATHEMATICS	73.45	86.05%	.00003*
BEGINNERS JAPANESE I	58.93	66.43%	.00003*
BUSINESS ETHICS	51.94	59.86%	.000031*
QUANTITATIVE METHODS FOR BUSINESS	50.29	62.41%	.00003*
ECONOMICS FOR INTERNATIONAL BUSINESS	60.09	69.09%	.000031*
BUSINESS AND COMPANY LAW	46.93	61.11%	< 0.00001*
MANAGING INNOVATION AND CREATIVITY	51.59	64.61%	.00003*
INTERACTIVE MARKETING	55.03	62.77%	.00003*
MARKETING COMMUNICATIONS	49.23	66.65%	< 0.00001*
NEW EMPLOYMENT SYSTEMS	56.74	64.74%	.00003*
ADVANCED MANAGEMENT ACCOUNTING	53.07	64.85%	.000031*
INTERNATIONAL BUSINESS	60.64	71.79%	.00003*
CONTEMPORARY WORKPLACE RELATIONS - CONFLICT AND COOPERATION	52.45	66.17%	.000031*
STRATEGIC BUSINESS DECISION MAKING	55.47	67.91%	.00003*
INFORMATION SYSTEMS FOR MANAGEMENT	53.58	66.69%	< 0.00001*
PRINCIPLES OF MARKETING	57.40	65.29%	.00003*
BEGINNERS FRENCH I	60.15	67.64%	.00003*
BEHAVIOURAL ASPECTS OF MARKETING	50.08	65.32%	< 0.00001*
SUPPLY CHAIN MANAGEMENT: STRATEGY AND DESIGN	50.20	67.50%	< 0.00001*
KNOWLEDGE MANAGEMENT	52.15	64.32%	.00003*

5.5 Discussion

We have found that the results of our preliminary analysis assessing each attribute independently using a Feature Selection ranking algorithm are in accordance with what has been found in other studies (e.g. [251], [252], [253], and [254]). Previous studies have found that Home students are associated with higher attainment than OS students ([253, 251] and [252]). In contrast, some studies [253, 255] found that there were no statistically significant differences in the class of degree obtained by OS students compared to Home students. However, this tended to be in disciplines such as agriculture, librarianship and information science, engineering and technology, mathematical sciences or combined studies. Still, there have been statistically significant differences in other specific disciplines [253] such as architecture, computer sciences, building and planning, social, economic and political studies, law, business and administrative studies. In the later subjects, Home students have higher levels of attainment than OS students.

Table 5.13: Year 2 and Year 3 most important modules in building each classifier (The second dataset).

Ensemble	Neural Net	Decision List	Logistic Regression
1. STRATEGIC BRAND MANAGEMENT. 2. INTERNATIONAL FINANCIAL SERVICES. 3. PERSONAL AND CORPORATE TAXATION. 4. BUSINESS ETHICS. 5. BUSINESS AND COMPANY LAW.	1. STRATEGIC BRAND MANAGEMENT. 2. INTERNATIONAL FINANCIAL SERVICES. 3. PERSONAL AND CORPORATE TAXATION. 4. FURTHER MATHEMATICS. 5. BUSINESS ETHICS. 6. QUANTITATIVE METHODS FOR BUSINESS. 7. BUSINESS AND COMPANY LAW. 8. MANAGING INNOVATION AND CREATIVITY. 9. MARKETING COMMUNICATIONS.	1. STRATEGIC BRAND MANAGEMENT. 2. INTERNATIONAL FINANCIAL SERVICES.	1. All Year 2 and Year 3 modules.

Additionally, our findings were consistent with other studies [251, 253] in terms of gender: female students are more likely to graduate with GH degrees than male students, although they are minorities in some disciplines such as science subjects compared to art subjects. For instance, our first dataset which relates to a science subject has 17% female students, compared to 83% male students; our second dataset has 39% females and 61% males.

Some studies [251] have found that students who come from areas with the lowest levels of participation in HE, and those who come from less affluent areas, are more likely to have lower attainment. In contrast, some other studies in [251] have found no statistically significant difference. Those findings are in agreement with our own findings in terms of the widening of participation: Home students who come from neighbourhoods with very low participation in HE are associated with lower attainment, but there is little difference between other groups (2-low, 3-medium and 4-high). The greatest differences that other studies [251] found in terms of attainment are between students who come from different types of schools, such as comprehensive/independent schools. We did not include this attribute in our data, since we do not have this information in the University Data Warehouse.

Other studies [253, 254] found that mature (21+) and/or full-time students have statistically significant higher levels of attainment than younger/part-time students respectively. The attribute "age at entry" was not statistically significant in our Feature Selection assessment because 88% of students were between the ages of 17 and 21, and we excluded the attribute full/part-time because all the students in the dataset were full-time learners.

We have been able to discover groups of students that have poor performance in terms of good honours grades. Those students are identifiable with some certainty as soon as they arrive by their general characteristics, i.e. gender, course enrolled on, nationality, maths

qualifications and widening participation level. Furthermore, they are more accurately identifiable at the end of year 1 when considering their performance on different modules in that year. We expected that including attributes from module performance would improve predictive accuracy. However, we assumed that particular modules may be found to be problematic when in fact poor performance appears to affect every module of year 1. Moreover, by applying the third experiment, we were able to identify the optional modules that were more relevant to the overall classification. They were DATABASE SYSTEMS, PRINCIPLES OF MARKETING, SOUND AND IMAGE I, INTRODUCTORY COMPUTER GRAPHICS, SOFTWARE DEVELOPMENT TECHNIQUES, NETWORKS, GRAPHICS II, SOFTWARE ENGINEERING II.

The poor performer group show some ability to marginally improve according to their year 2 and 3 averages so targeted intervention could give them enough impulse to achieve GH degrees. If the intervention could achieve a good lift in terms of GH rates, it will also positively affect the University as it will influence league table positions.

Our discovered patterns hold for two different datasets belonging to different schools with different admission strategies and teaching different disciplines. Schools operate quite independently of one another but the same patterns have emerged from both in terms of characteristics of low attainers. We believe this gives some validation to the patterns found.

Some of the immediately obvious interventions could be targeted at the OS students who are prominent in the under achieving group (over 19% in the first dataset and over 39% in the second dataset). Providing extra English language lessons to improve their comprehension and communication skills could achieve the desired effect. Additionally, all those found to be in the selected group of predicted poor performance could be approached by their academic advisers and offered remedial sessions. Remedial sessions could run in the summer remotely to revisit areas of the course where students have done poorly. This may improve their academic knowledge and ability and prepare them to undertake the second and third years from a stronger footing. The analysis did not uncover specific problem year 1 modules as the poor performers seemed to do poorly across the board and on all modules in relation to their peers. However, it uncovered some problematic year 2 and 3 modules that are important in building the overall classifier. Further analysis of module performance may help our overall aim of improving student outcomes, particularly for those highlighted problem modules. Our analysis could also be used to influence admission policies given the characteristics of predicted poor performers.

The next step of the analysis which is not yet included in this thesis is to include additional measure of engagement once they become available in the University data warehouse such as Blackboard activities which may give a measure of engagement.

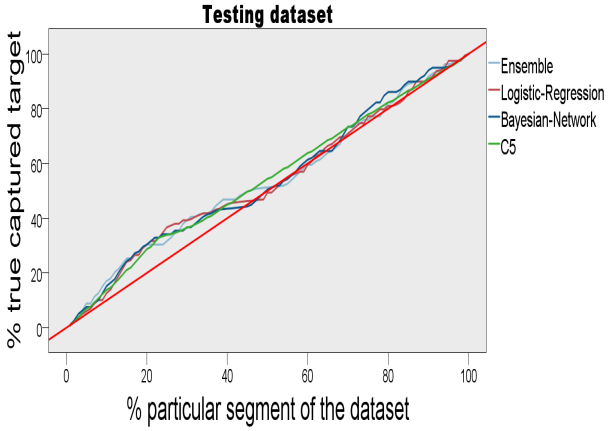
In terms of classifiers, there were no overall winners as different classifiers appear to be best in different experiments but their performance was very close and any differences appeared not statistically significant. The ensemble approach can encompass a compromise between different models. Used to target specific groups by selecting those with a high probability to belong to the target class, it represents knowledge in a usable format.

5.6 Summary

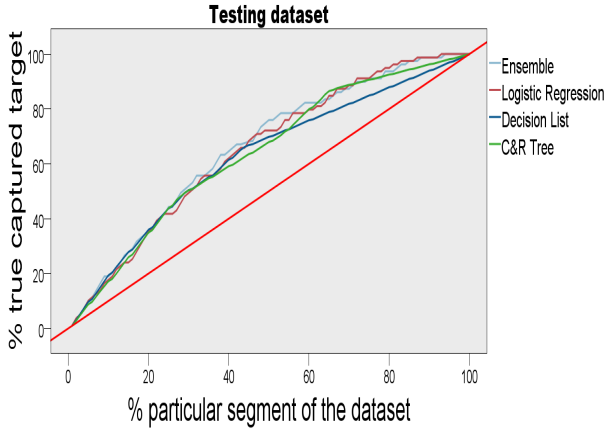
The primary goal of the work in this chapter was to obtain and prepare student data for analysis and to perform initial analysis by predicting students that are at high risk of not achieving a good honours degree, but more importantly, to identify this as early as possible in year 1 so that interventions can be proposed. We have been able to achieve this goal with reasonable accuracy by using classification models to highlight the students that are predicted to be low achievers with high probability. Simple models built with a few attributes known at the time of registration are sufficient to identify a group containing up to 57% of the low attainers with GH rates as low as 32.6%. When combining this with first year performance, we were able to identify 89% of the low attainers. The group identified had a GH rate of 21.97%. Moreover, the built models were able to uncover some year 2 and year 3 optional modules that seem more correlated to the overall outcome.

The next practical step in putting our results to the test, i.e deploying the knowledge uncovered, would be to recommend strategies based on this and measure performance improvements. This is not a feasible part of this thesis as it depends on external agents and we are not at liberty to implement changes. We do however investigate the attitude of key members of staff in relation to the study and our findings as that will uncover the obstacles in the implementation of the knowledge found in an educational data mining project such as ours. For this, we conduct a questionnaire survey that targets faculty members to understand their attitudes. Chapter 7 presents the results of our investigation into issues such as whether the University should act on the above findings to improve students outcomes; what could/should the University offer to those at risk, and whether assistance should be offered to those at risk or to all students.

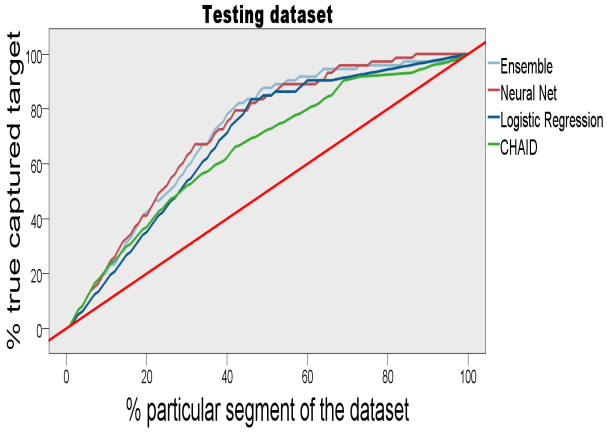
In the next chapter we address the more granular issue of performance prediction for elective modules. As for module prediction we encounter missing data, we address this issue and present how multiple imputation methods can be used in this context to improve the models.



(a) First Experiment



(b) Second Experiment

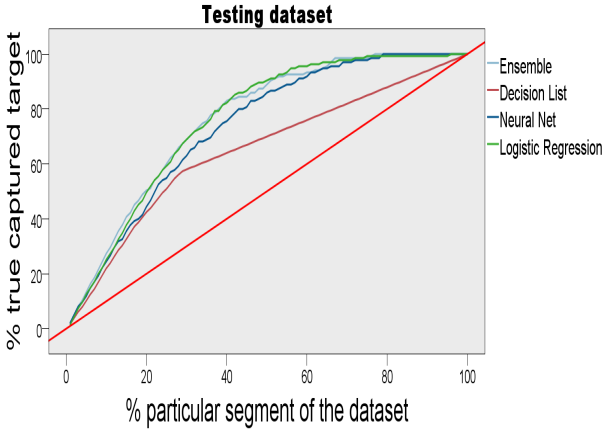


(c) Third Experiment

Figure 5.3: The gain chart showing the percentage of positive predictions that the model gains for each segment of the dataset predicted. This chart is based on the testing sample from **the first dataset**. The gap between the red line (no model) and each of the remaining lines(derived models) shows the percentage of correct target selection with the derived model over a random selection of targets. Note that x-axis is sorted by the probability of the target outcome, highest to lowest.

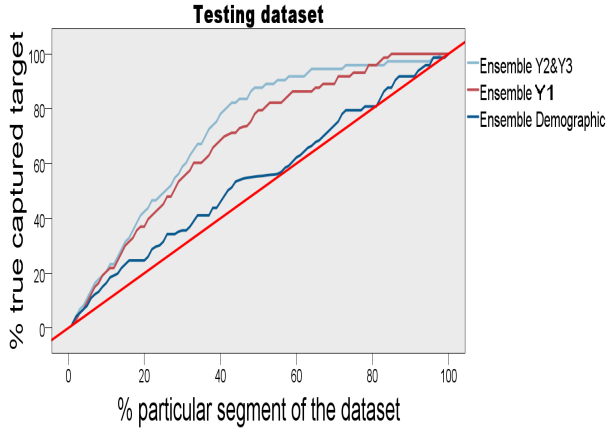
(a) First Experiment

(b) Second Experiment

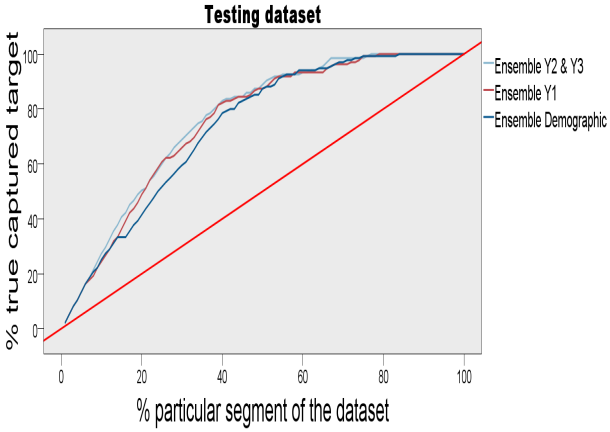


(c) Third Experiment

Figure 5.4: The gain chart shows the percentage of positive predictions that the model gains at each segment of the dataset. This chart is based on the testing sample from **the second dataset**. The gap between the red line (no model) and each of the remaining lines(derived models) shows the percentage of correct target selection with the derived model over a random selection of targets. Note that the x-axis is sorted by the probability of the target outcome, highest to lowest.



(a) First dataset



(b) Validation dataset

Figure 5.5: The gain chart shows the percentage of positive predictions that the ensemble model of each experiment gains at each segment of the dataset. This chart is based on the testing sample from the dataset. The gap between the red line (no model) and each of the remaining lines(derived models) shows the percentage of correct target selection with the derived model over a random selection of targets. Note that the dataset of x-axis is sorted by the probability of the target outcome, highest to lowest.

Chapter 6

Generating module-level performance predictions

6.1 Introduction

Nowadays, higher education institutions face many challenges, such as increases in student numbers [256] and diversity [257, 258], a considerable reduction in government funding [259], and a globally competitive education market [256, 260]. These challenges are forcing universities to re-think the best approaches to deliver and support education. The utilisation of advance analytics such as the prediction of students' performance to guide students through their choices and improve their outcomes could help with some of those challenges [261]. In this context, educational data mining, as discussed earlier in the thesis, is becoming an important area of research [262, 39].

For example, when applications are evaluated, the prediction of academic performance may help universities in finding applicants who are going to excel for a specific academic program [263]. The results produced from prediction systems for current students may be used to offer extra support such as tutoring resources and customised personal assistance as we discussed in the previous chapter. The outcomes of prediction can also be utilised by educators to identify the most appropriate teaching materials and actions for each category of students. In addition, performance prediction can assist students to make choices regarding courses and universities. Therefore, developing prediction tools should be beneficial for higher education institutions. Nevertheless, accurately predicting student module outcomes in practice is complex, due to the large number of factors that affect student academic performance, such

as demographics, social characteristics, previous learning, delivery methods, etc. [264].

Missing data is another important factor that can affect the accuracy of many predicting systems. We make an important point of addressing this in our study, because missing data is a common issue in educational data. However, to the best of our knowledge, many educational data mining empirical studies have not addressed how to deal with missing data adequately. Many data mining techniques were developed for complete datasets, and missing data was handled by sometimes questionable pre-processing methods. One such method is the removal of records with missing values. Another method is a simple imputation replacing missing values with the mean values of the affected variable. Such methods can introduce bias and have been critiqued in the literature. Little and Rubin [65, p.39] stated, *"We do not generally recommend any of them"*. Also, Wilkinson [265] cautioned against their use commenting that they are *"among the worst methods available for practical applications"*.

Over the last few decades, there have been considerable methodological improvements in the field of missing data analysis [266]. Particularly, the technique of multiple imputation (MI) is currently considered to be *"state of the art"* [267] and is the recommended method for imputation [268]. We apply two MI approaches, chained equations and expectation maximisation in the context of data mining to improve our module prediction systems, since module information is often missing, as many students do not make the same module choices.

Previously in chapter 5, we used prediction models to attempt to highlight students at risk of overall poor performance (i.e. failure to gain a good honours degree) using data collected by the business intelligence unit of the University of East Anglia. The purpose of this chapter is to tackle the more difficult task of module-level prediction using previous academic performance and student characteristics as discussed in chapter 3. In our higher education setting, module choices can be vast, taking a student through different career paths and presenting different challenges and opportunities. Module choices are sometimes made with guidance from a qualified faculty member. However, increasingly, students make their enrolment decisions on their own. Thus, the enrolment process mainly depends on the students' experience and the accessible information, but this is often inadequate to assess the time, effort and academic skills needed for each module. Higher education institutions usually make direct information, such as existing module descriptions, assessment patterns, schedules and instructors, available to students. Information about other students' experiences and outcomes from previous enrolment is often not made available. Therefore, it could be very beneficial to provide more information to the students regarding their predicted outcomes, given their characteristics and the outcomes of similar students, and further integrate this into a system that could help students make better enrolment-related decisions. Here

we focus on predicting module outcomes accurately.

This chapter, following some of the studies reviewed in Chapter 2, looks at predicting performance as both a regression and classification problem in order to understand if one approach can provide better results than the other. We focus on leading data mining algorithms for each task. Many of the studies have been conducted on a single dataset; we attempt to make our study more robust by using different datasets. First, we include two different schools, albeit within the same university, and then we complement that with a public dataset on student performance. As we are comparing multiple algorithms over multiple datasets we use the Friedman rank statistical test introduced by Demšar, which is the recommended statistical test in such scenarios [204] to establish statistically significant different performances. The contribution of this work is to examine the best data mining algorithms to accurately predict student performance in the context of extensive missing data, and to study the effect of the missing data on performance. For this, we apply multiple imputation by chained equations and expectation maximisation approaches, and Random Forest imputation in an ensemble data mining context, which appears to be novel in relation to the reviewed literature. We also experiment with increasing amounts of missing data in the publicly available complete datasets.

The rest of this chapter is organised as follow: Section 6.2 summarises the experimental work undertaken; the results are described in section 6.3; Section 6.4 includes the discussion of the results; lastly, Section 6.5 summarises this chapter.

6.2 Experimental Set up

In this set of experiments, we attempt to predict module performance as a mark using regression techniques, and as a categorical label (**G**ood **H**onours/ **N**ot **G**ood **H**onours) using classification techniques. This is to enable the comparison of both approaches for performance prediction.

6.2.1 Regression experiments

We applied and compared a number of regression prediction methods:

- Our first prediction (Simple Average) is just a baseline. The predicted mark is the average marks of the previous students who took the same module. Hence, this is a naive prediction, not taking into account any of the student's characteristics and simply looking at other students' past performance on a given module.

- The second prediction system is based on clustering. We use clustering techniques to partition students based on the similarities of their academic records, and then we use cluster average marks as predictors. Hence, for this method we take into consideration the student's characteristics and the performance of similar students.
- The third prediction system is based on the use of the regression algorithm Rpart. We applied Rpart using 10-fold cross validation. To avoid overfitting the data, we pruned the tree using the complexity parameter that was associated with the minimum cross-validated error [269].
- The fourth prediction model uses Random Forests [168].
- The fifth prediction system is SVM [169]. For both Random Forests and SVM, we used 10-fold cross validation.

To test how to handle missing data, we applied the Rpart, Random Forests and SVM algorithms four times on each dataset:

- First, we attempted to apply the algorithms on each dataset without imputation. However, for algorithms that cannot handle missing data (e.g. SVM), we used naive imputation, by replacing the numeric missing values with the average value for that attribute, and the non-numeric missing values with the mode (most commonly occurring) value for that attribute. We could not remove the records or attributes with missing data (i.e. perform complete case analysis) due to their large number, as discussed earlier in section 3.1.
- Second, we applied the algorithms on each dataset with single Random Forests imputation.
- Third, we applied them on each dataset with multiple imputation using the chained equations approach.
- Fourth, we also applied the algorithms on each dataset with multiple imputation using the expectation maximisation (EM) approach.

It should be noted that imputation was not possible for some columns in some of the datasets because there was insufficient available information. For the public datasets we have not applied the simple average prediction system, due to the absence of the previous students' records required to calculate this. We have also, for obvious reasons, not applied the algorithms with the imputation on the completed version of the public datasets. Also, note that we imputed the missing values on the test sets due to several reasons. First, some of the used algorithms such as SVM and RF do not accept missing values in the test set. Second, we believe that missing values should be handled in the data pre-processing phase.

Lastly, the basic assumption in machine learning is that training and testing sets are drawn from the same population, and therefore follow the same distribution [270].

Once our predictions were obtained according to the different methods, we computed the RMSE for each regression prediction. Then, we statistically compared the predictive accuracy of those different algorithms and techniques, with an emphasis on comparing the different missing data handling approaches.

Lastly, to measure the statistical significance of any detected differences between the mean of the RMSE computed in the previous step, we applied the Friedman rank test. Then, we presented the test results using critical difference diagrams.

6.2.2 Classification Experiments

Again, we applied and compared a number of classification methods. In fact, the same algorithms used for regression were used for classification since they are also applicable. However, we did not apply the clustering prediction system for classification, instead applying the C5.0 algorithm. Again, we applied the data mining algorithms (C5.0, Rpart, Random Forests and SVM) four times on each dataset (without imputation for C5.0 and Rpart or with naive imputation for Random Forests and SVM, with single Random Forests imputation, and with multiple imputation). We applied the algorithms using 10 fold cross-validation. Experiments were performed on both schools and the public datasets. As before, we could not apply the simple average prediction method for the public datasets, and did not apply the algorithms with the imputation on the completed version of these datasets.

We computed the accuracy for each classification method and used the Friedman rank test to test the statistical significance of any differences. Then, we presented the test results using critical difference diagrams.

6.3 Results

We performed our set of experiments using RStudio version 1.0.44 [271]. Before we started clustering the data, we used a heat map, as shown in Figure 6.1, to visualise the first dataset's distance matrix obtained using the Gower coefficient. The black scale (where distance $\leq$ 0.1) reflects strong similarity between student objects, and it scales through yellow, green and blue until it reaches the white colour (where distance $>$ 0.6) to reflect dissimilarity between student objects. Figure 6.1 shows two zones where yellow and black colours reflect the most similar students, and the other two zones of blue colour show the most dissimilar

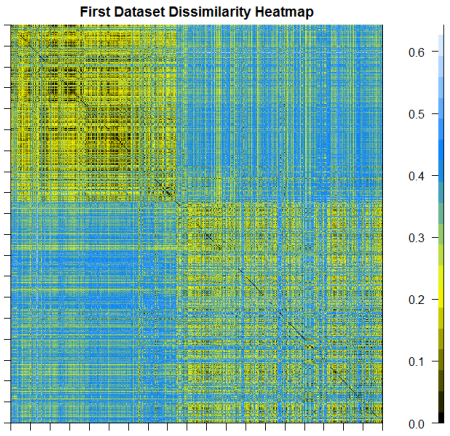


Figure 6.1: This heatmap shows the dissimilarity between students in the first-school dataset. The black scale reflects strong similarity ≤ 0.1 and scales through yellow until it reaches the white colour to reflect dissimilarity > 0.6.

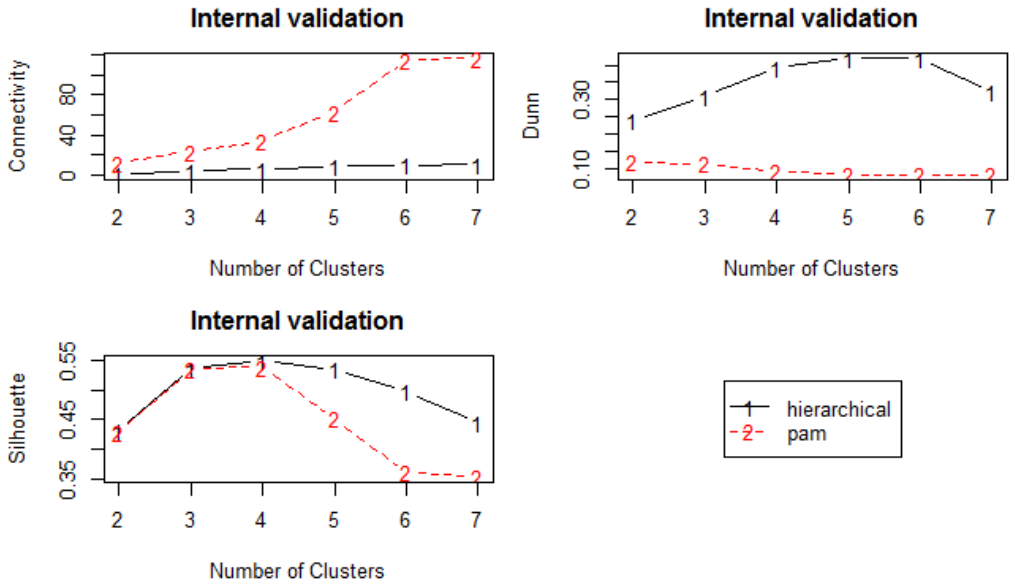


Figure 6.2: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the first-school dataset. The x-axis shows the number of clusters (from 2 to 7 clusters), while the y-axis shows the score of the validation test.

students. This provided us with an initial indication that we would be able to successfully cluster the first-school dataset.

When we compared the different approaches to clustering on the first dataset using internal validation techniques, we found as shown in Figure 6.2, that hierarchical clustering outperforms PAM by producing a best score of 0.1429 by Connectivity measure (note this has a

range between 0 and ∞ and should be minimised), 0.4194 by Dunn measure (note this has a range between 0 and ∞ and should be maximised), and 0.5492 by Silhouette measure (note this has a range between -1 and 1 and should be maximised). Figure 6.2 shows the score in relation to the cluster size (2 to 7 clusters), and it can be seen that the best cluster size was not consistent between different evaluation methods, as Connectivity scores best with cluster size 2, Dunn scores best with cluster size 5, and Silhouette scores best with cluster size 4. The difference between hierarchical clustering with 4 and 5 clusters is that the 5th cluster contains only three students who have not taken the selected optional modules. However, the other 4 clusters are equal. Later on we will present RMSE for a hierarchical clustering with size 4. We found, through using CPCC evaluation, that hierarchical clustering produced by average linkage was best, scoring 0.8943946, whereas the complete linkage and single linkage methods scored 0.8533753 and 0.8293649, respectively. We discounted Ward's method from the CPCC computation, as it is based on dissimilarity between the centroid of the cluster, not on the dissimilarity between the objects of the cluster [184].

Next, for the second-school dataset, we again used a heat map to visualise the distance matrix obtained using the Gower coefficient, as shown in Figure 6.3. The black scale (where distance ≤ 0.1) reflects strong similarity between student objects, and it scales through yellow until it reaches the white colour (where distance > 0.6) to reflect dissimilarity between student objects. However, this time Figure 6.3 does not show us clear distinguishable zones that can reflect the most similar and dissimilar students. This is an initial indication that the cluster results are worse than the results of the first-school dataset.

We also compared the different approaches to clustering on the second dataset using internal validation techniques. As shown in Figure 6.4, we found that hierarchical clustering outperforms PAM by producing a best score of 5.0329 by Connectivity measure, 0.2317 by Dunn measure, and 0.3049 by Silhouette measure. Figure 6.4 also shows the score in relation to the cluster size (2 to 7 clusters), and it can be seen that best cluster size was consistent this time between different evaluation methods, as all the validation methods score best with cluster size 2. Later, we will present RMSE for a hierarchical clustering with size 2. We found, through using CPCC evaluation, that hierarchical clustering produced by average linkage was best by scoring 0.6296929, whereas the complete linkage and single linkage methods scored 0.5480718 and 0.5446145, respectively. We also discounted Ward's method from the CPCC computation.

We present the mean and standard deviation of the 10 RMSE values obtained by cross validation for each prediction system in Table 6.1 for the first-school dataset, Table 6.2 for the second-school dataset, and Table 6.3 for the public datasets. We found that the prediction systems with EM multiple imputation are slightly better for a number of datasets, along

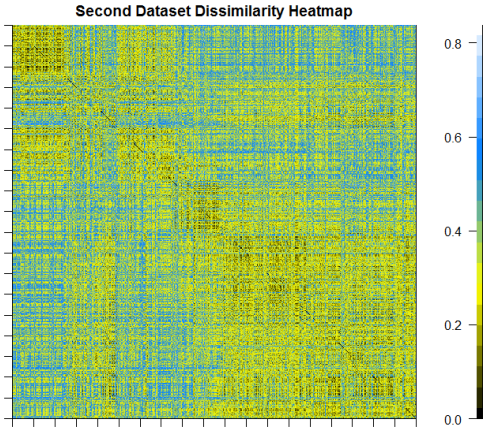


Figure 6.3: This heatmap shows the dissimilarity between students in the second-school dataset. The black scale reflects strong similarity ≤ 0.1, and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.6.

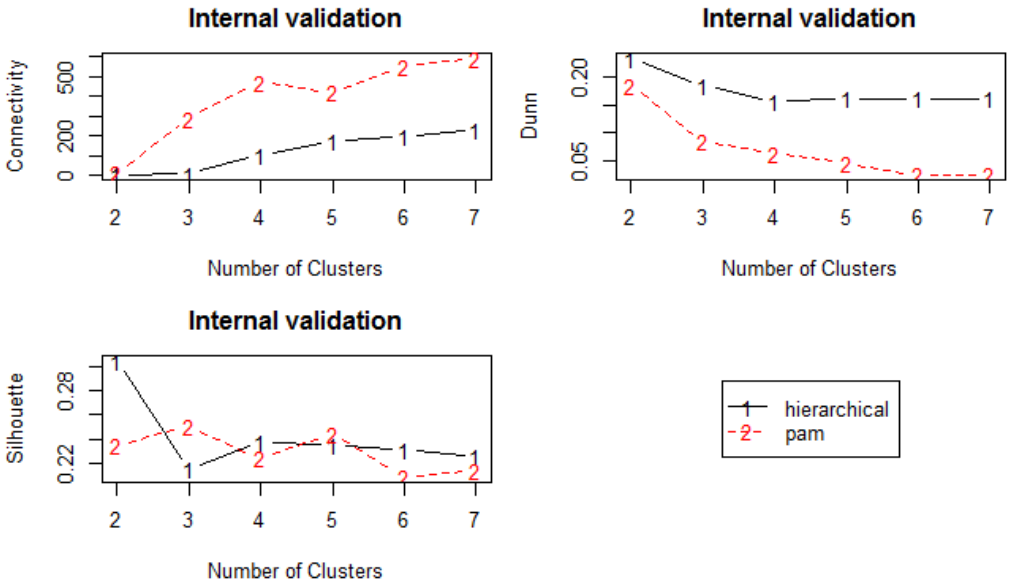


Figure 6.4: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the second-school dataset. The x-axis shows the number of clusters (from 2 to 7 clusters), while the y-axis shows the score of the validation test.

with the clustering approach. Particularly, we found that SVM with EM-based multiple imputation perform best for a number of datasets. The results also show that the simple average was associated with the worst performance, hence a predictive model based on students' characteristics and performance does show an advantage. For the public datasets, where we undertook the additional comparison between complete data and data with 25%

and 45% of values removed, we can observe that the removal of data did not result in a comparable deterioration of RMSE. In fact, statistical testing with a Wilcoxon signed rank test showed no statistically significant difference (p-value > 0.05). For the incomplete datasets, methods combined with multiple imputation provide the best results. Hence, models built with multiple imputation in the presence of large amounts of missing data do not show statistically significant deterioration with respect to the complete data.

To statistically validate the results across all the prediction systems we applied the Friedman rank test. The results were statistically significant ($p \leq 0.05$), 9.9920×10^{-16} according to both schools and the public datasets. Figure 6.5 shows the critical difference diagram resulting from the post-hoc Nemenyi test. The diagram summarises the differences in the average ranks of the 14 prediction systems over 16 datasets for both schools of study and the external educational institution, with the bars indicating cliques, within which there is no statistically significant difference in rank. From the average rank for each regression prediction system (lower rank means better performance) we can see that SVM with multiple imputation by chained equations(SVM_{MI}), followed by SVM with EM multiple imputation(SVM_{EM}), have the best performance across the 16 datasets that are associated with both our schools of study and the external school. We can also observe that multiple imputation methods are among the top ranked, regardless of the algorithm used, and imputation in general is beneficial, as the top 8 ranked methods involve some form of imputation. Figure 6.5 also shows that the prediction systems with imputation statistically significant outperformed the baseline, Rpart with RF imputation and the clustering method over all datasets.

Next, we present the mean and standard deviation of the 16 Accuracy values obtained by cross-validation for each classification prediction system in Table 6.4 for the first-school datasets, Table 6.5 for the second-school datasets, and Table 6.6 for the public datasets. We found that the prediction systems with EM multiple imputation are slightly better for a number of datasets. For the experiment on the public datasets, where data is removed at random, perhaps surprisingly, the removal of large amounts of data results in some methods achieving slightly higher accuracy (for the 25% and 45% incomplete datasets). This was found to be statistically significant better with a Wilcoxon signed rank test between POR and POR45 (p-value= 0.01953) and between Math and Math45 (p-value = 0.05248). Modelling with high levels of uncertainty does not result in any deterioration of the model, as the imputation techniques manage very good results. Again, the simple average prediction system was associated with the worst performance.

We computed F1-score in addition to the accuracy to evaluate the performance of the classification approach. However, we found that F1-score and accuracy provided similar

Table 6.1: Comparison of RMSE mean values for each prediction system for the first school datasets. The standard deviation is in brackets.

	DS	NW	IT	SA	SE
SimpleAvg	14.162(1.804)	13.685(1.293)	10.074(2.049)	8.154(1.708)	10.027(2.293)
Clustering	11.375(8.812)	**8.138**(7.631)	**6.002**(5.082)	8.141(4.118)	10.385(7.951)
Rpart	11.532(1.751)	12.060(2.067)	8.745(2.190)	7.266(1.052)	9.797(2.064)
Rpart_{RF}	12.530(2.127)	12.222(1.791)	9.001(1.720)	7.243(0.937)	9.426(1.945)
Rpart_{MI}	10.605(1.540)	11.516(1.893)	8.287(1.379)	7.196(1.344)	9.090(1.773)
Rpart_{EM}	10.207(1.382)	11.276(1.602)	7.463(1.335)	7.035(1.275)	8.612(1.859)
RF	11.112(1.674)	11.858(1.392)	8.597(1.452)	7.276(1.613)	9.449 (2.065)
RF_{RF}	10.035(1.821)	11.484(1.524)	8.0499(1.509)	7.218(1.162)	9.648(2.002)
RF_{MI}	10.459(1.450)	11.878(1.448)	7.955(1.681)	7.068 (1.003)	8.708(1.652)
RF_{EM}	**10.011**(1.437)	10.886(1.468)	7.353(1.279)	6.698(1.368)	8.557(1.781)
SVM	13.804(1.939)	13.502(1.315)	8.352(1.547)	8.154(1.742)	9.243(2.482)
SVM_{RF}	10.017(1.537)	11.425(1.709)	7.866(1.566)	7.179(1.382)	9.0821(1.840)
SVM_{MI}	10.293(1.355)	11.011(1.890)	7.854(1.573)	7.039(1.463)	8.966(1.863)
SVM_{EM}	10.286(1.625)	10.194(1.364)	7.335(1.335)	**6.594**(1.277)	**8.438**(1.732)

results. To make the result section easier to read and understand, we decided to move the F1-score results to Appendix H.

We applied the Friedman rank test, obtaining statistically significant results ($p \leq 0.05$), with a value of 7.6536×10^{-10}. Figure 6.6 shows the critical difference diagrams summarising the differences in the average ranks of the 17 prediction systems over 16 datasets. Again, multiple imputation methods obtained the lowest ranks, with SVM_{EM}, RF_{EM}, SVM_{MI}, RF_{MI} performing best among all the algorithms, and statistically significant better than the baseline. Overall, however, differences in performance were small and not statistically significant.

6.4 Discussion

In this study, we observe that SVM and RF with an ensemble, in the context of multiple imputation, can lead to promising results. There was no clear advantage between the classification and regression approaches. This contradicts a similar study [132] (reviewed earlier in chapter 2), which claimed that classification methods perform better than regression methods. We do observe, using the critical diagram, that performances are more differenti-

Table 6.2: Comparison of RMSE mean values for each prediction system for the second school datasets. The standard deviation is in brackets.

	EB	PT	IS	SM	FM
SimpleAvg	6.447(0.509)	11.782(1.323)	11.065(1.588)	8.218(1.106)	9.678(1.156)
Clustering	**4.034**(3.149)	13.080(7.589)	13.073(8.168)	**5.420**(3.979)	**5.854**(4.597)
Rpart	5.354(0.661)	9.861(1.393)	9.435(0.754)	6.883(0.999)	9.006(1.188)
Rpart_{RF}	5.426(0.685)	10.015(1.297)	9.525(0.807)	7.021(0.993)	8.778(1.457)
Rpart_{MI}	5.322(0.618)	9.785(1.286)	9.201(0.868)	6.839(1.027)	8.952(1.220)
Rpart_{EM}	5.264(0.698)	9.566(1.218)	9.323(0.805)	6.584(1.018)	8.957(1.226)
RF	5.090(0.704)	9.689(1.321)	**9.011**(0.957)	6.546(1.098)	8.740(1.415)
RF_{RF}	5.245(0.664)	9.938(1.366)	9.145(0.974)	6.567(1.121)	8.900(1.193)
RF_{MI}	5.245 (0.695)	9.557(1.212)	9.123(0.860)	6.504(1.220)	8.929(1.203)
RF_{EM}	5.133(0.636)	9.472(1.033)	9.103(0.934)	6.411(1.086)	9.042(1.156)
SVM	6.011(0.582)	9.734(1.332)	9.095(0.865)	6.304(1.107)	8.909(1.203)
SVM_{RF}	5.008(0.762)	9.722(1.308)	9.091(0.841)	6.311(1.095)	8.875(1.170)
SVM_{MI}	5.001(0.727)	9.561 (1.290)	9.051 (0.874)	6.266 (1.100)	8.772(1.244)
SVM_{EM}	4.991(0.703)	**9.360**(1.141)	9.149(0.824)	6.211(1.045)	8.798(1.354)

Table 6.3: Comparison of RMSE mean values for each prediction system for the publicly available datasets. The standard deviation is in brackets. NA stands for Not Applicable.

	Math	Math 25%	Math 45%	Por	Por 25%	Por 45%
SimpleAvg	NA	NA	NA	NA	NA	NA
Clustering	4.563(0.446)	4.565(0.445)	4.160(0.557)	3.222(0.331)	3.191(0.281)	3.191(0.325)
Rpart	**2.601**(0.415)	**2.506**(0.524)	2.882(0.588)	**1.706**(0.320)	1.889(0.373)	1.847(0.222)
Rpart_{RF}	NA	2.931(0.537)	2.817(0.680)	NA	1.918(0.324)	1.804(0.254)
Rpart_{MI}	NA	2.702(0.542)	2.757(0.533)	NA	1.898(0.302)	1.845(0.261)
Rpart_{EM}	NA	2.561(0.489)	2.963(0.646)	NA	1.854(0.302)	1.824(0.214)
RF	3.539(0.581)	3.796(0.481)	3.680(0.495)	2.205(0.310)	2.333(0.354)	2.294(0.299)
RF_{RF}	NA	2.729(0.477)	2.849(0.548)	NA	1.882(0.367)	1.917(0.327)
RF_{MI}	NA	2.658(0.462)	2.794(0.435)	NA	1.866(0.360)	**1.778**(0.344)
RF_{EM}	NA	2.532(0.473)	2.817(0.492)	NA	1.867(0.335)	1.831(0.301)
SVM	2.866(0.459)	2.924(0.499)	2.832(0.571)	1.819(0.401)	1.892(0.388)	1.856(0.354)
SVM_{RF}	NA	2.847(0.527)	2.820(0.567)	NA	1.881(0.388)	1.858(0.370)
SVM_{MI}	NA	2.831(0.505)	**2.749**(0.483)	NA	**1.839**(0.386)	1.801(0.376)
SVM_{EM}	NA	2.816(0.518)	2.840(0.520)	NA	1.870(0.395)	1.814(0.355)

ated with the regression approach. We believe this is because regression prediction provides finer grain answers compared to the binary output of the classification approach. We found

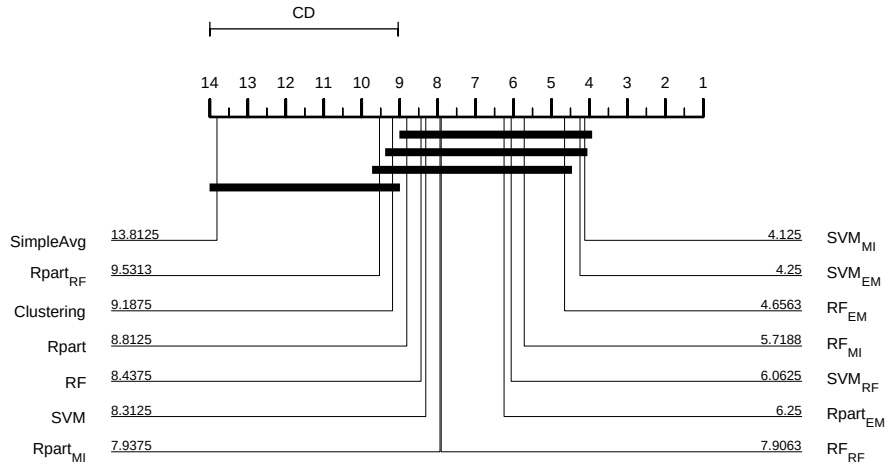


Figure 6.5: Critical difference diagram for the RMSE across the 16 datasets associated with both school of study and the external institution. The decimal number close to each prediction system is the value of its average rank used in the Friedman test computation.

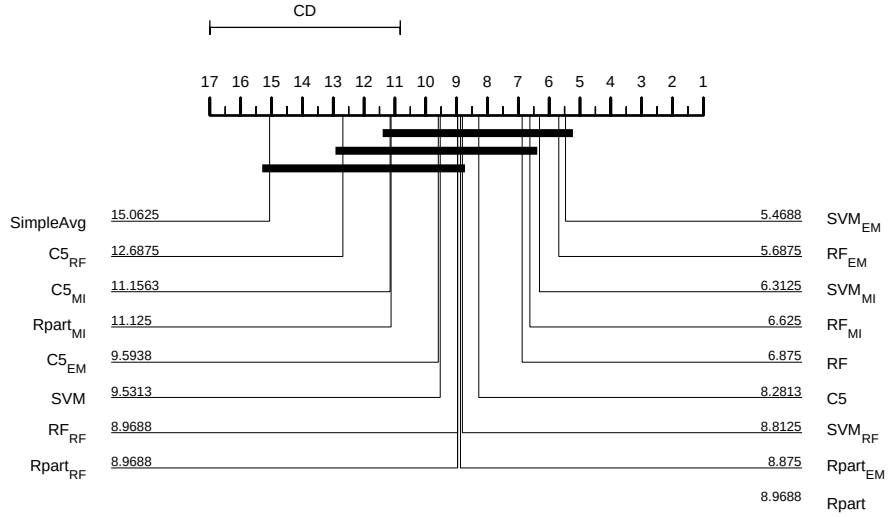


Figure 6.6: Critical difference diagram for classification accuracy across the 16 datasets associated with both schools of study and the external institution. The decimal number that is close to each prediction system is the value of its average rank used in the Friedman test computation.

Table 6.4: Comparison of Accuracy mean values for each prediction system for the first school datasets. The standard deviation is in brackets.

	DS	NW	IT	SA	SE
SimpleAvg	0.500(0.059)	0.583(0.086)	0.633(0.144)	0.567(0.107)	0.600(0.112)
Rpart	0.736(0.107)	0.697(0.090)	0.626(0.118)	0.659(0.114)	0.658(0.117)
Rpart_{RF}	0.714(0.088)	0.663(0.087)	0.642(0.104)	0.664(0.068)	**0.716**(0.100)
Rpart_{MI}	0.700(0.069)	0.681(0.061)	0.651(0.086)	0.652(0.075)	0.633(0.094)
Rpart_{EM}	0.709(0.068)	0.687(0.042)	0.759(0.058)	0.653(0.0348)	0.672(0.087)
C5	0.711(0.095)	0.633(0.068)	0.637(0.112)	0.636(0.105)	0.674(0.118)
C5_{RF}	0.694(0.090)	0.643(0.111)	0.663(0.112)	0.641(0.094)	0.584(0.098)
C5_{MI}	0.699(0.0598)	0.677(0.060)	0.636(0.077)	0.645(0.069)	0.619(0.097)
C5_{EM}	0.707(0.066)	0.678(0.0466)	0.740(0.064)	0.649(0.0390)	0.654(0.100)
RF	**0.750**(0.087)	0.653(0.103)	0.721(0.124)	0.673(0.088)	0.658(0.139)
RF_{RF}	0.747(0.069)	0.673(0.049)	0.679(0.150)	0.645(0.088)	0.674(0.128)
RF_{MI}	0.742(0.083)	0.717(0.091)	0.686(0.089)	0.684(0.077)	0.680(0.154)
RF_{EM}	0.748(0.074)	0.743(0.066)	**0.795**(0.076)	**0.716**(0.079)	0.686(0.135)
SVM	0.633(0.078)	0.577(0.097)	0.679(0.112)	0.600(0.102)	0.637(0.135)
SVM_{RF}	0.665(0.094)	0.720(0.076)	0.726(0.113)	0.645(0.064)	0.637(0.160)
SVM_{MI}	0.746(0.068)	0.729(0.070)	0.699(0.093)	0.673(0.068)	0.665(0.103)
SVM_{EM}	0.729(0.056)	**0.767**(0.048)	0.765(0.064)	0.700(0.046)	0.684(0.085)

that the results obtained for the classification and regression approaches, similarly, gave a small advantage to the SVM and RF algorithms with multiple imputation. The results associated with the baseline, a very simple average (naive) model, were the worst, as we may have expected for both classification and regression, so modelling produces significant improvements.

When attempting to evaluate the effect of missing data on performance by removing 25% and 45% of values from the publicly obtained complete dataset, we found that performance did not deteriorate in line with the percentage of missing data. In fact, for classification, the results improved slightly with the increase in missing data. Hence, this is an important conclusion, as it is often believed that missing data may have a noticeable negative effect on models, yet in our scenario of MCAR, missing data appears to have no noticeable effect on our ability to predict accurately. This is in line with the good results we obtained in the context of our datasets having up to 50% MAR data.

SVM with multiple imputation by chained equations and by EM are consistently associated with the top 5 best average ranks for both regression and classification. However, overall,

Table 6.5: Comparison of Accuracy mean values for each prediction system for the second school datasets. The standard deviation is in brackets.

	EB	PT	IS	SM	FM
SimpleAvg	0.732(0.047)	0.637(0.105)	0.499(0.137)	0.637(0.051)	0.797(0.020)
Rpart	0.788(0.041)	0.655 (0.057)	0.678(0.081)	0.743(0.034)	0.900(0.057)
Rpart_{RF}	0.786(0.028)	0.698(0.044)	0.673(0.070)	0.729(0.035)	0.906(0.054)
Rpart_{MI}	0.792(0.033)	0.679(0.052)	0.665(0.057)	0.728(0.033)	0.889(0.049)
Rpart_{EM}	0.792(0.0297)	0.721(0.056)	0.657(0.0661)	0.730(0.033)	0.890(0.047)
C5	0.797(0.038)	0.693(0.060)	0.655(0.089)	0.753(0.061)	**0.912(0.050)**
C5_{RF}	0.781(0.0438)	0.700(0.059)	0.632(0.110)	0.738(0.061)	**0.912(0.050)**
C5_{MI}	0.787(0.031)	0.689(0.049)	0.646(0.056)	0.745(0.046)	**0.912(0.050)**
C5_{EM}	0.791(0.030)	0.705(0.055)	0.644(0.075)	0.733(0.050)	0.911(0.050)
RF	0.805(0.043)	0.693(0.087)	0.680(0.091)	0.767(0.052)	**0.912(0.050)**
RF_{RF}	0.751(0.036)	**0.728**(0.034)	0.634(0.063)	0.764(0.055)	0.900(0.057)
RF_{MI}	0.767(0.049)	0.698(0.073)	0.675(0.066)	0.769(0.050)	0.896(0.060)
RF_{EM}	0.772(0.032)	0.719(0.075)	0.656(0.078)	**0.779**(0.051)	0.9(0.0542)
SVM	0.737(0.042)	0.702(0.106)	**0.693**(0.062)	0.772(0.061)	**0.912(0.050**)
SVM_{RF}	0.810(0.031)	0.715(0.097)	0.675(0.063)	0.778(0.062)	**0.912(0.050)**
SVM_{MI}	0.811(0.030)	0.699(0.084)	0.690(0.065)	0.761(0.052)	**0.912(0.050)**
SVM_{EM}	**0.813**(0.035)	0.726(0.0731)	0.681(0.063)	0.763(0.055)	**0.912(0.050**)

cliques in the critical difference diagrams tell us that the methods are not performing very differently to one another, with cliques showing no statistically significant difference between a number of algorithms. In this sense, our conclusion is that modelling performance by either regression or classification, and with any of the leading algorithms (SVM, RF, Rpart), can produce good results. Nevertheless, the ensemble approach together with multiple imputation and SVM or RF is novel, and produces consistently good results, so should be considered.

6.5 Summary

The primary goal of this chapter was to predict student performance at module level. However, it was also important to understand how best to apply the available algorithms in the context of missing data. To this aim, we experimented with multiple imputation in combination with an ensemble to improve prediction outcomes. We believe this is important, since very little is known about how to handle missing data in the educational data mining

Table 6.6: Comparison of Accuracy mean values for each prediction system for the publicly available datasets. The standard deviation is in brackets.

	Math	Math 25%	Math 45%	Por	Por 25%	Por 45%
SimpleAvg	NA	NA	NA	NA	NA	NA
Rpart	0.864(0.066)	0.879(0.048)	0.877(0.051)	0.845(0.043)	0.856(0.035)	0.852(0.040)
Rpart_{RF}	NA	0.885(0.058)	0.874(0.055)	NA	0.853(0.034)	0.873(0.032)
Rpart_{MI}	NA	0.878(0.046)	0.877(0.047)	NA	0.848(0.038)	0.856(0.032)
Rpart_{EM}	NA	0.879(0.054)	0.878(0.049)	NA	0.849(0.038)	0.859(0.024)
C5	0.854(0.042)	**0.887**(0.056)	0.885(0.065)	0.855(0.043)	0.859(0.048)	0.853(0.049)
C5_{RF}	NA	0.854(0.047)	0.862(0.047)	NA	0.852(0.041)	0.841(0.040)
C5_{MI}	NA	0.873(0.051)	0.876(0.040)	NA	0.852(0.038)	0.865(0.021)
C5_{EM}	NA	0.875(0.047)	0.879(0.043)	NA	0.850(0.038)	0.863(0.024)
RF	**0.869**(0.057)	0.869(0.044)	0.885(0.050)	**0.858**(0.039)	0.844(0.059)	0.861(0.026)
RF_{RF}	NA	0.882(0.053)	0.856(0.047)	NA	**0.873**(0.035)	0.863(0.045)
RF_{MI}	NA	0.881(0.044)	**0.886**(0.040)	NA	0.868(0.040)	0.880(0.023)
RF_{EM}	NA	0.875(0.052)	0.881(0.040)	NA	0.863(0.044)	**0.882**(0.025)
SVM	0.854(0.069)	0.877(0.0577)	0.836(0.061)	0.855(0.023)	0.861(0.036)	0.853(0.020)
SVM_{RF}	NA	0.872(0.055)	0.867(0.082)	NA	0.861(0.028)	0.848(0.040)
SVM_{MI}	NA	0.875(0.053)	0.871(0.050)	NA	0.872(0.027)	0.875(0.008)
SVM_{EM}	NA	0.873(0.047)	0.869(0.042)	NA	0.872(0.028)	0.869(0.019)

field. We learned from this novel study that ensemble approach combining with SVM and RF with multiple imputation could lead to potential outcomes.

In the next chapter we address a management study of the module choice problem from the standpoint of both of the students and a number of key staff members at UEA. We also address how to make use of any knowledge derived from the overall performance predictive models.

Chapter 7

From data to decisions - a management perspective

7.1 Introduction

As Rebecca Eynon has argued, *"the seemingly simple act of using numbers to describe the incredibly rich and complex process of how we learn could result in a range of consequences that vary from individual to individual, and thus decisions about how we want to develop and support such practices need careful consideration"* [272, p.408]. In this chapter, we investigate how to utilise data-driven models in the management of higher education (HE) institutions. For this purpose, we design a survey questionnaire and a number of interviews to understand student views. We also investigate how to utilise the available information from the university data warehouse and the data mining process to improve student outcomes in the context of a HE institution. We carry out several interviews with some of the key stakeholders to understand barriers to, and enablers of, change. We then analyse the collected data and propose recommendations for the final system.

Data can be described as an illustration of facts that can be collected, recorded and employed as a base for decision making [273, 274]. Gradually, data have become essential to both the theory and practice of higher education [275]. This use of data can be seen as part of a wider process of the 'datafication' of education [43, 276], in addition to many other aspects of society [277, 278]. Datafication has been defined as the ability to transform every aspect of life into computerised data, and to turn this data into something valuable (more detail in Ayankoya et al. [279]). Typically, higher education devotes much attention to two concerns

related to data. The first concern is data use by governments, government agencies, higher education institutions and their employees in the governance, management and evaluation of higher education. The second concern is data use by media organisations to produce 'league tables' of institutions and courses, which are seen as informing prospective students to allow them to select their institutions and programmes (see e.g., Ordorika and Lloyd [280]; Bougnol and Dula [281]).

Nevertheless, higher education has given much less attention to feeding data back to students within courses; for instance to aid elective module choice. As Neil Selwyn [275, p.71] has explained, students have 'data done to them' rather than being enabled to 'do data'. In this chapter, we use evidence from two schools that are associated with one British university to explore the implications of feeding back data, in the form of personalised predicted grades, to students. If, as Daniel [276] argues, predictive analytics are capable of offering institutions superior decisions and actionable insights built on data, is the same likely to be true of the students studying in those institutions? In spite of the rapidly increasing volume of educational big data research, higher education institutions have paid little attention to what the different social actors actually do with the outputs of their data-driven models and the resulting predictions, and how these data-driven decisions utilise and perhaps feed back into the creation of new data sets. As Donald MacKenzie [282, p.275] has written, in his ground-breaking study of the use of models in financial markets,

> *when confronted with a theory or model it is natural to ask: is it accurate? Keeping performativity in mind reminds us to also ask: if the model is adopted and used widely, what will its effects be? What will the use of the model do? (see also ONeil, [283] on the effects of algorithmic prediction).*

This chapter's main research motivation is to explore and present the potential effects of utilising available information in HE institutions and sophisticated data-driven models to inform student choice. This means using models to make data endogenous, rather than exogenous, to the HE system. We investigate what information the students or the academic staff believe is needed for a well-informed module choice and how the students report making their module choice.

The rest of this chapter is organised as follows: section 7.2 describes the findings of the study; section 7.3 includes the discussion of the results; lastly, section 7.4 summarises the chapter. The study's set-up has been explained in the Research Methodology Chapter in section 4.6.

7.2 Findings

7.2.1 Staff members' and students' perceptions of module choice

We can establish an initial picture of module choice based on the students' questionnaire survey responses. The responses provide important contextual information for the more nuanced qualitative interview data. The sample was reasonably representative of the population: 51.3% of respondents were male compared to 48.68% female respondents. The majority of the participants were home (UK) students (68%) but the sample also included international students from a number of other countries.

The current undergraduate students reported that they chose their optional modules mostly based on the information provided about each module by the school of study (76.7%) and on the opinion of the previous cohorts of students who had taken the same module (50.7%). The three most widely reported criteria taken into consideration while making their decisions were (in declining order): intrinsic interest in the module's topic, the type of assessment, and the student's (self-assessed) expected academic performance. It is therefore unsurprising that the majority of students (86.9%) were interested to know the predicted mark of their current modules and 77.5% thought that knowing their predicted mark in advance may have affected their optional module decisions. However, students identified a wide range of information that they believed would assist in module choice, including:

(a) An average mark based on the past few years of student marks.

(b) Personalised predicted student satisfaction rate based on students with similar personal characteristics.

(c) General satisfaction rate of students who took the same module in the past few years.

(d) Personalised predicted mark based on previous students with similar personal characteristics.

(e) General career opportunities associated with the module.

In what follows, we present the findings from the interviews on certain themes to help the readers easily access the findings and understand them. We should note that since the interviews were conducted anonymously, the students are referred to by numbers (1-28) and the staff members are referred to by letters. We have grouped student and staff responses in terms of three issues: the interaction of student decision making with the choice architecture offered by the school and the university; the moral and ethical concerns of students and staff about predictions; and staff and student attitudes towards the personalisation of information

regarding choice. We chose these three issues as our themes based on their importance in the literature that we have discussed in Section 2.10. In addition, these themes will assist us in producing our third contribution, which is how to utilise the knowledge derived from the exercise in the educational context both from the point of view of the students and the institution.

Decision making and the choice architecture

Choosing optional modules is typical of many decision-making situations we confront with a clear *"choice architecture."* We are given a menu of options (in this case, possible modules) and a set of rules to which we must conform (e.g., the number of modules we can choose, any restrictions based on requirements, any limitations on the specific combinations possible). However, we are also provided with some information concerning each menu item. This information may be directly concerning the module itself (e.g., the form of assessment or the curriculum it follows) or it may be about the experiences of previous cohorts (e.g., their performance, 'satisfaction' or subsequent career achievements). In our case, however, we are *hypothetically* providing a more refined form of information, personalised to the individual student and presented as a *prediction*, albeit one based on past data, rather than a retrospective account of prior performance.

A behavioural nudge is *"any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives"* [92, p.6]. Thus, this aspect of behavioural science can encourage individuals to alter their current ways of performing tasks, but maintains choice — something termed Liberal Paternalism [284, p.3]. Behavioural nudges often take the form of the provision of particular information to frame or anchor a decision. A personalised enrolment recommender system, such as the one we are exploring here, can be seen as a form of behavioural nudge. By providing information, in the form of a predicted mark or grade, the recommender potentially alters the student's behaviour, but does not restrict the student's choice in any way. However, to investigate this issue, we need to understand module choice from the student's perspective.

In interviews, we found that the majority (15 out of 28) of students reported prioritising their interest in, and passion for, the topic of an optional module over achieving a higher mark or other criteria. Typical comments from this group of students include: *"I do like learning rather than just focusing on getting high mark,"* or, *"If I was passionate about it [the optional module], it [the low predictive mark] wouldn't affect me, I wouldn't have thought it [the low predictive mark] would affect my decision, no"*. Student 26 stated, *"if it's not a very interesting module I personally don't think that would persuade me to do it[optional module] if it's something I really don't want to do, I won't do it I don't think"*. Student 5

argued as follows: *"If it was a module where I think it would be useful and interesting, I feel like I'd be more inclined to take a hit on a grade."* They would rather *"go for something that I would actually be able to try and pay attention to than just do something for the sake of getting a better grade in it."*

Students also reported using other criteria to inform their choice of module. Some students prioritised employability or future plans over achieving a higher mark. For example, one student said

> *I then changed my choices mid-semester, because I decided I wanted to apply for a PhD and I needed to do something that [was] more appropriate for that. I had to change my choices. I am fairly certain I am not going to get as good a mark now. But, the modules have been interesting and beneficial. (Student 16).*

Another student stated,

> *If I want to go in software development, I know the software engineering module is a must-do, because everything in that module is what employers look for. So if I was predicted a 2:1 as opposed to a first, then I'd still take it, just because I guess the learning aspects of it just throw away the mark, essentially. Because you get more from it, in that respect. (Student 12).*

Nevertheless, most students did report some interest in marks as one criterion among others in module choice.

Students did not necessarily see these criteria as discrete. Rather, criteria could interact with each other. For example, some students reported that, if they were interested in the subject matter of a particular module, then they would obviously enjoy working harder at it, and eventually would gain a better mark. For instance, Student 5, considering a module for which she had a high predictive mark, commented that: *"I ... don't really find it [optional module] as interesting and find it hard to pay attention in lectures, then that means overall I could end up doing worse because I am less interested"*. Student 4 remarked, *"I still think, because I'm interested in machine learning [an optional module], I would still pick it and try and obviously beat 40, 50 [low predictive mark]."*

Both students and staff argued that performance criteria (marks) were particularly important for some students in some situations. Specifically, we found that a significant number of students value the predictive mark information when they aim to achieve a higher outcome and when they are undecided about their module choice and are seeking 'tie break' information. For example, some students made the following statements.

> *If I was predicted anything underneath a 2:1, I probably wouldn't do it. I very honestly wouldn't do it. (Student 4).*

> *I personally chose my modules for this year [Year 3] based on how I thought I would perform. I originally picked things I knew I would do well based on feedback from other people. (Student 5).*
>
> *I originally picked [optional modules] based on how to get the best mark. In that case, it would have been useful to me to see that I was going to get 80% on something versus 60-something on a different one. Then I would go for the higher mark. (Student 6).*
>
> *My proposal of getting this degree is to get a high mark and just do really well and get a good degree. (Student 8).*
>
> *If it [recommender system] said, Oh, we think if you do this module, you're going to get 58, I think that would definitely put me off. Whether it's actually true or not, I think it would. Even if the module looked really fun and really engaging, I think it would definitely put me off, despite my best intentions. (Student 7).*

The following students' comments support the 'tie-break' approach to the predicted mark:

> *I think people do pick it [optional module] just based on the grade if [it is] their last choice or if they've got nothing else they want to do.(Student1).*
>
> *If I know I'm going to just pick it because I don't know what else to pick then I would see if I could get a higher mark in something else.(Student2).*
>
> *It [the low predictive mark] wouldn't put me off, I would say, unless I was already a bit undecided on it. (Student 3).*

Staff, rather than students, emphasised another situation in which marks were particularly salient: students at risk of failing. One senior lecturer (Academic JA) argued that predictive marks *"would help students whose performances are borderline or boundary (at risk), to perform a bit better."*

In general, however, staff tended to promote a 'balanced' view in which students should weigh up a number of criteria. For example, one school's published guidance argued that students should ask themselves the following questions before choosing modules:

- Am I **interested** in the subject of the module? Does the subject matter intellectually stimulate me?
- Do I have **practical** work experience or other experience that makes the subject matter of the module or the skills developed particularly relevant to me?
- Will the module help me to achieve my **career** aspirations?
- Will the module facilitate interaction with **employers**?
- Will the module help me to develop **skills and competencies** that are valued by employers?
- Will the module help me to gain **exemptions** from professional exams?

- Does the module play to my academic and practical strengths enabling me to achieve the **best possible marks** for my degree? [2014/15 module guidance for the second school of the study]

Some staff, including senior academics, expressed concerns that a predicted mark would focus excessive attention on academic performance criteria alone.

Students exhibited a range of reactions towards the predicted mark, in particular where it was, in their opinion, particularly high or low. For example, where students received a significantly low predicted mark, they reported that this could provoke or stimulate them to take a more proactive stance. These students sought to investigate in advance the issues that caused the high or low predictive mark, identifying and improving certain skills or knowledge that would be required for their module choice. This observation is evidenced by the following students' remarks:

> *It [the low predictive mark] would influence me in a better way . . . because it would give me that push to go and like speak with the module organiser and know about this module in a different prospective . . . I would be more aware like maybe there is a reason this [enrolment] recommender system told me that you might have a bad grade. (Student 22).*
>
> *If it's something you really want to do it would probably potentially make you more proactive to try and work out where you can do better. (Student 26).*
>
> *Okay, for me, first of all I would start to think why is it like that? Where is the reason why they would predict such a low score even though I went higher? I think in my opinion I would start to look where I actually could fail. Maybe there is something that I'm not really comfortable with and that is, like, decreasing the mark. Maybe if I start learning even before, strengthening my basics of that thing, maybe I would improve that. (Student 8).*

The second observed reaction is that students are willing to consider alternative modules that would be suitable for their degree outcome if they saw the predictive mark is significantly low. This was inferred from students' comments such as, *"that would be quite a positive influence because if I saw I had a low predicted, less than I expected, I would maybe ask myself why. Maybe I would think I would struggle with that, and maybe look at different ones to take instead."* (Student 23).

The third observed reaction, however, indicates that there are students who thought that a low predictive mark might discourage them from taking a module, in spite of finding it very interesting. For instance, Student 26 stated,

> *It would discourage you from doing it [optional module] even if you might then get a first, but don't know what you're going to get if they're [the*

recommender system] going to say you will just pass.

Reactions to a significantly high predicted mark also varied. Some of the interviewed students argued that they would need to work particularly hard to achieve this predicted mark and that the high predicted mark would motivate them to do this. Others, however, believed they would be relaxed when taking modules with a high predicted mark and may even not put in as much effort as they would usually do. Although they were aware that this was a risky approach, they believed that this would be their attitude to knowing the high predictive mark in advance.

Moreover, it has not escaped our notice that students and staff members indirectly indicate how the transparency (i.e. information communication) of the given choices by the proposed enrolment system could affect their decisions. Transparency, in this context, is defined as a user understanding of why a specific suggestion was given, particularly the relation between the input data and the outcome [285]. Students pointed out that the transparency of the proposed system might help them to trust the given recommendations. Students and staff members explained transparency based on their knowledge and the information given to them about having such a system.

Particularly, students were concerned about the data used to generate the predictions. For example, a student believes that if the enrolment recommender system explains that the suggested module is not suitable for him based on low grades in previously completed modules that are relevant to Maths, then he will disregard it because he is aware that Maths is not his strongest subject. Another student mentioned that it is important to know the exact information that has been used by the recommender system to produce certain recommendations or the reasons behind suggesting these particular modules. The reason for this is that she believed that if a student had a tough year that affected their academic performance, they would need to make sure the enrolment recommender system gave them the right suggestions for the next year. In her perspective, knowing the input of the given predictions would help those students to decide whether they wanted to accept the given suggestions or not. Similarly, student 28 noted that he was diagnosed with dyslexia, so he felt unsure whether the system would take his condition into consideration, unless it was stated clearly what inputs had been used to create the provided recommendations.

Accordingly, Academic JA commented that there are a lot of issues with the quality of the current data. For example, the data on student satisfaction and employability is poor; it is based on samples, self-reported and 'skewed' by over- and under-representation of particular groups. In his opinion, this could result in misleading predictions.

We found that staff members were concerned about another two points that also related to

the transparency of the proposed system. The first was the complexity of the algorithms. For example, Academic D strongly believes that it is challenging and complex to build such a system that would be able 'to get the discipline's specific themes coming out' with high accuracy. In his opinion, this is similar to the 'career choice recommenders' that, to his best knowledge, took 20 years of development to become effective. Second, staff were concerned with how predictions would be communicated, how to best word the results of the predictions and how much explanation should be given to students. Academics JA, JN and D all pointed out that it is challenging to both justify the predictions, find the precise words that reflect the predictions, yet keep communication short and clear enough while not misleading the students.

In summary of this issue, we found a range of different responses. Students displayed quite complex attitudes towards the prospect of a recommender, with different students using different criteria to inform their choice of module. Variation involved whether the student is interested in a particular criteria or not, how much he/she uses a particular criteria in their decision-making process, and how much a specific criteria could affect his/her decision. Therefore, we found that it is unclear whether the university could use the recommender to nudge students without significant and unpredictable impacts. We would need to perform further experiments to test the student responses. Students agree in expressing their concern regarding the transparency of the enrolment recommender system as they understand that the given recommendations will be a result of the information that has been fed to the system. From their perspective, if the university did not feed the system with enough information, particularly the personal information of the students, the system might recommend unsuitable choices. Next, we will discuss the ethical side of the predictions' transparency, and consider how the recommender might affect the student sense of responsibility for the choice of module.

Responsibility for choice and ethical questions

An enrolment recommender system can be seen as a tool that supports students through the process of choosing a module. In theory, students should be aware that they have the choice to accept, investigate or reject the given recommendations. In the liberal paternalist paradigm, students retain choice and, importantly, responsibility for that choice. However, there have been concerns in the literature about the impact of recommender systems (and other 'big data'-based decision support systems) on individuals' sense of the locus of control, and therefore their responsibility for their choice [99]. Students suggested a range of ways of potentially using the recommender and exhibited a range of attitudes towards their responsibility concerning module choice. Specifically, we observed those who clearly took full responsibility for their choice, those who deferred to statistical norms in their decision,

and those who deferred to technology when making decisions.

The majority of students (21 out of 28) clearly believed that they were primarily responsible for their choices and they would use the proposed enrolment recommender system, if at all, as a support tool and an additional information source. This means their decisions would not be based only, or even mainly, on the system. The following students' comments illustrate this point:

> *I would likely do some research on top of that [the recommender system's suggestions], just to consolidate it [his module choices] in my mind. (Student 7).*
>
> *... an informed opinion is always best if you come from a range of different resources ... I definitely would use other resources as well. (Student 11).*
>
> *I think it would probably be a tool to help me with my decision making, but it wouldn't be the primary thing that drove my decision making. (Student 12).*
>
> *I know that my degree is down to me. (Student 13).*

These students are also aware that the recommendations might not turn out to be 100% accurate or might not match what they would like to achieve from their degree. Because a software tool such as an enrolment recommender system may not be reliable, they argue, they would prefer to use a wider range of available information in addition to the proposed enrolment recommender system. One student mentioned 'exploring all the available options,' using 'a range of sources,' to make sure he made 'the best possible decision.' One student even expressed a fear that the system may divert students from modules that could be of interest to them.

Some of the students claimed that the enrolment recommender system would provide them with a secondary assurance of their module choices, primarily made using other criteria. Other students argued that they would use the enrolment recommender system out of curiosity, to have an initial idea about the elective modules in which they were predicted to perform well.

A smaller group of students expressed a very different attitude towards responsibility for module choice, deferring to a statistical result that was seen as 'objective.' There was some evidence that students who deferred to the numerical or statistical nature of the recommender's prediction might do so because of their degree background and, in particular, their familiarity with the underlying algorithms. For instance, a business management student noted that if the recommendations considered raw data and current marks, she would find it hard to disagree with the results of 'pure statistics.' Another student mentioned that if she observed the general average marks for the past years, then this would help her

in choosing her modules as she has 'faith' in statistics. Students who expressed these views also came closest to deferring to the recommender in module choice. Staff, even when they were positive about providing students with data-based predictions to enable them to make an informed choice, were still concerned about the lack of transparency in the process — a phenomenon known as 'black boxing.' They fear that students might focus heavily on numbers and that may impact the quality of their decisions. One senior academic argued that,

> *you just have to be aware that there's a whole load of messy complex stuff underlying that generation of the numbers, that involves how students work, how their work is assessed . . . Placing too much analytical emphasis on getting 68 in one module, compared to 62 in another, you begin to perhaps over-rely on the validity of those numbers, I think. (Academic N).*

A small number of interviewed students appeared to rely more on technology than numbers, leading them to accept the given recommendations relatively uncritically. For example, a student claimed she would willingly accept the suggestions although she could feel cheated if she did not achieve the predicted mark. Further, she argued, *"because you're only going to take a module really if you think you're going to achieve well in it. That's just natural"* (Student 3). It seems like a stochastic system to her; she knows that she might not win but she will still do it.

Students' perceptions were not restricted to their own attitudes, but extended to others. Some students expressed concern that other students would follow the recommendations of the technology without either a) considering the required efforts or b) challenging themselves to achieve the predicted mark. What is more, they expressed concern that the recommender system may perhaps steer them towards taking similar modules or a specific path. For example, those students pointed out that the enrolment recommender system might push them to concentrate on the predicted mark and the type of assessment since some students prefer modules with coursework rather than exams. According to those students' view, this may eventually allow them to doubt their abilities, or prevent them from looking at the bigger picture. They might fail to see how beneficial the optional modules would be for their future career, or for broadening their university experience and education, or for experiencing different modules that are more challenging and that push them out of their comfort zone.

These concerns led some students to emphasise the importance of framing the recommenders' outcomes in an appropriate way. In particular, students were concerned that an algorithmically predicted mark needed to be clearly linked to and qualified by an emphasis on

student effort. Staff had similar concerns. Academic JA saw a danger in having an enrolment recommender system in the university as students could utilise it as an 'abdication tool.' Therefore, he feels, the university should make sure that the students understand that the recommender is a decision support tool that will not 'make the decision for them.' Specifically, he linked this idea to students who preferred a 'safe choice' and who tended to 'drift' though their degree. He argued, *"nobody ever makes a decision with full information, it doesn't exist"*; he thought that students should be aware of the partial nature of all information, but was concerned that a recommender system could obscure this fact.

Staff were also concerned with the legal implications of the recommender in an increasingly market-oriented sector. Both Academic JA and Academic JN (another senior lecturer who also is a manager of an undergraduate programme in one of the university's school) mentioned the framing of students as paying customers of the university. Consequently, they argued, students can come to expect a guaranteed first or a 2:1 with little effort. If those students felt misled by a recommender or did not achieve their predicted mark in optional modules, they might seek legal redress. Therefore, the university should be careful about what they promise students, particularly students who abdicate responsibility by over-relying on statistics or technology. Academic JN used the metaphor of a crutch:

> *We're [the university] a crutch for them. When something goes wrong, it's our fault, not their fault. It's about responsibility and blame. I would really want them [students], for themselves, to think and reflect, rather than using this as some kind of artificial crutch.*

Moreover, both Academic D (who is a lecturer and a director of learning and teaching in one of the university's school) and Academic JN agree with the concerns of the interviewed students regarding the possibility of the recommender system steering them towards a particular academic path that might not be best for them. Therefore, Academic D believes that it would be technically challenging to personalise the recommendations with the right weighting of the fed information. Academic D, drawing on his extensive experience as an academic advisor, argued that *"it quite often takes a little digging around"* to pull out the underlying theme that interests students, and that he usually finds that students are not very good at making links *"between the stuff they're interested in and the words they're seeing in module descriptions"*.

These concerns that an enrolment recommender system may heavily influence the students' independence in making their choices tend to lead to an emphasis on the wider context in which the predictions are presented and interpreted. One staff member (Academic JN), thinks the recommender system may have a more advantageous role if it can become part of that broader spectrum of different sources of information and advice, such as involving the

academic advisors in discussing the results of the recommender systems. Accordingly, in the next theme, we will present our findings with regard to what the interviewees think about the role of academic advisors and personalisation in the enrolment recommender system.

Academic advisors and personalisation

A major feature of modern recommender systems is that they can be 'personalised,' using demographic and other data to create an individual prediction, which is seen as offering a personalised experience for the students [286]. In practice, personal data is used to find close demographic and performance matches from previous cohorts in order to predict a student's grade; the prediction is less for the person and more for a type of person — people like you. It is, nevertheless, experienced as a personal prediction or recommendation. Many companies, such as Amazon and eBay, utilise a personalised recommender systems to add value to their business [287].

They achieve this personalisation by having a large data science team that intensively focuses on their data. However, they have not published independent evaluations of the value that these technologies add to their business [287]. Interviews conducted with students suggest that personal module recommendations are seen as supporting the student's academic trajectory more than general recommendations based on cohort data.

The students raised the following perspectives on personalisation. We should note that since the enrolment recommender system is currently not available in the university, these interviewees' perspectives are based on their expectations and on the information provided during the interview about how the proposed enrolment recommender system would function.

Initially, students suggested that the given predictions and recommendations could not be personalised if the university did not have enough demographic or other personal information to feed the proposed system. Some students argued that the university does not collect enough data that they consider relevant to module choice. Examples of these 'absent' data are students' preferred subjects or preferred future careers, and social or health circumstances that a student might have gone through in a specific year. Students feel that the lack of these types of information may affect the personalisation as well as the validity of the provided recommendations. Several students, such as Student 9, suggested that the university should add a questionnaire to the proposed module recommender system to capture additional information, such as which modules students have enjoyed in the past or what their future career ambitions are. Student 9 stated:

> *those job websites which help you decide what job you should do in the future. They [the employment questionnaires] can take some time but, for me, it's worth it because it helps you find exactly what you want to do.*

Equally, Student 28 pointed out that the reason he thinks collecting additional information is essential for a personalised automated system is because he did very well in the 'Graphics1' module although he did not enjoy it. Therefore, if the university does not feed the system with information such as his future preferred module topics, there is a higher chance that the system may suggest taking the 'Graphics2' module based solely on Student 28's past grade.

Students mentioned that having personalised predictions and recommendations might aid them in reducing the amount of time they spend annually searching for information before making their module choice, and in appropriately planning their academic trajectory. This is because they believe that personalisation means automatically taking into account the requirements of each individual course. For example, Student 14 explained that each course within her school has a different list of optional modules available and the prerequisites for each optional module vary depending on the student's course. Hence, Student 14 has to manually check all the prerequisites herself and cross-reference all of the timetabled slots to ensure things do not clash. This often requires significant information search activity from the student. Similarly, Student 27 explained that his course of study includes modules from two different schools and that the communication between these two schools was not great during the enrolment period. Thus, Student 27 needed also to manually arrange all his optional modules and make sure they would not clash with his compulsory modules. As a consequence, he ended up doing an Arts module that is not related to his studies and which he considers will not benefit him in his future career. Therefore, he explained, the proposed automated system needed to be on a university level, or at least connected between related schools, in order to be useful for students like him.

Some student interviewees pointed out that a personal recommender system would make them more aware of their academic limits. Some considered this point as negative while others counted it as positive. Student 17, for example, who viewed this point negatively, explaining that she believed that if the students knew their academic weaknesses, they might avoid modules that could take them 'out of their comfort zone.' She also stated:

> *It [the personal recommendation] might put doubt in them [students] if they want to go and do something else so they probably won't challenge themselves and go for something else. But they might go for what is recommended when they prefer something else. But they end up doing worse than what they would have.*

Students mentioned two reasons why it would be valuable to know that the proposed enrolment system would not recommend specific modules due to their poor grades in previous similar modules. First, in their opinion, this awareness would help them to be proactive

and to academically prepare themselves in advance if they decided to take this specific challenging module. Secondly, this information might help them to avoid this specific module if they preferred to take a module that played better to their strengths.

Some students mentioned that a personalised recommender could affect capacity management issues for a module. For example, a general recommender system may suggest the same module to every student, which could create overcrowding in the module's labs and lectures, which as a result would affect the quality of student learning.

Student interviewees were aware that personalised recommendations would not be sufficient, on their own, to support their module decision-making process. Many students felt strongly that they needed both a personalised enrolment system and general statistical mark data, in addition to the enrolment support that they currently have. In their opinion, this would be enough information to make a rational, informed decision without blindly following the recommender system. For example:

> *It's like there are always deviations to that [the personalised modelling with an algorithm] and exceptions. I kind of feel while the recommender system might help, I would still want to investigate, and make sure that I'm making the module choices that I want. (Student 24).*

The traditional way of personalising module choice information, and taking into account those 'deviations and exceptions', has been through the personal advisor system. Many students mentioned that they value having a discussion with their academic advisors who know them well enough, as students believe that these academic advisors will provide them with the individualised recommendations that they need. Considering the role of the academic advisor also raises the issue of discursive or deliberative models of decision making, in contrast to a simple calculative rational model or the focus on bias and 'automatic' decision making.

Many students were aware that, due to the large number of students, an academic advisor may not know all of his or her student advisees very well. Therefore, having a high-quality personalised automated system could provide some of the needed personalisation that an academic advisor may not be able to offer, especially if he/she does not know the student very well. Some students argued that if individuals are more engaged with the university, then there is a higher chance that their academic advisors will be acquainted with their educational background, strengths and weaknesses. As a result, this type of student will derive more benefit from the advisor during the current enrolment period than less engaged students. On the contrary, students perceived that shy individuals, or students who did not engage well, might benefit more from an automated algorithmic enrolment system.

The student interviewees also offered some insight into the interaction of the advisor sys-

tem with the proposed recommender. We observed from the interviews that two types of student behaviour were reported. First, there is a group of students who believe that they would need to speak to their academic advisor as well as consulting the proposed enrolment system. These students advanced a number of reasons for making use of their advisors, such as reassurance, as an additional source of information before they make their final decision, and as an essential support opportunity to familiarise themselves with a different and more experienced perspective that might assist them in making their module decision. For example, Student 7 stated,

> *I speak to my advisor as well, even if they are not related to the module, just to see what their feelings are. Basically just tap into as many different ways as I can before I come to a decision.*

A second group of students did not enthuse about speaking to their academic advisor if they believe that the advisor did not know them. They expressed concern that their advisors would indirectly push students towards their own personal academic preferences. Thus, those students were unable to trust their advisor's guidance. For instance, Student 15 explained that the reason she does not communicate with her academic advisor is that at their second meeting her academic advisor was still using her surname as her first name, which gave her the uncomfortable feeling that her academic advisor did not know her. Hence, Student 15 was not able to accept the advice of her academic advisor, although she was aware that the given advice might be valuable for her studies.

Students and staff both noted that specific lecturers can act as informal academic advisors and that students might prefer to discuss their module choice with their lecturers if they believe that they can understand them better than their assigned academic advisors. Student 21, for example, said,

> *I talk to my lecturers about it [the module choice] rather than going to my advisor because they know, I feel like they know me more because I see them every week.*

Equally, Student28 explained that he preferred to discuss his module choice with one of his lecturers. He believes this particular lecturer understands him more because they have been working on a project together.

Both staff and students were also aware that the quality of advising varied. Many described an ideal type of 'good advisor'. Academic H, for example, referred to the good academic advisors who are always available, good listeners, and able to provide the right support, combined with a bit of 'tough love' to get the students through their courses. Furthermore, Academic JN explained that good advisors are the ones who do not dictate to the students what they should be doing, but instead attempt to get them to reflect on themselves,

helping them to become aware of their strengths, weaknesses, preferences and dislikes, and to consider their preferred future career or industry. Students who are assigned to this first type of 'good' academic advisor are the fortunate ones. An automated system would complement and not replace the academic advisors for these students.

However, a second 'ideal type' was also recognised. This type of academic advisor is one who is not able to give good advice, or cannot get on well with their students. The students who are assigned to this type of academic advisors tend to not have a strong relationship with them. Some reasoned that having an enrolment system that provides information, support, and which could stimulate individuals' thinking may perhaps be helpful for these students. However, staff tended to see this as a poor outcome as students should have both systems. For example, Academic H argued that a personalised recommender system should not replace academic advising since it all depends on the student's context and that 'in an ideal world in the university we wouldn't have any of these contextual inequalities.'

Students pointed out that the benefit of having an academic advisor compared to a personalised enrolment system is that they can reach their advisors any time during the year, if required; they made this comment because they believed that an enrolment system would only be available during the enrolment period. Students mentioned that an academic advisor is important and cannot be replaced by an automated system, particularly in assisting them with solving module issues that might affect their studies. They argued that there is a chance that the personalised enrolment system would suggest odd or unsatisfactory module recommendations. Hence, they would need their academic advisors to discuss with them these types of recommendations, rather than solely basing their decision on the given information of the automated system.

Many simpler approaches to decision making assume that the decision makers, whatever other information they do or do not have access to, are fully informed about their own preferences. Staff do not necessarily share this view: Academic D, for example (as stated earlier), explained his concern regarding the complexity of employing personalisation features in such a system, as according to him, it would be quite challenging to draw out the underlying themes that interest students. Most staff used similar arguments, suggesting that the complexity of the different needs of each student meant that a personalised automated system with a more limited knowledge base would not be sufficient by itself, and that an academic advisor would be needed to complement such a system. This view suggests that staff, as well as some students, favoured an element of discursive or deliberative decision making in which preferences are emergent from the decision-making process rather than a pre-existing input to that process.

7.2.2 Staff members' perception of Chapter 5 findings

As presented in Chapter 5, we have analysed students' first year data. We were able to identify students with poor performance in terms of 'good honours' outcomes with reasonable accuracy. As part of this chapter, we have extended Chapter 5's study by investigating how to utilise the resulting information from data-driven models in the management of higher education institutions. We undertook this task by asking the staff members the following questions during the conducted interviews.

- How, from your personal and professional point of view, **should** the University act on those findings in terms of improving students' outcomes?
- What **could/should** the University offer to those at risk?
- Should assistance be offered to all students or only those at risk?

According to these questions, the following findings are presented in four themes.

The staff interviewees started by expressing their thoughts regarding whether the university should or should not act on discovered information. Academic G believed that once this information had been created, would be 'ethically unacceptable not to act upon it'. Further, he argued that the university requests that students pay roughly nine thousand pounds per year of study, and thus, if the students realised that particular patterns of their performance could be identified, they would have a reasonable case to expect this information to be shared with them. Further, he argued, they could reasonably expect it to be used to improve their outcomes and ensure that they could achieve their full potential. Drawing on his work experience and recent conferences he had attended on student analytics, Academic G argued that within three years or so, utilising the results of data-driven models to support students would be *"expected as standard."* He stated, *"I suspect you will begin to see it as a tick box on an application ... Like Wi-Fi in residences"*; he believed that students would consider it as a way to protect their investment in tuition fees. Academic JA agreed, adding that the university should act on this finding, to help the students *"to get the best value"* out their yearly tuition fees. Academic H saw the recommender as an opportunity to provide an equal chance for students to achieve what they would like out of their degree in terms of degree marks and useful knowledge.

The academic interviewees pointed out that acting on the findings has an 'enormous operational value', being particularly useful for certain groups of students, such as:

- students from disadvantaged or non-traditional backgrounds;
- students with learning difficulties or mental health issues; and,
- mature students who often have caring responsibilities.

Regarding the importance of improving the performance for these types of students, Academic H stated that *"the difference between 58 and 61 is all the difference in the world."* More specifically, Academics G and H explained that the university's goal under the government's widening participation agenda was to target the recruitment of students from postcodes with traditionally low higher education participation rates. This group of students, in particular, could benefit from close monitoring of their performance and receiving suitable interventions, which might prevent them from dropping out of university and could help to generate a sense of belonging in the university. Academic H argued that, in her school of study, there are a higher than normal proportion of students who are mature and with learning difficulties. This group of students find it much harder to achieve 2:1 so currently, they are *"in the 2:2 bracket."* However, Academic H explained further, *"if you're going to go into teacher training, the difference between a 2:1 and a 2:2 is getting onto the course or probably not getting onto the course, because teacher training, particularly here, is incredibly competitive."* Any assistance with ensuring the required grade that a recommender could provide would be particularly beneficial for this group.

A possible counter argument that was explored by some interviewees concerns the possibility that a recommender could promote 'grade inflation', which refers to the claim that achieving particular grades has become easier over time. Academic G argued that the university should not worry about this grade inflation argument, because the university is able to demonstrate positively the processes that they have undertaken to achieve the end results — for example, how they have *"gone through module by module, looked through data, understood where marking is out of kilter, understood where something isn't quite right, and actually resolved those problem areas."*

Academic N explained that students' performance is important because it could affect the university. The proportion of undergraduate students achieving 'Good Honours' degrees *"feeds into a couple of the league tables,"* and also reflects the general level of academic attainment. He also added that from his work experience, high performance of students is associated with *"the quality of student intake and level of academic staffing."* Hence, monitoring performance and acting upon the findings of analysis of performance is valuable to the university. However, he was cautious about relying solely on performance data, stressing the importance of the wider context of the school of study and the subject being studied. Students' performance should be compared with that of their peers in the same subject in other universities. Alternative reasons for poor academic performance, such as lack of engagement/attendance, should also be explored. Only after taking this wider context into account, should firm decisions about additional support be made.

In contrast, Academic C believes that there is no need to act upon those findings except if

those students are failing. This is because it is going to affect the university's drop-out rate, which he saw as indicating a waste of the university's resources. In his opinion, achieving a good honours degree is not essential compared with having a good education and experience at university. He argued that the university should be concerned more about providing modules that the students would enjoy learning from, instead of providing an intervention to boost their performance. He also did not believe that performing poorly in a number of modules would affect the students' future employability, as there are other important factors that count towards employability, such as having a particular skills or the ability to work well with within a team. Moreover, Academic D and Academic JE pointed out that the university should study first the drivers of that poor performance before deciding to act upon them. For example, Academic D argued that, if students were performing poorly due to personal and family issues, then the university could not support them in dealing with their problems at an emotional level; hence, they would not need to act upon those findings. Although for this particular point, Academic JA disagreed by arguing that the university does provide counselling and financial support for students with family and emotional issues, in order to help to improve their performance.

Interviewees then expressed their thoughts regarding how they should act upon the received findings. Academic JA thinks that any assistance the students received, based on the analysis of the data, should be an offer and not compulsory. This is for several reasons. First, providing compulsory activities for selected students would be difficult to staff and timetable. Second, students were seen to be more likely to respond well to assistance if they believed the university was concerned with helping them to get the best value out of their tuition, rather than *"forcing"* them to improve outcomes. Third, Academics C and N added, the students may simply not want the intervention, because they intended to achieve different goals during their life at university, such as taking modules that they would enjoy but not necessarily obtain a high mark in, playing certain sports, becoming involved in the social curriculum, or other sorts of opportunities.

Academic G showed his concern over the consistency of handling the knowledge derived from the data-driven models. He believed that the university should have a systematic approach of how *"to accept, organise, manage and understand the resulting information, and then staged interventions with students"*. He added that the university should have a clear *"understanding of what an adequate intervention looks like,"* before they offer any assistance to students. Academic N argued that if analysis showed students were at risk of failing, and the results could be linked to the characteristics of the students, then the university might need to make more substantial changes. These might take the form of revising their admission standards or investigating whether there were any structural features of the course

of study or assessments that were leading some students to fail.

Next, the interviewees discussed what they could or should offer to students at risk. Academic JA suggested that a reasonable basic approach to start with is to privately communicate with students who have been predicted a poor performance, to inform them of the prediction, and how this may affect their future employability chances. However, since the prediction is not a guarantee to be associated with high accuracy for all students, Academic JA emphasised that there was a need to explain to the students that the prediction was based on past performances by similar students — *"students like you in the past."* Academic H also believed that effective assistance should start by having *"an open dialogue"* between teachers and students, for example, asking a student at risk *"Have you managed to do the reading? Have you got any questions?"* She preferred this approach, because she believed that each student's needs were different. *"The missing pieces in the jigsaw puzzle are different for each student, and the instructor is only able to assist them based on their needs."* However, Academic H also explained that there were limits to this communication approach in helping students, especially where modules involved a large number of students, such as 250 or 300.

Moreover, Academic H stated the university could offer additional tutorials, or further chances in formative work, or assistance with finding learning sources. Academic JE believed that the lecturer could direct weak students to the Student Support Services team and their personal advisors, as she believes they are currently providing sufficient assistance. She also suggested having a tutor for each year of the course of study, with a particular remit to assist weaker students, since students need different kinds of support for each year of their study. Alternatively, Academic JA suggested that the university could offer a summer school model to those students predicted to be at risk of failing, although he believed this may not be the best approach since not all students would be *"failing the same way, or in the same places, or for the same reasons, or the same issues."* Academic JA argued that the best assistance from his perspective may be to have a personalised automated recommender system that would be able to recommend different solutions, such as extra classes, regular meetings with a personal advisor, or somebody with professional training in supporting particular types of learning. However, he explained that the recommended solution should be appropriate for the student's issue. For example, in the business school, students at risk could face a range of issues, such as a lack of motivation, difficulties with modules that involve logical/mathematical topics, the inability to write clearly, or an issue with group work and so on. Any recommender would therefore need much more information about the student.

In contrast, Academic D thinks that instead of providing assistance, there should be tough-

ness towards certain students who struggle. A least some students, he argued, should be told that they *"are not in a fit state to complete a degree."* This is because he believes that students who are struggling throughout their course, taking a lot of time and *"not succeeding at all is not productive for anybody involved."*

Last, the interviewees expressed their thoughts on whether the intervention offer should be offered to all students or only those at risk. Academic JA pointed out that this point is a morally difficult argument. He explained further that from an equality point of view, *"everybody payed the same, so everybody should get the same"*, except for some scholarship students. However, he believes that since education is a public good, then it is acceptable that some students obtain more resources than others, as long as they are actually attaining their educational goals. He offered a metaphor to explain:

> *If you're in a restaurant the menu is an offer, and people can pay the same amount of money, but have a very different meal. If you don't want to eat everything that's on your plate, you don't have to. If you don't want help or assistance, then you don't have to have it. (Academic JA).*

In addition, Academic JA explained that offering help to every student is against the idea of personalisation, *"as personalisation of education is about offering people what they need, not the whole menu."* He and Academic H added that often in reality two types of students obtain the most assistance in education. First, there are students who have very high performance, since they demand a lot of attention and have good *"help-seeking behaviour."* Second, there are students who are at risk of failing, since they draw attention to themselves. In contrast, the middle students often receive the least aid in their education as they perform well enough to avoid attention for being at risk of failing but not well enough to attract attention as high performers. This may of course also suit the students as they may not engage strongly, or they might view university as a *"rite of passage,"* having other more important interests that they want to achieve instead of high performance.

Academic G and Academic JE added that in reality, the university cannot afford to provide additional assistance to every student, and the offer should go to *"where it's going to be most valuable."* This is because all students already have access to their personal advisors and the Student Support Services team.

7.3 Discussion

Interviews with students and staff indicate a generally positive attitude towards an enrolment recommender system based on academic performance. Staff members showed an acceptance of having such a tool as they thought it could help them mainly to focus on

improving the management information around learning, teaching, and the student experience. Students welcomed the prospect of an enrolment recommender type system as an additional source of information to support module choice and, potentially, as a way of easing access to information to support module choice. As Student 15 argued, the required information that would help her in making the decision about her module choice is currently located in many different places, such as the Blackboard (the university's Virtual Learning Environment), the university's student management system, emails, the university's main website, and in material distributed at the annual 'module fair.' Therefore, it required an enormous amount of time to go through all of the sources and check the needed information to make her module decision; this was made more onerous by her current study workload, as the enrolment period coincided with the student's busiest time of the year. As a result, Student 15, as well as many other students, believed that having a recommender system would catalyse the university to include all the necessary information in one accessible location, which would be more convenient and would help students to be more receptive to module choice.

We believe from our findings that the university could build such a system to help them effectively utilise the available information, and as a result nudge students to choose modules that would aid them in achieving better academic outcomes. We also should note that there is a debate regarding whether nudging is ethical for a free society or not (more detail can be found in [284, 288]). Our findings suggest that, in our case at least, there is no neutral choice architecture for student to make fully independent decisions regarding module choice. This is due to the fact that the universities are responsible for providing the students with all the required information to make their module choice. Students do not have enough information to make this kind of decision by themselves. One of the students supports using this kind of behavioural nudging:

> *I don't think it's [recommender system's prediction] a risk because in school when you have your mock GCSEs and that would be like your predicted mark it didn't really affect the way I did my real ones.*

Besides, the enrolment recommender system should not lead students to make the 'wrong' decision, because they will always have the choice to accept or reject the given recommendations. In light of the findings above, it is clear that the interviewed students were generally aware that they should not blindly follow the system's recommendations. Because the student interviewees exhibited complex and varied criteria for, and approaches to, module choice, the provision of predicted grades could affect their module choice in a range of ways. Therefore, we are aware that we need further experiments to test their responses at a behavioural level.

The lack of transparency in the process by which the given predictions and recommendations were created was a concern for both staff and students. Therefore, if the university decided to implement such as system, they would need to very carefully word the actual predictions, clarifying and simplifying the framing and the presentation of the information for students to understand. The transparency of an enrolment system could be increased by following a discursive approach through involving human interaction during the enrolment process by thoroughly discussing and interpreting the results of the recommender system with the students. Providing personalisation features might add great value to the given predictions, although most interviewees believed that human advisor interaction would still be required.

Could feeding information, based on the university's available data, back to students via an enrolment recommender system (providing 'a little knowledge'), be regarded as 'a dangerous thing?' Next, our discussion will address two points: 'feeding back with little knowledge' and the risks involved.

Our findings suggest that students value every additional piece of information during the enrolment period. This is because, as students pointed out, an informed decision is associated with a variety of different resources. For example, Student 15 stated,

> *I think that everyone who I've spoken to about module choices would like more information about it and if this [enrolment system] can provide that extra bit of information then that would be very helpful.*

Also, Student 25 mentioned,

> *Well, an informed opinion is always best if you come from a range of different resources. Obviously, I've learned that through coursework research so I think there isn't such a thing as having too many points of information about something.*

In terms of the student decision-making process, there is an indication that providing students with such knowledge might affect their decision regarding module choice. However, we observed that each student followed one of the three different approaches of decision-making process (explained earlier in the literature chapter 2.10.2). First, there are students who perform the reflective system method by looking at all the available information and using reasoning and logic to make their module choice against established criteria and preferences. For these students, the prediction represented an additional data point that would be taken into account in the decision-making process. Second, there are students who seemed to follow the automatic system approach. The decisions of this type of student could be greatly affected by predictions of performance, being effectively nudged by the provided knowledge. Third, there are students who follow the discursive approach, seeking to discuss the recommended module choice with their academic advisor or their preferred instructor or even their

classmates, as this might help them stimulate their thinking regarding their module choice. For these students, predictions could form the basis of conversations, and may help to stimulate consideration of alternatives in a process of discovery, as much about themselves and their own preferences as about the modules they are choosing. The university staff members tended to believe that implementing this kind of decision support system would organise the data warehouse's available data in a way that would better support students in achieving their aims. However, they also showed some reservations regarding providing students with such knowledge.

In terms of the data, we noted that the staff and students agree that there are currently data quality issues since it appears there are additional data, such as employability and student satisfaction data, to which students wish to have access in order to make a well-informed choice. For example, Student 6 stated,

> *Student satisfaction would be a massive one for me, because it gives you a good idea of whether that module is something that people have enjoyed. Yes, sometimes there'll be a bit of an issue. Sometimes, some years, it goes a little bit wonky, but it'll give you an idea of what's going on.*

However, according to the staff interviews, these data are of poor quality in the data warehouse of the UEA. Academic JN and Academic N mentioned that the university collects course-level satisfaction data each year. One such data collection comes from the NSS (National Study Survey), which is a 22-question survey on different aspects such as academic support, quality of teaching, planning and organisation, assessment and feedback. This survey should be completed by all the final year undergraduate students and it is published nationally. The second data collection is a similar internal survey that the university developed for its first and second year undergraduate students, which has recently been replaced by the nationally comparable UK Engagement Survey (UKES) for non-finalists. These latter surveys were seen to act as an early warning for university management members to tackle any issues or difficulties arising. These surveys do not, however, ask questions at the module level. Individual-level responses are not available for the NSS, and despite the efforts of the university members to encourage students to participate in these optional surveys, there are no guarantees of achieving an acceptable response rate each year. We should note, for example, that the university does not feed back to the students the results of the internal surveys.

Some universities, or more often the Student Union or representation bodies, create an 'alternative prospectus' that describes students' experiences of courses, modules and lecturers, from a student perspective[1]. The official university source of data that could be used is the

[1] `https://www.applytocambridge.com/`

Student Evaluation of Teaching (SET) data from module evaluations. Practice varies across institutions, and within institutions, in how this data is captured and treated. Response rates vary widely and are often low. In general, the data is not released to students and there are a range of significant issues related to the use of this data [289].

The Destinations of Leavers of Higher Education (DLHE) survey [2] asks students about their employment and other activities approximately six months after graduation. There is also the much less widely used Longitudinal Destinations of Leavers of Higher Education (LDLHE) survey, which asks similar questions but roughly 3 and half years after the end of studies and which is undertaken every two years. Most attention to these surveys has been again at the level of the institutions or courses, in particular focusing on the choice of STEM (Science, Technology, Engineering and Maths) subjects [290]. Unlike the module marks, the DHLE survey is not completed by every student so the data comprises a sample and the LDHLE survey is an even smaller sample drawn from the DHLE. The DHLE that is available to an institution could therefore theoretically be linked to modules to create employability scores for modules, but again the data quality issues are significant.

In contrast, the quality of marks data in the university is high. According to Academic JA, this is because this data is completely controlled by the university. The university controls the assessment process through the exam boards and so on. Students are concerned about their marks and they are likely to identify any errors. Also, the learning and teaching service spends a great amount of time on creating and managing student marks data. We should note that we focused on performance data in our study due to its high quality. This does not mean marks data is, or should be, a more important issue for the students or the university than data on satisfaction or employability. Academic JA argued that that there was a risk of a street-light effect in which excessive attention is given to marks data because of its high data quality, and thus the quality of the data comes to drive the decision-making process.

Alternative argument is that utilising the available data to build such a system will be beneficial for the university in terms of data quality. Demand from students for a wider range of data, the immediate examples being again satisfaction and employability, would drive investment in improving these sources to tackle the challenges of data governance and data quality. Historically, changes to the coding of data were undertaken in an ad hoc manner, reducing comparability and compromising data quality. Academic G mentioned that having such an enrolment system along with the other proposed systems might increase the number of people depending on these datasets, which will lead to stronger data management and quality, as the university will be required to curate the data more formally.

[2] `http://www.hefce.ac.uk/lt/dlhe/`

We perceived from our interviews that the university does not wish to rush resolving the missing data issue; instead they prefer to develop a data structure that will be fit for the next fifteen or twenty years.

In terms of the riskiness of providing students with this partial knowledge, we realise that using a model to directly inform student choice in this way may also have other unforeseen effects. For example, it may raise several ethical issues that we may need to address in more detail in future studies [291, 292]. For example, Pariser [293] discussed how features such as personalisation allow prompt access to more related information, but they cause complex ethical questions and fragment the public in concerning ways. Legislation has already been proposed to restrict the collection and retention of data, mostly over concerns about privacy [294]. Interpretation is the focus of data analysis: data, in spite of its size, is subject to biases and limitations. Therefore, there is a necessity to understand and outline the biases and limitations of the available data to avoid misinterpretation [292, 283]. Mittelstadt et al. [291] have discussed a range of ethical concerns caused by the use of algorithms. They label one of these concerns as 'transformative effects', which occur when algorithms influence how we understand the world, and transform its social and political organisation (see also MacKenzie [282]).

In terms of investigating how the university saw the potential for utilising the results derived from Chapter 5, the interviewees provided a different range of perspectives. Some staff believe that since the information is available, then morally the university should act upon on it. University staff also tended to see the information as of particular benefit to certain types of students, to help students get the best value out of their tuition fees, and lastly, it would benefit the league tables and the general level of academic attainment. Other staff were more cautious, believing that is important that the university first investigates the causes or drivers of poor performance, so that appropriate support can be put in place. Finally, a few staff noted that there were other opportunities that students might want to pursue, other than higher marks, especially if they were not failing. The staff members agreed that the assistance should be an offer, not compulsory, and that the university should put in place a consistent and systematic approach to respond to predicted 'underperformance' and provide interventions to avert poor outcomes. The interviewees provided a range of suggestions about the form that an offer to help students at risk might take. They all agreed that extra assistance could legitimately be provided to students at risk only, as there were other resources available to all students such as personal advisors and Student Support Services.

We believe that this part of the findings is essential, since it shows that having a model with a high level of predictive accuracy is not, in itself, enough; it can raise a number of quite

complex questions about what could or should be done in response to this information.

7.4 Summary

The primary goal of this chapter has been to examine the value of utilising the available student data, particularly performance data, and to investigate whether feeding back data analysis to students, in the form of individualised predictions of future grades, is acceptable to the main stakeholders, students and staff, and to explore the potential for unforeseen consequences. For this purpose, we carried out a questionnaire survey that collected data from 59 students and we conducted interviews with 28 students and seven university staff with key roles in the module choice process, in order to understand attitudes and perspectives. We believe that this is important since, to our best of our knowledge, none of the published relevant studies (reviewed in Chapter 2) that focus on predicting module performance have addressed the management and student aspects of the given predictions; instead they have focused solely on the technical aspect of building accurate predictions.

Overall, staff and students show an indication of acceptance towards having the provided partial knowledge as an addition to, rather than a replacement for, existing sources of support. However, they showed some concern that it could be a dangerous thing too. Some were concerned that in the absence of a wider range of knowledge, people might overemphasise what knowledge they do have. In particular, the worry was that a marks-based recommender could lead students to overemphasise marks at the expense of other criteria such as intellectual curiosity, employability or satisfaction. Others were concerned that there is not enough knowledge about the individual student going into the prediction in the first place. However, adding more personal information may raise ethical issues such as privacy. It also raises the possibility of misguiding the students and makes explaining the output of the recommender in a transparent manner more challenging.

We also found that individuals vary in how they make their module choices. Some of the students will look at all the given information, then make their decision in a logical and a reasoning way. Other students might prefer to discuss the given information with an experienced academic supervisor or instructor before selecting their module choices. Lastly, there are some individuals who will trust information that is seen as having the authority of statistics or technology. This latter group of students are perhaps the individuals of most concern. Therefore, the university has to be cautious about selecting the knowledge that they can feed to those students and the way of presenting that additional knowledge. In addition, the university should alert the students that the proposed system is just a tool to

aid their decision-making process and will not make the module choices for them.

In this chapter, we also extended Chapter 5's work by studying the perceptions of some of the university staff members on how to utilise the knowledge derived from Chapter 5 to improve students' outcomes.

Our understanding of 'the long term societal effects of datafication' [295] remains poor. The stability of complex systems, many with endogenous feedback loops, cannot be taken for granted and there is always a possibility of unanticipated and negative outcomes.

Through our study we have uncovered the complexity of the educational system in which educational data mining operates. We have uncovered in particular that, while in other areas such as the commercial arena, a recommender system may be easy to implement, in the educational context, there are many areas of concern that need to be explored, not least because choices made through education have very long-lasting effects for individuals, and education is now a very highly priced commodity. We have uncovered that the data that institutions currently collect is probably not deemed sufficient to support trustworthy recommender systems. Furthermore, we have also discovered that there should be an emphasis on transparent models and that the presentation of results is not straightforward. It probably needs to be accompanied by a more established means of providing academic advice, where any recommendations can be discussed with students and the appropriate emphasis put on them. We have therefore uncovered that educational data mining requires very careful assessment of data quality, environment, stakeholders, and ethical considerations before any implementation is put into place. We have also determined that simple models are probably not adequate since decisions are very complex and rely on many different factors, some of which we may not be able to model adequately with the available information.

Chapter 8

Conclusions and Further Research

In this chapter, we look back at the aims and objectives of the research formulated in chapter 1, and we review how the research undertaken has achieved those, and the contributions that they represent. We also outline the recommendations that result from this study. Section 8.1 summarises the study undertaken and the findings, in addition to producing some recommendations. Section 8.2 discusses some of the limitations of our study. It also provides pointers to additional related research that should be conducted in the EDM field.

8.1 Conclusions

The key goal of this study is to explore how data collected routinely by universities can be used in the context of educational data mining to enhance student experiences and outcomes. Overall, we explored and experimented with several analytical techniques that can be employed to predict and improve student outcomes in the context of Higher Education, both at the programme level and at the module level. We focused on DM techniques for prediction, and a qualitative approach to gain an understanding of how the information derived from the predictions could be put to some use. The key reason for moving from data prediction or "technical aspects" to decisions or "management aspects", as was stated earlier, is that having a highly accurate prediction system in itself is not enough for the institutions and students. The derived knowledge should be translated and used to improve decision making and performance. This essential element equates to model deployment and although because of the time frame of the thesis we could not become involved with the

actual deployment, nor did we have adequate data access, we could at least investigate some of the possible consequences of model deployment via our management study. This also enabled us to understand the complexities of the educational setting and the additional requirements that may make some of the traditional approaches to providing recommendations based on information from other individuals who interact with a system inapplicable to this context

The thesis can be summarised as follows. In Chapter 2, we surveyed the literature to obtain useful insights about EDM, decision making, choice, data and related studies. This gave us a number of ideas for our own experimental work, including addressing prediction as a classification and regression problem, using many state-of-the-art algorithms and finding ways to successfully deal with missing data.

In Chapter 3, we described the process that we followed to acquire the data for the predictive models. This included the data extracted from the university and from the machine learning repository. We also described the data we acquired through interviews and questionnaires for our management study.

In Chapter 4, we attempted to explain how our results chapters are connected. We also introduced the algorithms and methods used at each stage and our own approaches, including a clustering approach used to create a prediction for module performance, a multiple imputation approach to deal with missing data, and our approach to complete the management aspect of our research. Furthermore, we described the evaluation criteria used and the statistical testing methods applied. We also the included ethical considerations of the thesis as these are important when analysing student data.

In Chapters 5 and 6 we presented the results of our prediction modelling at the degree outcome and module outcome level respectively. Finally, in Chapter 7 we presented the management study to consider how the knowledge extracted from the predictive models could be embedded to the satisfaction of staff and students.

Next, we explain in more detail how the work in the thesis has met the initial objectives of our research, and eventually attained the contributions of the study.

Objective 1. We will extract and prepare student data for analysis and report on its quality. Some of the effort required for this objective is presented through Chapter 3, where we highlight the operations necessary to prepare the data for an analysis. Much effort, which we do not elaborate on, was first required to extract the relevant student data from the data warehouse using querying tools. Initial data exploration led to data filtering, including both filtering student records which did not meet the entry requirements (because of missing data or anomalies) or filtering attributes (e.g. those on attendance and

engagement) because of quality problems. A reasonably clean and consistent version of the data was achieved, which was ready for analysis and contained a number of attributes including some engineered attributes (e.g. on comparative performance). We were also able to advise the University of the need to improve data collection in relation to engagement, employment and other aspects that could be important in the future if EDM is to be applied in earnest.

Objective 2. We will perform initial analysis to identify students who are at risk of obtaining poor outcomes using data mining methods.

We addressed this in Chapter 5, where we tackled the more general problem of performance prediction and highlighted year 1 students associated with poor performance, so that suitable interventions could be suggested to improve their outcomes. We defined good performance as a binary approach, "Good Honours" versus "Not Good Honours", in accordance with current UK university aspirations. We explored the available performance and demographic variables in relation to the overall Good Honours rate. For this we used a feature selection ranking algorithm based on the Pearson chi-square test. Then, we used nine classifiers and combined them using an ensemble to develop the best possible model. We attempted to build the model, using three feature sets: first, the generic characteristics of students, which are available at registration only, and then we added the performance of year 1. Lastly, we investigated optional module choices in year 2 and 3 in relation to the overall outcomes. We took into account the difficulty of each module by comparing the performance of each student with his/her peer group. We did not employ the available engagement data such as library services and attendance monitoring information, due to quality issues. In this chapter we compared the results across two schools of study associated with the UEA.

The main findings of the initial analysis show that we were able to uncover groups of students that corresponded with poor performance in terms of Good Honours degrees, identifying 57% of the low attainers with GH rates as low as 32.6%. We were also able to identify specific characteristics known at registration associated with poor performance. We found that the marks for year 1 modules between GH and NGH students are statistically significant different, although year 1 modules do not count towards the overall outcome. Therefore, adding year 1 performance improved the models: 89% of the low attainers were identified with 21.97% GH rate. The accuracy of the built model was improved by adding the third feature set, as expected. Also, by adding this feature set, we discovered some problematic optional modules that are related to the overall classification. We did not uncover any year 1 modules that were specifically problematic; instead, low attainers performed badly across all year 1 modules. We found that poor performers showed some marginal progress according to their year 2 and 3 average. Thus, we conclude that an intervention targeted at this

group could provide some lift to the GH rate and affect those students positively. Such an intervention may also have a positive effect on the university's league table position as that takes into account the GH rate. Our discovered patterns were similar for the two datasets, although each dataset belonged to a different school of study and therefore was associated with teaching a different discipline with a different admission strategy. We conclude therefore that it is possible to use predictive modelling from routinely collected data to highlight performance issues early on.

Objective 3. We will then investigate how to construct a robust predictive model for module performance, which could be deployed as part of a future enrolment support system. Our initial recommendations will be based on potential student performance on a module so for this we will present a comparison of different predictive models that could be used in the context of module outcome prediction. As part of this effort, we will investigate methodological issues that arise in educational data mining models.

In Chapter 6, we tackled the more granular problem of module performance prediction using the historical academic performance and characteristics of students. We believe this is important, as we could provide students with additional knowledge that might help them with their module enrolment, since module choice could affect their overall degree by presenting different opportunities and challenges. Particularly, we addressed module performance prediction as both a regression and classification problem. We compared multiple algorithms (such as Simple Average prediction, SVM, Random Forest, Rpart, C5 and clustering) over multiple datasets. We used the Friedman rank statistical test to explore statistically significant different performances. We made this study more robust by using different datasets associated with two schools of study within one university and two public datasets that are related to a different educational institution.

In approaching module performance prediction, we encountered missing data problems. We studied the effect of missing data and proposed a novel approach for this as this subject has not been well explored in the EDM literature, according to our review. We applied multiple imputation by chained equation and expectation maximisation, and Random Forest imputation in an ensemble data mining context. We also experimented with increasing amounts of missing data in the public complete datasets by removing 25% and 45% of their values.

The key findings were that the ensemble approach combined with multiple imputation and SVM or RF produces consistently good results in all cases and for both classification and regression — it is therefore recommended. The modelling performance by either classification or regression with any of the leading DM algorithms (RF, SVM, Rpart) can also provide

good results, since overall no statistically significant differences were shown between a number of algorithms. The baseline prediction (Simple Average model) produced the worst performance compared to the modelling techniques, so employing a model is recommended. The performance of the regression approach was more differentiated, as it produces finer grain results, compared to the binary classification. Lastly, by comparing the performance of the complete public datasets with those obtained by removing 25% and 45% of their values, we found that missing data obtained with an MCAR mechanism has no noticeable effect on the prediction accuracy.

Objective 4. We will perform robust experiments by using a number of datasets with different characteristics.

Our experimentation data came from two different schools in the same university. This was to contrast the results in two slightly different environments. It was not possible to access other educational datasets for the first set of experiments as universities work within ethical and privacy constraints that prevent them from making data on their students publicly available. However, for the problem of module performance prediction we did find some publicly available datasets, which we used to enhance the robustness of the methodological experiments for how to handle missing data. When considering each dataset (each module in each school becomes a separate dataset) from the two schools plus the publicly available data, it becomes possible to apply statistical tests that take into account multiple algorithms over multiple datasets.

Objective 5. We will investigate qualitatively, by means of interviews, the views of students and staff on deploying the knowledge found, for example as part of an enrolment recommender system or a programme of remedial action for students at risk of poor outcomes

In Chapter 7, we focused on investigating the step after data modelling, which in our context is how to utilise the knowledge derived from the previous exercises to inform students and staff. So we investigated whether the university should act on the findings, and what measures they could take. Such management aspects have often been neglected in other technical studies on performance prediction. However, the investigation is very important because it helps us to uncover aspects of the educational environment that make it very unique and requiring special consideration. The complexity of decisions, the long-term impacts of any decisions and the high capital costs of education, as well as the ethical considerations, mean that systems that may be applied in other settings (e.g. a commercial setting) may not be acceptable in this setting. Much of this complexity was uncovered through our management study.

We conducted mainly a qualitative approach, in particular by using semi-structured interviews, combined with a questionnaire survey. The purpose of the questionnaire survey was to grasp an initial understanding of undergraduate student views on module choice and to recruit students for the interviews. The survey participants were 59 students. The interview participants were 28 students and 7 staff members from different roles in the university. The student participants in this study were associated with the same schools of study as those in the previous chapters, for consistency purposes.

The key findings of this chapter showed a general positive attitude towards having additional knowledge – as an addition to, rather than a substitute for, existing resources. There was also a positive attitude towards having a future enrolment recommender system tool partially based on academic performance. Nevertheless, there were some reservations that it could have potentially dangerous unintended consequences. Reservations included the concern that students might focus overly on the the provided knowledge, such as marks, at the expense of other essential criteria such as employability or intellectual curiosity. Another reservation is that there is insufficient information about individual students to produce good predictions in the first place.

The study suggested that adequately employing the available information in such a system might result in nudging students to select modules that would help them achieve better academic outcomes. However, students were varied in how they selected their modules:

- Some of the students made their decisions in a logical and a reasoning way.
- Some students might prefer to discuss the given information with an experienced academic supervisor or instructor before making their choice.
- Other individuals favour information that is seen as bearing the authority of statistics or technology. This group of students are possibly the individuals of most concern. Therefore, the university has to be careful about choosing the knowledge that can be fed to those individuals and the way it is displayed since it may have a disproportional weight in their decisions.

Both staff and students showed concerns regarding the lack of transparency about the process by which the given predictions and recommendations were produced. Most interviewees stated that a human advisor interaction would still be required, despite adding personalisation to the given predictions.

Based on the study of this chapter, we recommend that if the university were to implement such an enrolment tool, they should take special measures. In particular, recommendations may have to be presented in a very cautious way, perhaps in the context of the advising system and using well-designed language and communications methods. We therefore

recommend employing a discursive approach by including human interaction during the enrolment process by thoroughly discussing and interpreting the given results of the recommender system tool with the students, and also discussing the methods by which the predictions are arrived at, as this could enhance the transparency of the enrolment system. Also, we suggest that the university should ensure that the students understand that the proposed system is just a tool to assist their decision-making process and provide them with further information, but that it should not make the module choice for them.

It may also be necessary to enhance the data collection with good-quality data on employability, engagement and student satisfaction. The proposed system would then need to combine all these factors to produce information at many levels. Furthermore, the proposed system should act as a portal to present all the necessary information, to save the students time, searching for relevant information in different places. For example, the proposed system could present for a given module choice: description of module, number of students enrolled previously, average grades in previous years, student satisfaction measures, student engagement measures, employability scores, related modules (modules that may be considered in combination with the module being considered or that follow from it) and then a predictive score. Students would then make choices based on complex criteria with all the relevant information. Given the complexity of the information presented, there would definitely be a need for interaction with experts (i.e. advisers) to interpret the given information.

The main findings in terms of how the university staff viewed the potential for interventions, based on predictions of poor performance presented in Chapter 5, included:

- Some staff think that the university has a moral obligation to act upon predictions of poor performance.
- Some believe that the information will be beneficial to certain types of students, and support them in getting best value out of their annual tuitions fees.
- Other staff were more cautious, by pointing out the university should first examine the drivers or the causes of poor performance, so that suitable assistance can be put in place.
- A few staff pointed out that there were other outcomes that students might want to achieve, not related to high marks, especially if they are not at risk of failing.

In addition, our staff interviewees provided a range of ideas about the form that an offer to assist students at risk might take. Staff members agreed that the support should be optional, not mandatory. The university should have a consistent and systematic approach to handle predicted 'under-performance' and provide interventions to prevent poor outcomes. The

additional support could be provided to students at risk only, since there are other resources obtainable by all students.

This part of the study showed us that for less complex scenarios, it may be easier to implement coherent approaches to deploy the knowledge found by modelling. However, having a model with a high level of predictive accuracy is not, in itself, sufficient because it can cause a number of quite complex questions to arise, such as what could or should be done in response to this information, who should benefit from any actions and where are the resources to implement any required strategies. Once more, analysing the management aspects through the stakeholders' attitudes has unveiled the complexity of the educational setting.

By successfully addressing all the objectives above, we have achieved the contributions of our research, stated earlier in Chapter 1.

8.2 Limitations and Future work

This research has highlighted a number of areas that could be explored in the future; these are:

- Although this thesis showed that the quality of performance data is high, other types of data are not collected adequately. In the future it will be necessary to improve data collection on these aspects, e.g. employability, engagement, etc., and use them in the modelling to understand how other factors may affect student choice and outcomes. For example, engagement could be measured by interaction with the Blackboard, a system used by the university to present information to students. However, interviewed staff argued that the university is not currently putting a great deal of effort into this area, so instead they want to slowly build a data warehouse that can be fit for the next fifteen or twenty years. In practice, rich information may be necessary to create acceptable recommender systems, so future research should focus on how to efficiently and effectively collect rich information of high quality for all aspects of student choice and attainment.
- In Chapter 5, we found some optional modules that had a large impact on the overall classification. One strategy could have been to survey the faculty members and students associated with those in order to explore their perceptions, as this might present some validation of the results of the predictive models. It may also have been useful to implement some remedial approaches for students at risk of poor outcomes and observe their effect and the views of those involved. However, this was considered to

be outside the scope of the project.

- We found from our study, particularly in Chapter 6, that data MCAR seems to have no noticeable effect on the accuracy of the predictions. Therefore, we could investigate in future work whether the MAR and MNAR mechanisms produce a similar effect. Our own results on MAR data seem to also show good performance in the context of large amounts of missing data. This could be done using benchmark complete datasets; however this is also considered to be outside the scope of our thesis as we are focused on the analysis of educational data.
- We did not apply a number of approaches that are often used in recommender systems to produce predictions. Those are based on matrix factorisation techniques, low rank approximation techniques, and the traditional recommender methods, which include collaborative filtering, content-based and hybrid filtering techniques. Ideally we would have implemented such approaches to investigate their worth against predictive models. However, we attempted to focus on predictive models, which required many experiments to be run, especially in the context of multiple imputation, so we were limited in our scope to implement other approaches. We believe this would be a good area of future work.
- Student interviewees exhibited complex and varied criteria for, and approaches to, module choice. The provision of predicted grades could affect their module choice in a range of ways. Therefore, we are aware that we need further experimentation to test their responses. Particularly, we could implement an actual prototype of the enrolment system and test the responses to it. This may be best performed as a long-term project in which a form of randomised control study is performed, with some students receiving assistance with module choices and other students receiving no assistance. Overall outcomes could then be compared. However, this would require complex ethical approval and a long follow up, which was outside the scope of our project.
- In terms of the ethical considerations, we were limited to using mostly data associated with one UK university, which may prevent us from generalising our conclusions as we would wish. The application of similar models to other educational settings and datasets would be advantageous. Ethical considerations for projects such as this are complex and they become even more complex when considering implementation issues, hence a study of ethical implications of data analytics in an educational setting in itself may also be appropriate.
- It may also be valuable to extend interviews and modelling to other schools of study,

to fully understand how different disciplines affect students' views and attitudes.

Lastly, we believe our research has added value to the EDM field as we have gone beyond the analytic aspects of creating models of student performance to the management aspects of how such models may be acceptable to those concerned. This has showed the complexity of implementing analytical approaches, and the necessary aspects that should be investigated, such as getting the view of stakeholders, looking at the data quality issues and considering all ethical questions. From the technical aspect, our research has addressed the missing data problem, which is a neglected area in the EDM field.

Appendices

Appendix A

Public Dataset Attributes

Table 1: The preprocessed student related variables

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
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).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Mean Squared (RMSE) is a popular metric (Witten and Frank 2005). A high PCC (i.e. near 100%) suggests a good classifier, while a regressor should present a low global error (i.e. RMSE close to zero). These metrics can be computed using the equations:

$$\begin{aligned} \Phi(i) &= \begin{cases} 1 & , \quad \text{if } y_i = \widehat{y}_i \\ 0 & , \quad \text{else} \end{cases} \\ PCC &= \textstyle\sum_{i=1}^{N} \Phi(i)/N \times 100\,(\%) \\ RMSE &= \sqrt{\textstyle\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2/N} \end{aligned} \tag{1}$$

where $\widehat{y}_i$ denotes the predicted value for the i-th example.
In this work, the Mathematics and Portuguese grades (i.e. G3 of Table 1) will be modeled using three supervised approaches:

1. **Bin**ary classification – *pass* if G3$\geq$10, else *fail*;
2. **5-L**evel classification – based on the Erasmus[1] grade conversion system (Table 2);
3. **Reg**ression – the G3 value (numeric output between 0 and 20).

Figure 1 plots the respective histograms.
Several DM algorithms, each one with its own purposes and capabilities, have been proposed for classification and regression tasks. The Decision Tree (DT) is a

[1] European exchange programme that enables student exchange in 31 countries.

Appendix B

Interviews Coding Nodes

Name	Sources	References	Created On	Created By
Recommender	9	28	4/6/2017 11:48 AM	ZA
Acceptance Reccomender	35	97	4/10/2017 4:02 PM	ZA
Responsibility for choice	31	84	4/6/2017 12:08 PM	ZA
Over relay on technology	16	24	4/6/2017 12:08 PM	ZA
Over rely on numbers	5	6	4/6/2017 12:08 PM	ZA
Assurance	25	52	4/11/2017 2:49 AM	ZA
Substitution or supplementation	33	48	4/6/2017 12:10 PM	ZA
Personlisation	24	43	4/11/2017 2:41 AM	ZA
Capacity issue	2	2	4/30/2017 2:09 PM	ZA
Disproving the prediction -counter performa	21	36	4/6/2017 12:09 PM	ZA
Knowing more about students (interest , care	23	31	4/6/2017 12:06 PM	ZA
Trust	17	29	4/6/2017 12:06 PM	ZA
Reliability (version 1.0)	19	31	4/6/2017 12:07 PM	ZA
Authorial voice (uni ,staff)	3	3	4/6/2017 12:07 PM	ZA
Reduce info search	21	27	4/6/2017 12:06 PM	ZA
Become aware of other choices	15	21	4/11/2017 3:26 AM	ZA
personal info in context	17	20	4/6/2017 12:05 PM	ZA
Curiosity- what does it say	10	16	4/6/2017 12:09 PM	ZA
Affect of Predicted mark	8	9	4/11/2017 3:09 AM	ZA
Low prediction	21	25	4/11/2017 3:10 AM	ZA
Poistive	14	15	4/11/2017 3:12 AM	ZA
Negative	10	13	4/11/2017 3:12 AM	ZA
High prediction	20	21	4/11/2017 3:11 AM	ZA
Positive	15	16	4/11/2017 3:16 AM	ZA
Negative	10	10	4/11/2017 3:17 AM	ZA
unique system (stand-out)	6	6	4/23/2017 2:40 PM	ZA
helplecturerPromotingThierModules	3	3	4/28/2017 1:47 PM	ZA
Preparation	5	6	4/27/2017 1:58 PM	ZA
Choice	11	20	4/6/2017 11:48 AM	ZA
Ethics	20	32	4/6/2017 11:52 AM	ZA
Influence Criteria	12	24	4/6/2017 11:52 AM	ZA
Interest	27	62	4/6/2017 12:02 PM	ZA
Career direction	30	54	4/6/2017 12:01 PM	ZA

Marks		25	52	4/6/2017 12:01 PM	ZA
Challenge		22	32	4/6/2017 12:02 PM	ZA
Lecturer or staff		17	29	4/6/2017 12:03 PM	ZA
students satisfaction		19	23	4/6/2017 12:00 PM	ZA
Assessment		8	11	4/6/2017 12:04 PM	ZA
CW or exams		11	15	4/6/2017 12:04 PM	ZA
team or group work		1	1	4/6/2017 12:05 PM	ZA
ModulesConnection		7	7	4/23/2017 1:10 PM	ZA
ModuleDescription		6	7	4/30/2017 12:41 PM	ZA
Time table		2	2	4/6/2017 12:03 PM	ZA
Exemption		1	1	4/6/2017 12:00 PM	ZA
Constraints		13	15	4/6/2017 11:50 AM	ZA
lack of information		28	51	4/6/2017 11:57 AM	ZA
time table		7	10	4/6/2017 11:56 AM	ZA
pre-requisite		5	7	4/6/2017 11:57 AM	ZA
Information Sources		6	11	4/6/2017 11:52 AM	ZA
Peers		20	32	4/6/2017 11:58 AM	ZA
Advisors		17	20	4/6/2017 11:58 AM	ZA
Parent-Family		2	2	4/6/2017 11:59 AM	ZA
Risk		8	10	4/6/2017 11:51 AM	ZA
Variety		7	9	4/6/2017 11:51 AM	ZA
DegreeShaping		6	9	4/27/2017 2:36 PM	ZA
OverlookModulefairEmail		6	8	4/23/2017 11:39 AM	ZA
ImproveStudentOutcome		7	19	5/6/2017 11:16 AM	ZA
supportRiskStudent		6	10	5/6/2017 11:24 AM	ZA
EqualitySupport		6	8	5/6/2017 11:27 AM	ZA
admissionPolicy		4	6	5/6/2017 11:31 AM	ZA
Data		5	9	4/6/2017 11:48 AM	ZA
Feedback		23	37	4/6/2017 11:53 AM	ZA
Data Quality		13	18	4/6/2017 11:54 AM	ZA
DataUse		5	12	5/7/2017 2:44 PM	ZA
missing data		3	4	4/6/2017 11:55 AM	ZA

Appendix C

Interviews Duration

Interviewee Alias Name	Date	Duration
Student 1	15/03/2017	0:07:49
Student 2	13/03/2017	0:15:56
Student 3	13/03/2017	0:18:04
Student 4	16/03/2017	0:10:52
Student 5	14/03/2016	0:20:25
Student 6	16/03/2017	0:16:43
Student 7	11/03/2017	0:13:29
Student 8	21/03/2017	0:19:49
Student 9	01/01/2012	0:16:30
Student 10	13/03/2017	0:13:57
Student 11	13/03/2017	0:14:51
Student 12	19/03/2017	0:13:27
Student 13	22/03/2017	0:13:06
Student 14	23/03/2017	0:15:05
Student 15	29/03/2017	0:17:43
Student 16	13/03/2017	0:12:02
Student 17	16/03/2017	0:16:51
Student 18	20/03/2017	0:08:00
Student 19	22/03/2017	0:10:53
Student 20	20/03/2017	0:07:00
Student 21	22/03/2017	0:09:29
Student 22	16/03/2017	0:13:06
Student 23	24/03/2017	0:08:19
Student 24	24/03/2017	0:17:26
Student 25	23/03/2017	0:19:50
Student 26	22/03/2017	0:14:10
Student 27	23/03/2017	0:14:36
Student 28	29/03/2017	0:15:26
Academic H	31/03/2017	0:37:32
Academic D	04/01/2017	0:35:19
Academic N	29/03/2017	0:41:18
Academic G	24/03/2017	0:35:05
Academic JA	12/03/2017	1:04:06
Academic C	20/03/2017	1:08:24
Academic JE	22/03/2017	0:37:45

Appendix D

Ethic Checklist form

School of Computing Sciences

Research Ethics Check

Name …….. UG / PGT / PGR / (S)RA / Faculty / Other

Title of project (80 chars. max.) ……...

Name of Supervisor / PI / Lab leader:……

A. Does the research use an interview or questionnaire survey? **Yes No**

If so, does it:

- Ask for any personal information? **Yes No**
- Ask personal questions other than those from published surveys/questionnaires? **Yes No**
- Use questions on age, gender or ethnicity other than those in widespread use? **Yes No**
- Ask other personal or sensitive questions? **Yes No**

B. Does the research offer advice or guidance to people? **Yes No**

- Are you using a validated knowledge base? **Yes No**
- Are you (or your collaborators) formally qualified to give the advice or guidance? **Yes No**

C. Does the research involve children, vulnerable adults or their carers? **Yes No**

- If so, have you obtained the relevant VBA checks? **Yes No**

D. Does the research record or observe people's behaviour? **Yes No**

- Does it replicate other published studies? **Yes No**
- Are these recent and culturally compatible? **Yes No**

E. Has this research been previously considered by another REC? **Yes No**

If so, please provide full details in the research protocol.

F. Does the research involve the analysis of personal data collected by others? **Yes No**

If so, please describe the arrangements made to ensure confidentiality, security, … in the research protocol.

G. Will the researcher carry out fieldwork alone while away from UEA? **Yes No**

If so, please describe the arrangements made to ensure the researcher's safety in the research protocol.

H. Will participants be paid or offered a reward for participating? **Yes No**

If so, please describe, in the research protocol, the arrangements made to record the names and addresses of everybody receiving a payment.

I. Data management

- Does the research collect or use sensitive data? (e.g. commercially confidential, military, …) **Yes No**
- Does the research use existing confidential data? (e.g. medical records) **Yes No**
- Is the research covered by the consent given when the data were collected? **Yes No**
- Are special arrangements needed for the storage (10 years) of the data? **Yes No**

J. Attachments

☐ Project synopsis ☐ Research protocol ☐ Questionnaire ☐ Other forms

Approval (Chair of CMP-REC)

Approved **Yes No** Signature **……..** Date……..……

Please return completed form to CMP Office, S2.45

Notes for guidance

Any research, dissertation or project carried out at UEA that involves working with people or animals - either directly or indirectly - must obtain ethics approval before work starts. Failure to do so is a Research Misconduct matter.

Many applications can be processed quickly, but work that falls outside the scope of CMP-REC (a sub-committee of the UEA REC) will be referred elsewhere. Work that involves medical patients, or NHS staff issues that may affect health and well-being, must be approved by a NHS REC and Research Governance Committee. Work with NHS staff on non-sensitive matters (e.g. use of IT) needs CMP-REC ethics approval and NHS Research Governance approval. Plenty of time must be allowed for these processes.

The most important issues in considering the ethical dimensions of a project are:

- **Appropriateness of methods**. Are the methods proposed appropriate (e.g. not unduly intrusive, or time-consuming) for the gains in knowledge and understanding expected,
- **Experimental subjects and consent**. These are *indicative* topics to be addressed in the research protocol:
  How will you recruit subjects?
  How many will be recruited? (justified in relation to the aims of the survey and the analysis methods)
  How will you obtain the informed consent of your subjects?
  How will they be informed of their options to withdraw and of any risks or benefits from participating?

Attachments

Project synopsis. The committee needs to have an understanding of the scope and aims of the project; these should be provided in the project synopsis. The project synopsis is usually no more than two paragraphs long.

Research protocol. This describes the experimental or survey methods and procedures to be used; it should be written in sufficient detail to (in principle) allow a reasonably competent researcher to complete the experimental or survey work with no additional information or guidance.

Questionnaire. Copies of all questionnaires, interview forms etc. must be attached. The questionnaire should provide participants with sufficient information about the project and questionnaire to allow them to decide whether or not to participate, what will happen to the information they provide, what will happen if they withdraw part way through, contact details of the investigator and supervisor (or Head of School)

Other documents. Any other participant information sheets, consent forms, etc. that will be used in the research

Sections

A. Interview or questionnaire survey. This covers all face-to-face or web-based surveys, systematic programmes of interviews, comparison tasks, etc. You do not need to complete this form if you are **only** carrying out a requirements gathering interview with a single stakeholder for whom you are designing a system.

B. Advice and guidance. Answer Yes to this if your work will produce advice for people on matters that may directly affect their health or well-being, e.g. exercise or diet. Answer No to this question if one of the outcomes of your work will be some suggestions about how a website or business process might be improved, etc.

C. Work with children or vulnerable adults or their carers. If you answer Yes to this question (see https://www.gov.uk/disclosure-barring-service-check/overview), you must explain fully in the research protocol how this work will be carried out. You will also need to be aware of the University's policies on research with children and with people who may fall within the scope of the Mental Capacity Act 2005.

D. Recording or observing behaviour. This covers thinking aloud, speech, lip-reading experiments, etc.

E. Previous applications. A copy of the submission, the REC applied to, the date and outcome.

F. Analysis of personal data. The research protocol should explain the nature of the data, how anonymity will be ensured (if appropriate), any contracts, non-disclosure agreements or limitations on the use of the data, …

G. Safety of researcher(s). Does the work involve exposure to risks beyond those involved in everyday life in the UK? (e.g. unwanted attention from overseas police authorities for work which would be unremarkable in the UK) If so, appropriate arrangements must be made to reduce the risks where this is practicable and to ensure that there is a system for positively reporting the safe completion of each research session or activity.

H. Payment. UK tax regulations require that the University keeps details of all payments made. The list of payees' details should be kept securely, and it should be designed so that research subjects' confidentiality is preserved.

I. Data. These are indicative questions, covering topics that need to be addressed in the research protocol.
(See also https://intranet.uea.ac.uk/ren/Research+Data+Management)

What observational or behavioural data will be collected? How?
Will the data be made available to other studies? How?
How will experimental subjects be informed of these issues?
For secondary analyses, is the work covered by the consent obtained when the data were collected?
What is the data storage plan?

Appendix E

Using Student Data Ethical Documents

Project Synopsis

This research project will involve data analysis of part of the UEA Students' data that have been collected by the UEA Business Intelligence Unit. The Analysis will be done by applying data mining techniques to build models or extract patterns that can explain the discrepancy on students' outcome. This research will try to produce recommendations that will improve the UEA management process, as well as inspecting the data analysis techniques that may be more useful in that context. The data use will be 'fair', because the outcome will not impact the individuals. The data will not be shared with any other organization. In addition, the research results will be anonymized; thus it will not be possible to identify individual students from any results published.

The UEA Business Intelligence Unit will create specific 'Research Views' in the data warehouse that will give the researcher access to the data she needs, but without students' name. The UEA information compliance manager (David Palmer) will add an alteration request to their list of proposed changes on Student Data Protection Notice for 2014 summer. This alteration means adding a specific line to the Data Protection statement of students on registration, noting that students' data may be used for research purposes that are in line with broad corporate objectives. This clearly will include all future students' data but as the information compliance manager state this alteration will also consider valid to all previous students.

Lastly, the researcher has put a data management plan in place.

Please Contact the following if further confirmation/clarification is required.

- Dr. Garrick Fincham (Business Intelligence Unit) : g.fincham@uea.ac.uk
- Dave Palmer (Information Compliance Manager): david.palmer@uea.ac.uk

Data Management Plan

Project Title

Analytics and Information Management in Higher Education

Research Student

Zahyah Hamed Alharbi (School of Computing Sciences).

Project Supervisory Team

I. Dr. Beatriz de la Iglesia (CMP).
II. Dr. James Cornford (NBS).
III. Dr. Garrick Fincham (UEA BI Unit).

Project Duration

Start Date: 1st April 2014.
End Date: 1st April 2017.

Data Collection

What data will you collect or create?

I will analyse UEA existing students' data that has been collected by the university during the process of admission, registration, UG and PGT study, and also on students' destination after graduation. The data will contain details of performance, employability, and will include attributes such as gender, subject of study, classification, etc. The data has already been collected by UEA Business Intelligence Unit and is managed by the,. The project is concerned with the data mining analysis of such data to extract insightful information from the raw data, but also with the management aspects of utilising the information extracted in the data mining process to improve decision-making within an educational organisation. Educational Data Mining is a growing field of research so we want to investigate how data is being utilised to improve student outcomes and how results of any data mining analysis can be fed back to the institution to create tangible benefits for both students and the institution. When necessary the UEA data will be used as a case study to understand the application of educational data mining to real data.

How will the data be collected or created?

The UEA Business Intelligence (BI) Unit collects and curates data on students' admissions and performance including employability. The BI Unit will create specific 'Research Views' that can provide the researcher with the data she needs while maintaining anonymity by hiding details such as student name and other identifying characteristics. The BI Unit will maintain ownership of the data at all times. Some linking identifiers may be left to permit linking of data but will be protected by the BI Unit to ensure that the data remains anonymous.

Documentation and Metadata

What documentation and metadata will accompany the data?

The data will be derived from the students' information that has been collected by UEA BI Unit and it will be provided as Excel sheets containing the restricted views. Some documentation may be provided as Word documents. For example, a data dictionary will be provided as metadata which will explain the meaning of the different fields collected, with their expected ranges and where necessary coding information.

Ethics and Legal Compliance

How will you manage any ethical issues?

The data will not be shared with any other individual or organization. During the project, the researcher and the primary supervisor team, will be the people with access to this data. Unpublished results may be shared with (1) the supervision team, subject to any confidentiality restrictions and (2) Internal and external assesors and markers, subject to the completion of the appropriate confidentiality agreement

In addition, any research results will be fully anonymized; thus no individual students will be identifiable. Publication of any results will be with consent from the BI Unit. They will oversee any publication and will have a look at any article before submission to approve it.

The researcher has also applied to get an ethical approval for the project. Furthermore, the data will be saved in an encrypted format and only in secure storage as specified below.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The intellectual property for the project is owned by UEA BI Unit.

Storage and Backup

How will the data be stored and backed up during the research?

The researcher will be storing the data in The Central File Store (CFS or U: drive), because as the UEA IT support's staff stated that it is secured and centrally managed storage which is available to users with access to PC/MAC on the UEA network, or externally via VPN; and It is provided by IBM N5600 Nas Gateways located in UEA data centres, with back end storage from IBM DS5000s virtualized using IBM SAN Volume Controllers. Also, postgraduate users get 10GB of storage with an option to increase this by purchasing additional space.

With regards to the data backup, UEA IT support staff stated that will be snapshots taken regularly throughout the day, as the following schedule illustrate.

Schedule	Created	Number retained
Hourly	8am, 12pm and 4pm	4
Nightly	Each night at midnight	7

Tape backups consist of nightly differentials and weekend fulls which are retained for 28 days.

The Data is backed up using IBM Tivoli Storage Manager (TSM) which writes to IBM 3584 robotic tape libraries. Data is vaulted between libraries in physically separate data centres to provide off site disaster recovery. All hardware is hosted in two data centres with UPS power protection and air conditioning. All services are monitored by ITCS operations team, who routinely replace failed components with no service disruption due to in built resilience. Also, the researcher can restore their own files. This process can also be done by the UEA IT Helpdesk on request, and if the file that needs recovering is not in the snapshots and has been moved to tape.

How will you manage access and security?

The permissions of the central file store are restricted to the researcher specific UEA user account, and nobody else will have access to this storage area. It is mapped to a Windows drive when the researcher log on to a UEA Windows PC using their UEA username.
There is no encryption on the CFS itself but the researcher can use a third party application called **TrueCrypt** to encrypt a container file on the CFS.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The individual data views that are created to be used for analysis will be destroyed or maintained in accordance with the UEA research data policy as appropriate given confidentiality concerns and the requirements of the data owners, the BI Intelligence Unit, which will retain ownership and control of the data at all times.

What is the long-term preservation plan for the dataset?

At the end of the project, all the data files generated will revert back to the BI Unit who will decide on what data could be made available following the UEA Research Data and which should be destroyed.

Data Sharing

How will you share the data?

The researcher will not share the data with any corporations or individuals external to UEA. The researcher and her supervisory team will be the only ones with access the raw data with external and internal assessors or markers having access to unpublished results as specified earlier in the document . Also, all published data will be anonymized. In addition, the information compliance manager will add an alteration request to their list of proposed changes on Student data protection Notice for 2014 summer. This alteration will make students aware that their data may be used for research purposes.

Are any restrictions on data sharing required?

Yes, the data must not be shared with any organization or individuals outside UEA.

Responsibilities and Resources

Who will be responsible for data management?

The researcher will be responsible for carrying out the actions required by this plan and report them to the project supervisor as appropriate. This plan will be review by the researcher with her supervisor every 6 months and update if needed.

Appendix F

Questionnaire Survey Ethics Documents

Project Synopsis

There is currently an increasing interest in educational data mining. Higher Education institutions require an improvement of their educational quality to be more competitive; therefore the application of data mining in this setting is becoming very interesting to both university administrators and researchers. Recommender systems are widely utilised in various areas, mainly in e-commerce to support customer decisions. Lately, they are also employed in learning tasks such as recommending appropriate modules, books, papers etc. to the learners (students). In this research, we investigate the use of data mining techniques in an educational setting to highlight performance problems early on and propose remedial actions. We also investigate recommender systems in a higher educational setting. We propose a recommender system that may guide students towards better module choices to increase their chances of a good outcome, based on their prior performance and other similar students' prior performances. We compare different prediction models in the context of recommender systems. We validate our results by utilising data relating to students with different characteristic from different schools. We will also investigate how to make the recommender system acceptable to students, and how to utilise the available information to improve students' outcomes. Our research's end results will enable us to provide recommendation about the quality of the used data to improve the University data warehouse, about the technical aspects of building a recommender system and about the management aspects of deploying such system.

In this stage of the ongoing PhD research, we aim to survey the students who already made their module choices about their attitudes towards a recommender system, and what additional information they think it should be available to help them make a better decisions regarding their modules choices and how they have made their module choices. The collected data use will be 'fair', because the outcome will not impact the individuals involved. The data will not be shared with any other organization. In addition, the research results will be anonymized; thus it will not be possible to identify individual students from any results published.

Research Protocol

Appropriateness of Methods

The method used for collecting data for this survey will not be unduly intrusive, as students are completely free to choose whether to participate in the survey or not. The survey will be sent by email

to Year 3 and Year 2 students since they have experienced the process of module choices. The email will include the URL of the survey, which has been created using SurveyMonkey website. Then, Year 3 and Year 2 students will be totally free to choose to participate in the survey.

It is expected that completing the survey honestly should take on average, 10 – 15 minutes. This will be stated clearly before the participants start the survey. The survey will be piloted with two students from each school (CMP and NBS).

Sample size

Initially, the researcher aims to collect data associated with 120 participants. Sample will include students from Computing Science School (CMP) and Norwich Business School (NBS) in the University of East Anglia (UEA).

Gaining informed consent

On the survey's website, possible participants will see an online informed agreement, on the Welcome page and before the start of the survey. This includes information about how the survey data will be used, data confidentiality, how participants can withdraw from the survey at any time, and the incentive to participate. The participants will be free to agree to participate. If the participant consents to participate, he/she is advised to click 'Next` to start the survey.

Informing participants of their option to withdraw

As mentioned in the previous section, the informed consent page includes a statement signifying that participants can withdraw from the survey at any time.

Informing participants of risks or benefits of participating

The survey Welcome page will state clearly that the survey has no associated risks and how the collected data will be used. Moreover, the Welcome page will notify participants that they will enter a prize draw for a £50 Amazon.co.uk voucher if they successfully complete the survey.

Data

The researcher will not collect observational nor behavioural data. There will not be any video or audio recording, nor any computer screen recording as the purpose of this survey is just to explore the views of students regarding their module choices process. The participants are allow to take the survey from any device connected to the internet. The survey answers will be automatically

collected by the Survey Monkey tool. The researcher will be able to export the collected data to spread sheets as numbers and texts.

Ensuring data confidentiality and anonymity

The researcher will not collect personally identifiable information. There will only be an optional question as clearly stated in the welcome page of the survey, asking the participants to enter their email address if they are interested in being interviewed in the future in a different stage of the researcher's PhD (The researcher will apply for ethical approval separately prior to any recruitment for the future interview stage). The data will not be shared with any other organization. In addition, the research results will be anonymized; thus it will not be possible to identify individual students from any results published. The data will be saved in The Central File Store (CFS or U: drive), because as the UEA IT support's staff stated that it is secured and centrally managed storage which is available to users with access to PC/MAC on the UEA network, or externally via VPN. After ten years according to UEA procedures, the data will be disposed of. Also, if the data spread sheets files must be printed, hard copies will be locked securely in a drawer in the researcher's desk at UEA.

Moreover, the prize draw will be done through different URL link that is not related to the survey URL link to assure the anonymity of the participants. The prize draw URL will appear at the end of the survey for the participants. The participants will be free to click the prize draw URL link. They will be asked to enter any preferred emailed address, as stated clearly in the Welcome page.

Restricting data access

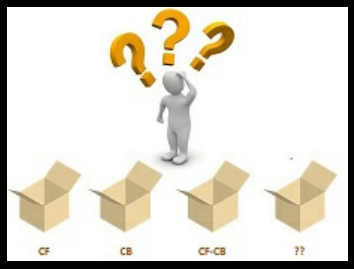
The collected data will only be accessible to the researcher and supervisors. It will also be stored and backed-up as the researcher mentioned in the data management plan that had been put in place in 2014 and approved by the Ethics committee for different stage of this PhD research. Please see the attached 'Data Management Plan document`, section 'Storage and Backup`.

Restricting data use

The Responses data will only be utilised for analysis and writing the PhD thesis. The analysis results may be published, but all data will be anonymous and individuals will not be identifiable.

Informing subjects of ethical issues

In the welcome page of the survey, the participants are advised to contact the researcher or the researchers' supervisors via the UEA email address that have been specified to them, in case of any issues or if assistance is needed.



Students' View on the Module Enrolment Process

Dear participant,

Thank you for agreeing to take part in this important survey of exploring students' thoughts and opinions about their optional module enrolment process, Your contribution is highly appreciated.

Who am I ?

My name is Zahyah Alharbi. I am currently pursuing my PhD degree in Computing Sciences School at UEA, and as part of my PhD thesis, I am doing a quantitative study on Year 3 & Years 2 students' views of how they chose their optional modules and what criteria could affect their decisions during that process.

How long will the survey take?

This survey should take approximately from 10 to 15 minutes to complete (depending on each participant).

When and Where?

This survey can be taken from any devices that connected to the internet. The survey will be available until 28 February ,2017.

What can I expect if I participate?

You will answer 19 short survey questions. We will collect the response data and analyse them to answer our research questions. We will also ask if you are interested in being interviewed.

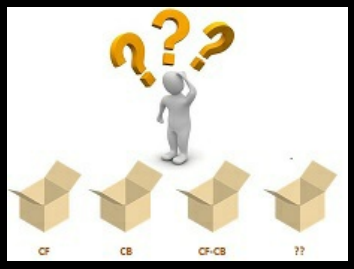
Be assured we will not require your name and your survey data will be kept confidential and anonymous. There are no risks associated with the study but you can withdraw any time while you are performing the survey.

You will also be entered into a prize draw for a £50 Amazon voucher if you successfully complete our survey. To enter this prize draw you will asked to enter any preferred email address; however you will be asked to enter your preferred email address in a different page that is not related to this survey to ensure your anonymity.

Who can I contact?

If you need any additional information, please contact me at Z.alharbi@uea.ac.uk or my supervisors Mr James Cornford at j.cornford@uea.ac.uk or Dr Beatriz de la Iglesia at b.iglesia@uea.ac.uk

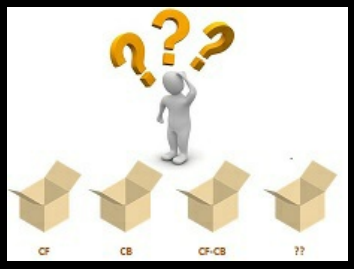
If you consent to participate in this survey based on what has been stated in this page. Please click 'Next'.



Students' View on the Module Enrolment Process

* 1. What is your school of study?

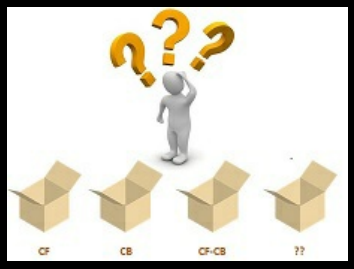
- Norwich Business School.
- School of Computing Sciences.



Students' View on the Module Enrolment Process

* 2. What is your course of study?

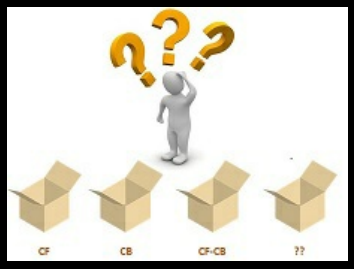
- ○ Actuarial Sciences
- ○ Applied Computing Science
- ○ Business Information Systems
- ○ Business Statistics
- ○ Computer Graphics, Imaging and Multimedia
- ○ Computer Systems Engineering
- ○ Computing Science
- ○ Other (please specify)



Students' View on the Module Enrolment Process

* 3. What is your course of study?

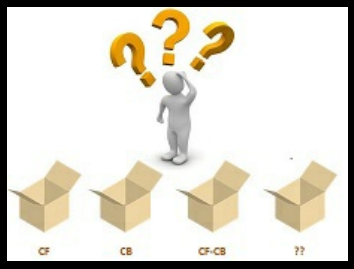
- ○ Accounting and Finance
- ○ Accounting and Management
- ○ Business Finance and Management
- ○ Business Management or Management
- ○ Marketing and Management
- ○ Other (please specify)



Students' View on the Module Enrolment Process

* 4. What is your year of study?

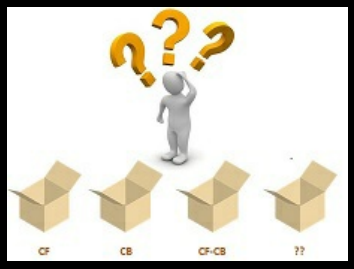
- ◯ Year 2
- ◯ Year 3



Students' View on the Module Enrolment Process

* 5. What is your gender?

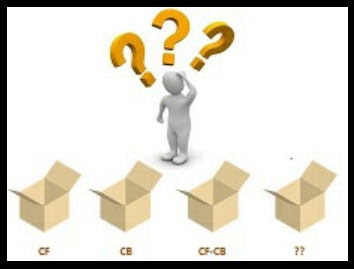
- ◯ Female
- ◯ Male



Students' View on the Module Enrolment Process

* 6. What is your age?

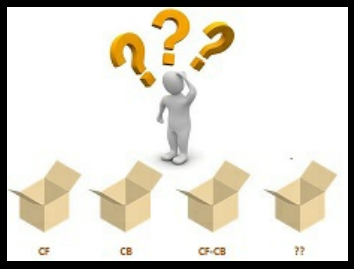
- ◯ 17-21
- ◯ 22+



Students' View on the Module Enrolment Process

* 7. What is your fee status?

- ◯ Home student
- ◯ European Union student
- ◯ Overseas student
- ◯ Don't know / prefer not to say

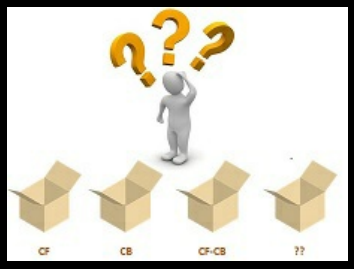


Students' View on the Module Enrolment Process

* 8. What source/s of information did you consider when you chose your optional modules? (Check all that apply)

- [ ] Outline of each module.
- [ ] Module details on e-vision catalogue (https://evision.uea.ac.uk).
- [ ] Opinion of students who have experienced the module.
- [ ] Recommendation of your academic adviser.
- [ ] Module information day /fair.

Other (please specify)

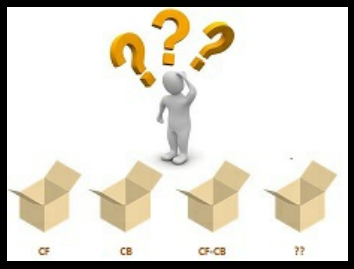


Students' View on the Module Enrolment Process

* 9. What criteria did you consider when you chose your optional module(s)? (Check all that apply)

- [ ] Interest in the topic of the module.
- [ ] Type of assessment in each module.
- [ ] Expected instructor of the module.
- [ ] Friends who could take the same module.
- [ ] Opinion of parents.
- [ ] Relevance to expected career options.
- [ ] Expected academic performance.
- [ ] Eligibility for exemption from professional exams.
- [ ] Expected time slot of the lecture.

Other (please specify)

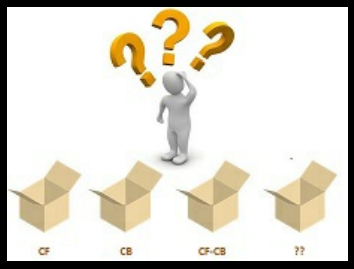


Students' View on the Module Enrolment Process

* 10. Please rank the following criteria based on your priority when choosing your optional module(s)? (1 being the highest priority and 9 being the lowest priority)

Please note each rank number is allow to be chosen once.

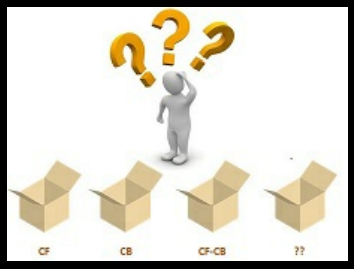
- Interest in the topic of the module.
- Type of assessment in each module.
- Expected instructor of the module.
- Friends who could take the same module.
- Opinion of parents.
- Relevance to expected career options.
- Expected academic performance.
- Eligibility for exemption from professional exams.
- Expected time slot of the lecture.



Students' View on the Module Enrolment Process

* 11. What optional module(s) have you chosen? (please enter at least 1 module)

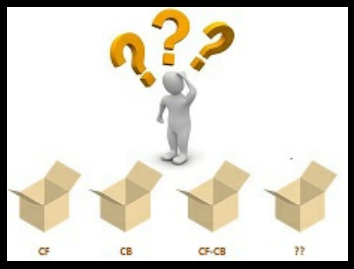
Module 1	
Module 2 (optional)	
Module 3 (optional)	
Module 4 (optional)	
Module 5 (optional)	
Module 6 (optional)	



Students' View on the Module Enrolment Process

* 12. Recently, universities have started to use advanced data mining techniques to provide personalised predictions of module performance. For example, in the process of your module choice, the university enrolment system could show you your predicted mark based on previous students with similar personal characteristics. **How much would you value this personalised prediction of module performance in making your module choice(s)?**

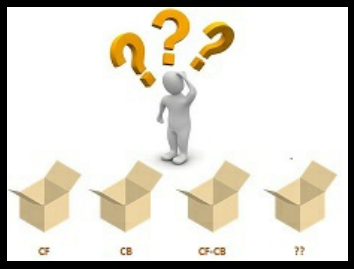
- Extremely valuable
- Very valuable
- Moderately valuable
- Slightly valuable
- Not at all valuable



Students' View on the Module Enrolment Process

* 13. Would you be interested to know your personal predicted marks for the modules you are currently studying?

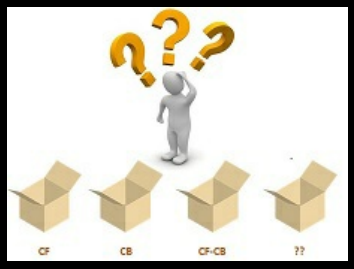
- ◯ No
- ◯ Yes



Students' View on the Module Enrolment Process

* 14. Do you think that knowing your personal predicted marks would have affected your decisions in choosing your optional modules?

- ◯ No
- ◯ Yes

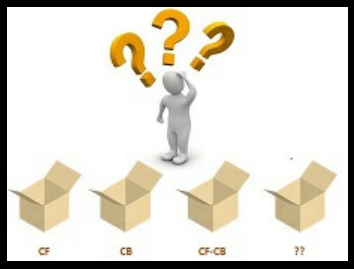


Students' View on the Module Enrolment Process

* 15. Currently, universities can provide personal enrolment recommender programmes that can suggest optional modules based on a student's expected performance, expected student satisfaction and his /her career choices. The recommendation would be personalised, that is based on previous students with similar personal characteristics.

Would you have been interested to have had such a broadly-based programme during your module enrolment process?

- ○ No
- ○ Yes

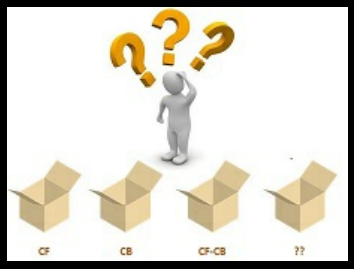


Students' View on the Module Enrolment Process

* 16. In your opinion, what additional information would be helpful in choosing your optional module(s)? (Check all that apply)

- [ ] An average mark based on the past few years of students marks.
- [ ] Your predicted mark based on previous students with similar personal characteristics.
- [ ] General career opportunities associated with the module.
- [ ] Personalised career opportunities based on previous students with similar personal characteristics who took the module.
- [ ] General satisfaction rate of students who took the same module in the past few years.
- [ ] Your predicted satisfaction rate based on students with similar personal characteristics.
- [ ] An average post graduation salary of students who took the module.

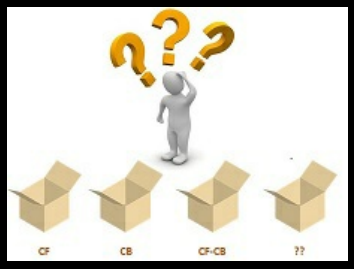
Other (please specify)



Students' View on the Module Enrolment Process

* 17. In your opinion, would it be helpful to know the previous students' evaluation of the module instructor during your module enrolment process?

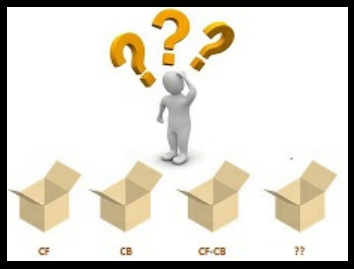
- ○ No
- ○ Yes



Students' View on the Module Enrolment Process

* 18. On a scale from 1 (being the lowest) to 7 (being the highest), how useful would the following information be in making your module choices?

	1 (lowest)	2	3	4	5	6	7 (highest)
An average mark based on the past few years of students marks.	○	○	○	○	○	○	○
Your predicted mark based on previous students with similar personal characteristics.	○	○	○	○	○	○	○
General career opportunities associated with the module.	○	○	○	○	○	○	○
Personalised career opportunities based on previous students with similar personal characteristics who took the module.	○	○	○	○	○	○	○
General satisfaction rate of students who took the same module in the past few years.	○	○	○	○	○	○	○
Please select the third scale (circle).	○	○	○	○	○	○	○
Your predicted satisfaction rate based on students with similar personal characteristics.	○	○	○	○	○	○	○
An average post graduation salary of students who took the module.	○	○	○	○	○	○	○



Students' View on the Module Enrolment Process

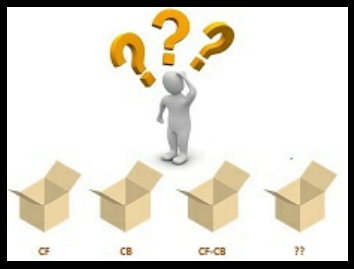
* 19. Would you be willing to be interviewed for 20 minutes to discuss the process of optional module enrolment and to find out what your predicted marks are?

(1- You will be asked to give your informed consent before the interview.

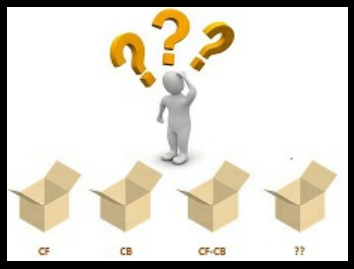
2- We will not require your name and all your recorded data will be anonymous.

3- A **£10** Amazon voucher will be emailed to the selected students after they complete their interview.)

- ◯ Yes
- ◯ No



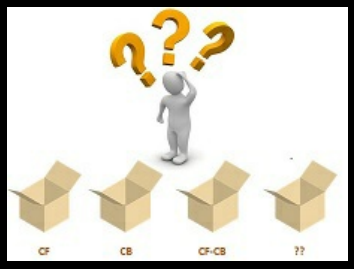
20. If yes, at what email address would you like to be contacted? (optional)



Students' View on the Module Enrolment Process

This question is only for the pilot study's respondents.

21. Do you have any comments, other questions you think we should add, or concerns?



Students' View on the Module Enrolment Process

Thank you for completing our survey !

please click Prize Draw Survey to enter our prize draw.

The Prize Draw

This survey is to collect the responders' email address for our prize draw.

Be assured this survey is not connected to the previous one, which means your answers in the main survey is completely anonymous.

1. At what email address would you like to be contacted, if you win the prize?

Thank you !

Appendix G

Interviews Ethics Documents

Project Synopsis

There is currently an increasing interest in educational data mining. Higher Education institutions require an improvement of their educational quality to be more competitive; therefore the application of data mining in this setting is becoming very interesting to both university administrators and researchers. Recommender systems are widely utilised in various areas, mainly in e-commerce to support customer decisions. Lately, they are also employed in learning tasks such as recommending appropriate modules, books, papers etc. to the learners (students). In this research, we investigate the use of data mining techniques in an educational setting to highlight performance problems early on and propose remedial actions. We also investigate recommender systems in a higher educational setting. We propose a recommender system that may guide students towards better module choices to increase their chances of a good outcome, based on their prior performance and other similar students' prior performances. We compare different prediction models in the context of recommender systems. We validate our results by utilising data relating to students with different characteristic from different schools. We will also investigate how to make the recommender system acceptable to students, and how to utilise the available information to improve students' outcomes. Our research's end results will enable us to provide recommendation about the quality of the used data to improve the University data warehouse, about the technical aspects of building a recommender system and about the management aspects of deploying such system.

In this stage of the ongoing PhD research, we aim to interview some consenting students who already made their module choices about their thoughts towards a personalised recommender system, and what additional information they think it should be available to help them make a better decisions regarding their modules choices and how they have made their module choices. We also aim to interview some academic staff at various levels about their thoughts regards the use of personalised recommender system. The collected data use will be 'fair', because the outcome will not impact the individuals involved. The data will not be shared with any other organization. In addition, the research results will be anonymized; thus it will not be possible to identify individual students or staff from any results published.

Research Protocol

Appropriateness of Methods

The method used for collecting data for this interviews will not be unduly intrusive, as students and academic are completely free to choose whether to participate in the interview or not. An email will be sent to both selected academic staff and selected Year 3 and Year 2 students since they have experienced the process of module choices. The selection of students is based on their acceptance to be interviewed in a previous stage of this research. The selection of staff is based on their role in the management and delivery of undergraduate teaching and learning. The email will include introduction of the interview topic, the proposed questions, and the consent information. Then, the students and the staff will be totally free to choose to participate in the interview. If they accept to participate, a second email will be sent to schedule an appointment that is convenient for them.

It is expected that the interview should take on average 20 – 30 minutes with student interviewees and approximately 45 minutes with the academic staff. This will be stated clearly before the participants start the interview. The interviews will be piloted with one students and one academic staff. The interview will be digitally recorded and transcribed by the researcher. The participant will be informed clearly about the digital recording and the written transcription in the consent form before starting the interview. If the participant decline to be recorded then he/she will not be able to participate in this research. A copy from the written transcription will be sent to the participants if they wish to check that their views have been appropriately presented.

Sample size

Initially, the researcher aims to collect data associated with 27 student participants and 7 academic staff participants. The sample will include students from Computing Science School (CMP) and Norwich Business School (NBS) in the University of East Anglia (UEA). The sample will also include UEA academic staff at various levels.

Gaining informed consent

Before the start of the interview, the participants will read the consent form(s) and ask questions if needed. The consent form includes information about how the interview data will be used, data confidentiality, how participants can withdraw from the interview at any time, how long the interview should take and the incentive to participate. The participants will be free to agree to participate. If the participant consents to participate, he/she is advised to sign and date the consent form. A copy of the

signed and dated consent form will be given to the participant and the original dated and signed copy will be saved in a secure place in the research file.

Informing participants of their option to withdraw or to decline answering a question

As mentioned in the previous section, the informed consent form includes a statement signifying that participants can end the interview at any time. Also, it state that participants are free to decline answering any of the interview questions.

Informing participants of risks or benefits of participating

The interview consent form will state clearly that the interview has no associated risks and how the collected data will be used. Moreover, the consent form will notify student participants that they will receive a £10 Amazon.co.uk voucher if they complete the interview. The consent form will also notify staff participants that they will not be paid for their participation.

Data

The researcher will not collect observational nor behavioural data. There will not be any video recording, nor any computer screen recording. However, there will be audio recording as the purpose of this interview is just to explore the views of selected students and academic staff regarding the uses of personalised recommender system. A written transcription (using Microsoft Word software) will be made and sent to the participants if they wish to check that their views have been appropriately presented. Then, the researcher will be able to analyse the written transcription using **NVivo** which is **qualitative** data analysis software.

Feedback to participants

As mentioned previously, the collected data will be sent to the participants if they wish to check that their views have been appropriately presented.

Ensuring data confidentiality and anonymity

The researcher will not collect personally identifiable information. The participants name will appear only in the consent form and is clearly stated in the consent form that their names are not associated with the interview audio recording, nor the written transcription nor to any research materials. The data will not be shared with any other organization. In addition, the research results will be anonymized; thus it will not be possible to identify individual students nor staff from any results published. The data will be saved in The Central File Store (CFS or U: drive), because as the UEA IT support's staff stated that it is secured and centrally managed storage which is available to

users with access to PC/MAC on the UEA network, or externally via VPN. After ten years according to UEA procedures, the data will be disposed of. Also, if the data spread sheets files must be printed, hard copies will be locked securely in a drawer in the researcher's desk at UEA.

Moreover, the student participants will be asked to provide the researcher with their preferred email to send their incentive voucher to it. However, it will be clearly stated in the consent form that the provided email will not be linked to any of the research materials. The participants will be free to provide their preferred email.

There is additional consent form will be provided to academic staff participants only about mentioning their general role in the management and delivery of undergraduate teaching and learning at UEA in the research result. If the staff participant consents to give permission to mention their general role in the result, the researcher will be able to use the general role information in the research result if needed. However, if the staff participant does not consent to give permission to mention their general role in the PhD result, then the researcher will avoid use the role information and assure the anonymization of the staff participant general role in any of the research materials.

Restricting data access

The collected data will only be accessible to the researcher and supervisors. It will also be stored and backed-up as the researcher mentioned in the data management plan that had been put in place in 2014 and approved by the Ethics committee for different stage of this PhD research. Please see the attached 'Data Management Plan document`, section 'Storage and Backup`.

Restricting data use

The "Responses" data will only be utilised for analysis and writing the PhD thesis. The analysis of the results may be published, but all data will be anonymous and individuals will not be identifiable.

Informing subjects of ethical issues

Before the interview starts, the participants are verbally advised to contact the researcher or the researchers' supervisors via the UEA email address that have been specified to them, in case of any issues or if assistance is needed.

Proposed Topic Guide for Student Interview

The following are the student interview questions:

1. What is your course of study?
2. Are you pleased with your current module choices?
3. In your opinion, do you think the University provides enough information to help you make your module choice? (If yes/no, why?)
4. If the University decided to implement a Personalised Recommender System, do you think you would like to use it?
5. Would you still use other resources along with the Recommender System to make your module choice? If yes, which and why?
6. Do you think your decision would be mainly based on the Recommender System? If so, why?
7. Is there any other information you want to see used within a Recommender System?
8. What, from your personal point of view, would be the potential advantages/ enablers, if any, of a personalised recommender system for the student?
9. What, from your personal point of view, would be the potential disadvantages/ challenges, if any, of a personalised recommender system for the student?
10. From a student perspective, do you have any other thoughts about the acceptability of a recommender system for undergraduate module choice in the University?
11. Would you like to find what people with your characteristic predictive mark would have been in any of your elective modules?
12. By knowing that predictive mark hypothetically, do you think it would have influenced your choice of module?
13. Do you consider that influence is positive or negative? If so how?
    a. Imagine the predicted mark is high (based on what you consider high as it is vary for each student), would it have influenced your choice of module, Do you consider that influence is positive or negative? If so how?
    b. However, imagine the predicted mark is low (based on what you consider low as it is vary for each student), would it have influenced your choice of module, Do you consider that influence is positive or negative? If so how?

Proposed Topic Guide for Staff Interview

Introduction

Universities have recently become much more interested in the use of various forms of data and the techniques of predictive analytics - what is sometimes called big data- to support decision making by, for example, improving the targeting resources, identifying potential opportunities for "early intervention". In general, these techniques have been used by university managers to support decision making and resources allocation. There is also some interest in using these techniques to support students' choice of university and course. However, these techniques could also be used by students to support their decision-making within their course of study, as well as their choice of course of study. In this research, we are focusing on the main formal decision that many undergraduate students face: the choice of elective modules.

Using data from UEA's Data Warehouse, we have been able to create a reasonably accurate model that can provide an individualised prediction of student academic performance on any given module (for which data exists) for two Schools of study. The model can provide a statement of the form, 'on the basis of past student performance, a student matching the characteristics which we have for you can be predicted to have a score of X in this module'. Such software is referred to generically as a recommender system. However, sharing such information with students raises a number of practical and ethical issues and there may be complex and potentially harmful unanticipated outcomes. Before universities adopt such technologies, or permit their development using university data, we want to understand these issues more clearly.

We are surveying students who have already made their module choices about their attitudes towards a recommender system and we will conduct short interviews with selected, consenting students which will include offering them the opportunity to find out their predictive outcome and analyse if they think that may have altered their module choices.

We are also concerned to understand the attitudes of academic staff at various levels, to the use of personalised recommender system.

Next, we introduced the academic staff interview questions:

1. Can you tell us a little about your general role in the management and delivery of undergraduate teaching and learning at UEA? (Or what use do you have for 'data' and analytics in your role?
2. Can you tell us about how your role relates to undergraduate student module choice?
3. Can you tell us about what kinds of information are currently provided to support student module choice within UEA generally or within your School? Can we have access to copies of any materials used last year or those that are proposed for this year to assist with module choice?
4. Can you tell us what do you believe students ***do*** use to make module choices?
5. What information do you believe students ***should*** use to make an informed module choice?
6. What, from your personal and professional point of view, would be the potential advantages/ enablers, if any, of a personalised recommender system for the university and for student?

7. What, form your personal and professional point of view, would be the potential disadvantages/ challenges, if any, of a personalised recommender system for the university and for student?
8. Do you have any other thoughts about the acceptability of a recommender system for undergraduate module choice at UEA?
9. By analysing first year 1 data, we may be able to identify students with poor performance in terms of good honour outcomes with reasonable accuracy. How, from your personal and professional point of view, **should** the University act on those findings in term of improving students' outcomes? What **could/should** the University offer to those at risk? Should assistance be offered to all students or those at risk?

University of East Anglia

Consent Form for Interviews - (Student Participant)

I volunteer to participate in a PhD research study conducted by Zahyah Alharbi from the University of East Anglia. I understand that the research is designed to collect information about the acceptance of using Personalised Recommender System for module choice. I will be one of approximately 27 people being interviewed for this research.

1. My participation in this interview is voluntary. I may withdraw at any time without penalty. If I withdraw from the interview, no one on campus or elsewhere will be told.

2. I understand that if I feel uncomfortable in any way during the interview session, I have the right to decline to answer any question or to end the interview.

3. I understand that after I complete my interview I will receive an Amazon.co.uk email voucher in the amount of **£10.** I agree to give the researcher my preferred email to receive the incentive voucher. The researcher will not link the provided email to any of the research materials.

4. The interview will last approximately 20 -30 minutes. Notes will be written during the interview. An audio recording of the interview and written transcription will be made. If I don't want to be recorded, I will not be able to participate in this research.
5. I understand that the information collected in this interview is for PhD purposes only and there are no risks associated with the study.
6. I understand that the researcher will not identify me by name in any reports/publications/thesis using information obtained from this interview, and that my confidentiality as a participant in this research will remain strictly secure. Subsequent uses of data and records will be subject to standard data use policies which protect the anonymity of individuals and institutions.
7. I understand that my name in the consent form is not associated with the interview audio recording, the written transcription nor to any research materials.
8. I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this research.

_____	_____	_____
Name of participant	Date	Signature
_____	_____	_____
Name of researcher	Date	Signature

***Copies**: Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form. A copy of the signed and dated consent form should be placed in the research file which must be kept in a secure location.*

University of East Anglia

Consent Form for Interviews - (Staff Participant)

I volunteer to participate in a PhD research study conducted by Zahyah Alharbi from the University of East Anglia. I understand that the research is designed to collect information about the acceptance of using Personalised Recommender System for module choice. I will be one of approximately 27 people being interviewed for this research.

1. My participation in this interview is voluntary. I may withdraw at any time without penalty. If I withdraw from the interview, no one on campus or elsewhere will be told.

2. I understand that if I feel uncomfortable in any way during the interview session, I have the right to decline to answer any question or to end the interview.

3. I understand that I will not be paid for my participation in this interview.

4. The interview will last approximately 45 minutes. Notes will be written during the interview. An audio recording of the interview and written transcription will be made. If I don't want to be recorded, I will not be able to participate in this research.
5. I understand that the information collected in this interview is for PhD purposes only and there are no risks associated with the study.
6. I understand that the researcher will not identify me by name in any reports/publications/thesis using information obtained from this interview, and that my confidentiality as a participant in this research will remain strictly secure. Subsequent uses of data and records will be subject to standard data use policies which protect the anonymity of individuals and institutions.
7. I understand that my name in the consent form is not associated with the interview audio recording, the written transcription nor to any research materials.
8. I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this research.

_____	_____	_____
Name of participant	Date	Signature
_____	_____	_____
Name of researcher	Date	Signature

***Copies**: Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form. A copy of the signed and dated consent form should be placed in the research file which must be kept in a secure location.*

University of East Anglia

Consent Form for Staff Participant only

I confirm that I give permission for the researcher to mention my general role in the management and delivery of undergraduate teaching and learning at UEA in the written research results.

_____	_____	_____
Name of participant	Date	Signature
_____	_____	_____
Name of researcher	Date	Signature

***Copies**: Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form. A copy of the signed and dated consent form should be placed in the research file which must be kept in a secure location.*

Appendix H

F1-score Results

Table H.1: Comparison of F1-score mean values for each prediction system for the first school datasets. The standard deviation is in brackets.

	DS	NW	IT	SA	SE
SimpleAvg	0.217(0.095)	0.571(0.122)	0.211(0.180)	0.501(0.153)	0.571(0.158)
Rpart	0.729(0.106)	0.754(0.092)	0.154(0.253)	0.717(0.115)	0.720(0.118)
Rpart_{RF}	0.688(0.137)	0.716(0.080)	0.451(0.147)	0.726(0.060)	0.779(0.084)
Rpart_{MI}	0.676(0.093)	0.732(0.065)	0.500(0.098)	0.718(0.070)	0.701(0.099)
Rpart_{EM}	0.651(0.101)	0.739(0.043)	0.615(0.058)	0.682(0.061)	0.714(0.086)
C5	0.681(0.120)	0.700(0.082)	0.438(0.150)	0.707(0.103)	0.750(0.123)
C5_{RF}	0.670(0.110)	0.690(0.111)	0.522(0.122)	0.703(0.102)	0.660(0.127)
C5_{MI}	0.673(0.086)	0.728(0.068)	0.470(0.077)	0.710(0.075)	0.692(0.107)
C5_{EM}	0.685(0.088)	0.732(0.0532)	0.600(0.066)	0.688(0.057)	0.711(0.114)
RF	0.725(0.103)	0.751(0.080)	0.540(0.160)	0.763(0.077)	0.741(0.132)
RF_{RF}	0.710(0.120)	0.717(0.052)	0.499(0.216)	0.699(0.084)	0.715(0.146)
RF_{MI}	0.726(0.049)	0.758(0.085)	0.530(0.114)	0.736(0.070)	0.718(0.192)
RF_{EM}	0.719(0.102)	0.788(0.054)	0.674(0.080)	0.760(0.080)	0.728(0.169)
SVM	0.647(0.064)	0.707(0.084)	0.436(0.146)	0.728(0.087)	0.715(0.125)
SVM_{RF}	0.718(0.132)	0.764(0.073)	0.556(0.173)	0.723(0.066)	0.695(0.163)
SVM_{MI}	0.717(0.092)	0.776(0.063)	0.526(0.106)	0.748(0.067)	0.732(0.112)
SVM_{EM}	0.682(0.082)	0.805(0.040)	0.624(0.067)	0.756(0.052)	0.732(0.095)

Table H.2: Comparison of F1-score mean values for each prediction system for the second school datasets. The standard deviation is in brackets.

	EB	PT	IS	SM	FM
SimpleAvg	0.851(0.019)	0.568(0.093)	0.460(0.231)	0.691(0.053)	0.884(0.042)
Rpart	0.859(0.029)	0.754(0.063)	0.735(0.085)	0.816(0.029)	0.946(0.033)
Rpart_{RF}	0.859(0.022)	0.768(0.045)	0.742(0.078)	0.806(0.030)	0.950(0.030)
Rpart_{MI}	0.863(0.027)	0.755(0.057)	0.737(0.067)	0.807(0.029)	0.940(0.028)
Rpart_{EM}	0.864(0.023)	0.787(0.059)	0.728(0.071)	0.805(0.028)	0.940(0.0271)
C5	0.870(0.027)	0.766(0.059)	0.718(0.093)	0.825(0.046)	0.953(0.028)
C5_{RF}	0.855(0.037)	0.773(0.059)	0.710(0.092)	0.813(0.044)	0.953(0.028)
C5_{MI}	0.861(0.024)	0.763(0.053)	0.723(0.060)	0.818(0.039)	0.953(0.028)
C5_{EM}	0.863(0.023)	0.775(0.057)	0.720(0.072)	0.809(0.038)	0.952(0.028)
RF	0.877(0.029)	0.767(0.093)	0.752(0.078)	0.839(0.038)	0.953(0.028)
RF_{RF}	0.856(0.033)	0.788(0.028)	0.714(0.057)	0.835(0.0189)	0.946(0.032)
RF_{MI}	0.863(0.034)	0.767(0.071)	0.738(0.070)	0.835(0.037)	0.944(0.034)
RF_{EM}	0.862(0.024)	0.779(0.077)	0.722(0.077)	0.843(0.036)	0.946(0.030)
SVM	0.846(0.027)	0.777(0.094)	0.768(0.061)	0.844(0.041)	0.953(0.028)
SVM_{RF}	0.879(0.024)	0.786(0.088)	0.757(0.057)	0.842(0.037)	0.953(0.028)
SVM_{MI}	0.880 (0.021)	0.777 (0.081)	0.768 (0.061)	0.836 (0.035)	0.953(0.028)
SVM_{EM}	0.881(0.025)	0.796(0.069)	0.754(0.062)	0.837(0.037)	0.953(0.028)

Table H.3: Comparison of F1-score mean values for each prediction system for the publicly available datasets. The standard deviation is in brackets.

	Math	Math 25%	Math 45%	Por	Por 25%	Por 45%
SimpleAvg	NA	NA	NA	NA	NA	NA
Rpart	0.818(0.115)	0.846(0.075)	0.839(0.091)	0.850(0.040)	0.857(0.038)	0.856(0.040)
Rpart_{RF}	NA	0.857(0.078)	0.843(0.065)	NA	0.859(0.028)	0.881(0.030)
Rpart_{MI}	NA	0.845(0.058)	0.846(0.060)	NA	0.853(0.040)	0.862(0.034)
Rpart_{EM}	NA	0.850(0.060)	0.850(0.054)	NA	0.853(0.042)	0.868(0.027)
C5	0.808(0.092)	0.865(0.061)	0.857(0.087)	0.862(0.044)	0.865(0.052)	0.858(0.047)
C5_{RF}	NA	0.816(0.063)	0.829(0.055)	NA	0.861(0.040)	0.849(0.040)
C5_{MI}	NA	0.838(0.072)	0.843(0.054)	NA	0.859(0.040)	0.874(0.022)
C5_{EM}	NA	0.844(0.056)	0.847(0.052)	NA	0.859(0.039)	0.870(0.028)
RF	0.824(0.111)	0.832(0.063)	0.851(0.076)	0.873(0.040)	0.858(0.059)	0.873(0.032)
RF_{RF}	NA	0.836(0.111)	0.812(0.083)	NA	0.882(0.033)	0.871(0.038)
RF_{MI}	NA	0.846(0.064)	0.855(0.056)	NA	0.876(0.038)	0.888(0.022)
RF_{EM}	NA	0.839(0.074)	0.844(0.062)	NA	0.869(0.045)	0.888(0.027)
SVM	0.791(0.147)	0.826(0.112)	0.783(0.109)	0.867(0.031)	0.869(0.046)	0.859(0.035)
SVM_{RF}	NA	0.835(0.083)	0.820(0.142)	NA	0.871(0.032)	0.856(0.045)
SVM_{MI}	NA	0.834(0.092)	0.827(0.095)	NA	0.882(0.035)	0.883(0.018)
SVM_{EM}	NA	0.829(0.083)	0.824(0.080)	NA	0.881(0.029)	0.877(0.029)

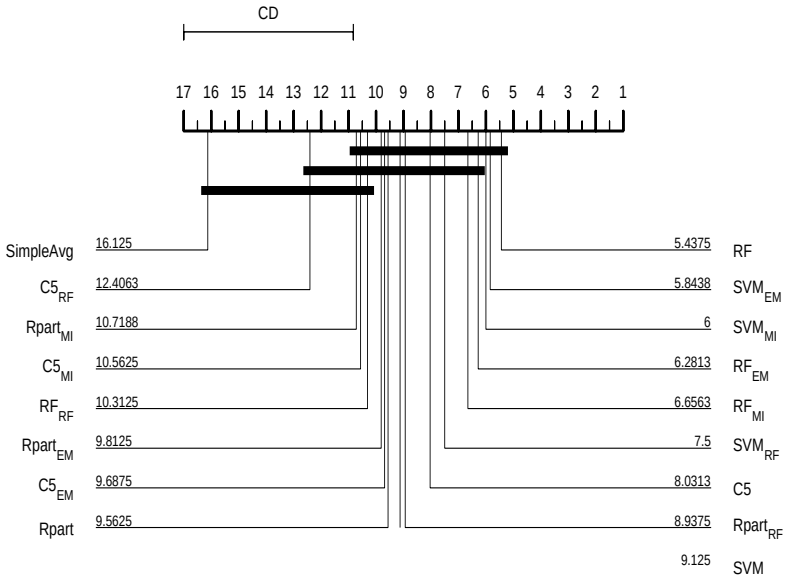


Figure H.1: Critical difference diagram for the classification F1-score across the 16 datasets associated with both schools of study and the external institution. The decimal number that close to each prediction system is the values of its average rank that is used in the Friedman test computation.

Appendix I

External Datasets Clustering Results

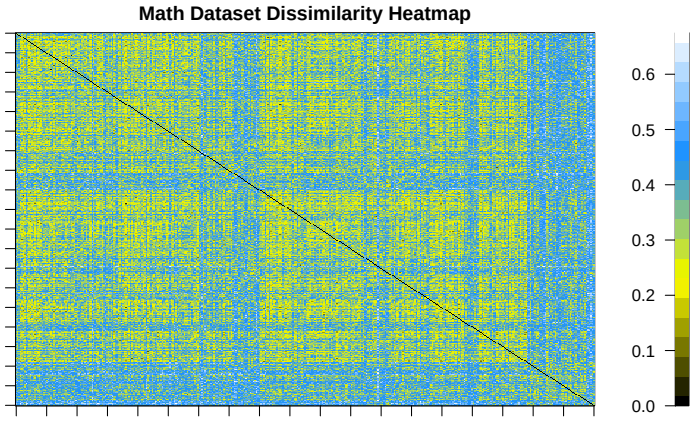


Figure I.1: This heatmap shows the dissimilarity between students in the Maths subject *complete* dataset. The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.6.

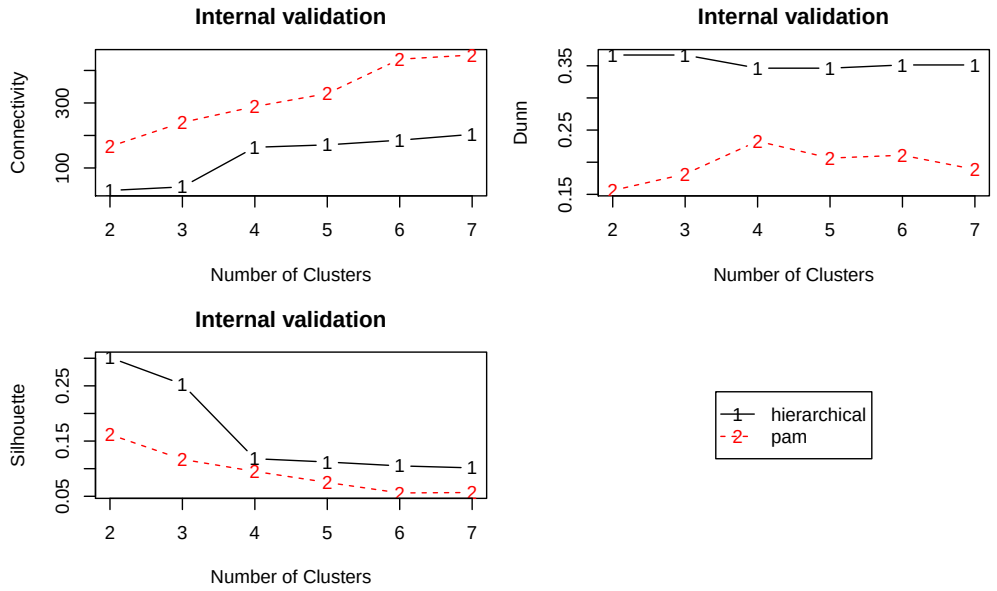


Figure I.2: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Maths subject *complete* dataset. The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

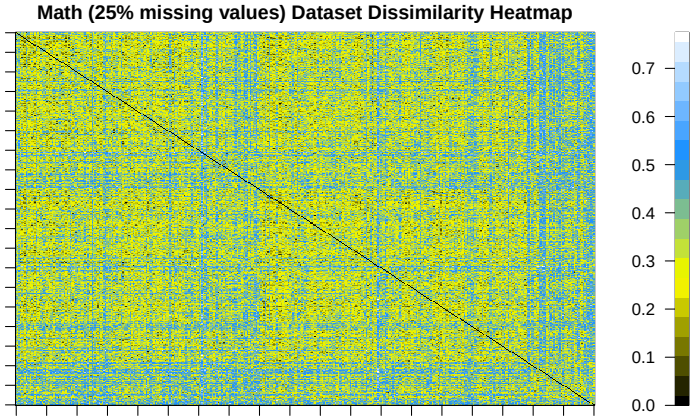


Figure I.3: This heatmap shows the dissimilarity between students in the Maths subject dataset (*include 25% missing values*). The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.7.

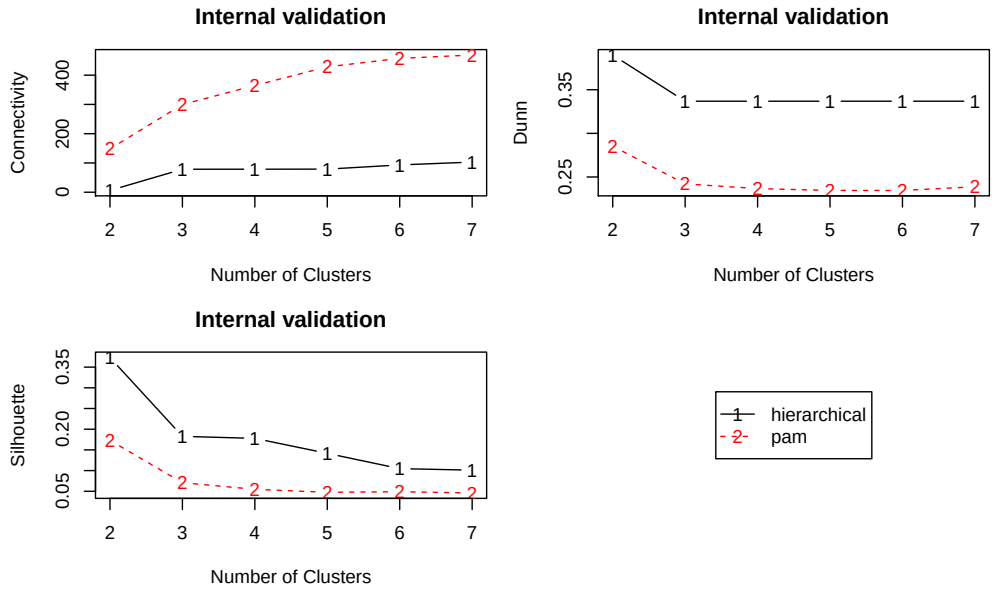


Figure I.4: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Maths subject dataset (*include 25% missing values*). The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

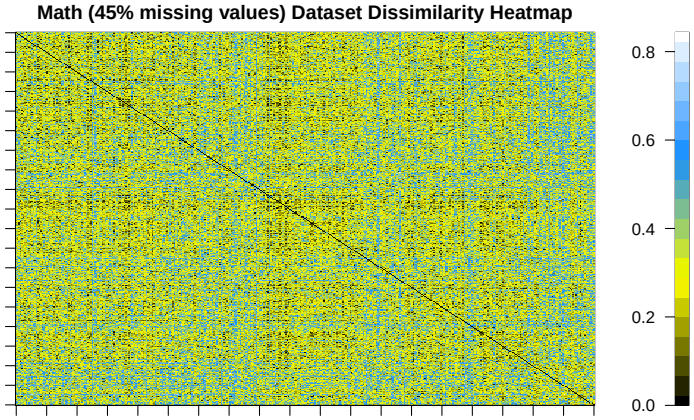


Figure I.5: This heatmap shows the dissimilarity between students in the Maths subject dataset (*include 45% missing values*). The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.8.

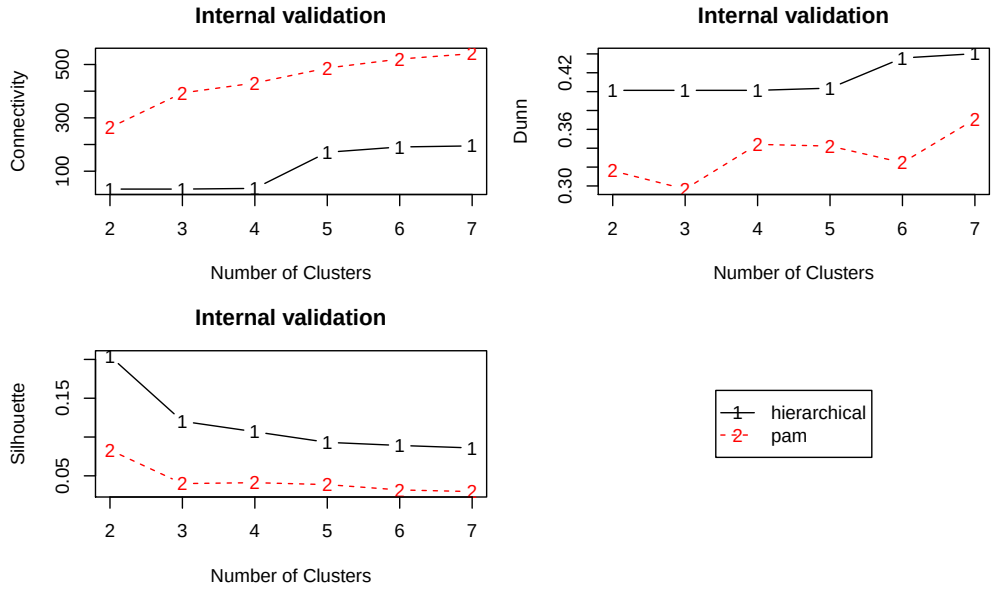


Figure I.6: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Maths subject dataset (*include 45% missing values*). The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

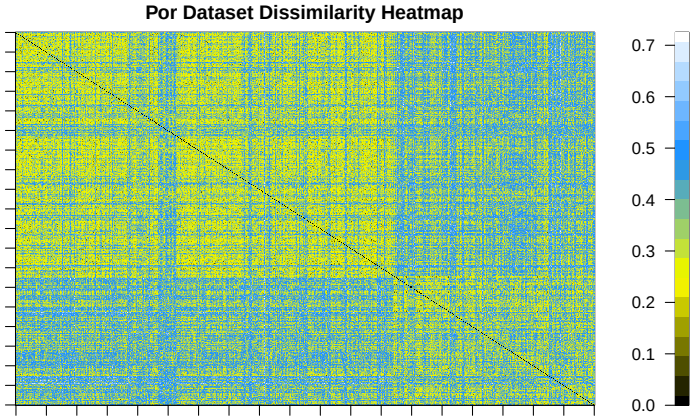


Figure I.7: This heatmap shows the dissimilarity between students in the Portuguese subject *complete* dataset. The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.7.

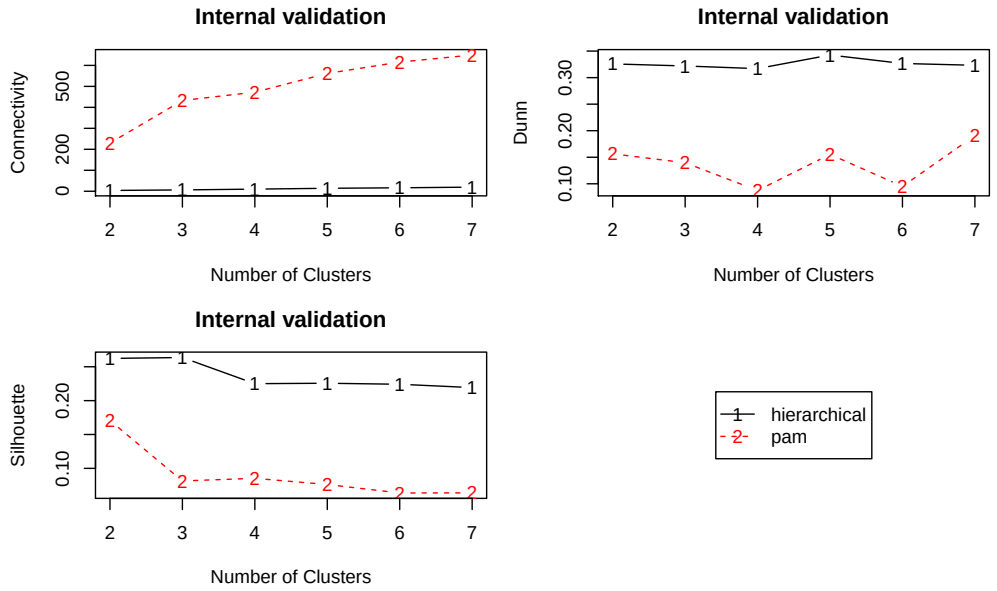


Figure I.8: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Portuguese subject *complete* dataset. The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

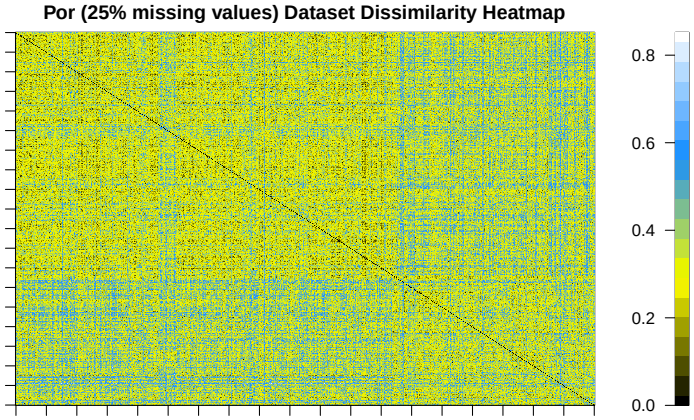


Figure I.9: This heatmap shows the dissimilarity between students in the Portuguese subject dataset (*include 25% missing values*). The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.8.

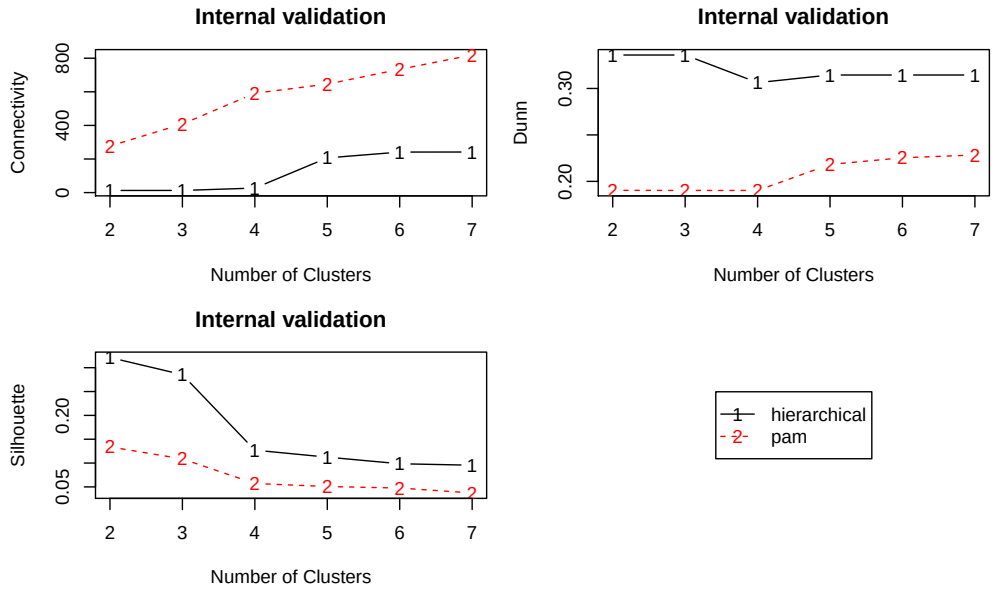


Figure I.10: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Portuguese subject dataset (*include 25% missing values*). The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

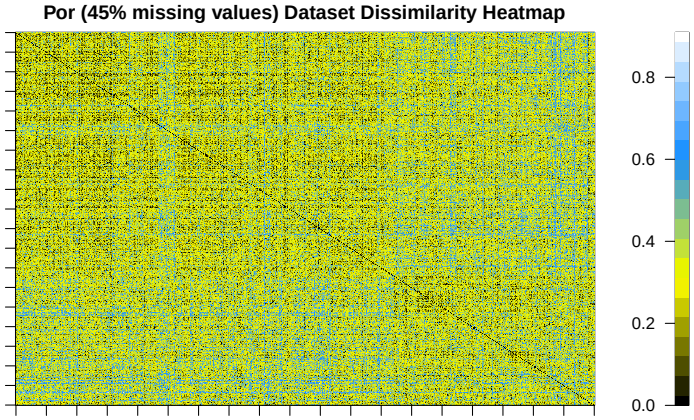


Figure I.11: This heatmap shows the dissimilarity between students in the Portuguese subject dataset (*include 45% missing values*). The black scale reflects strong similarity ≤ 0.1 and it scales through yellow until it reaches the white colour to reflect dissimilarity > 0.8.

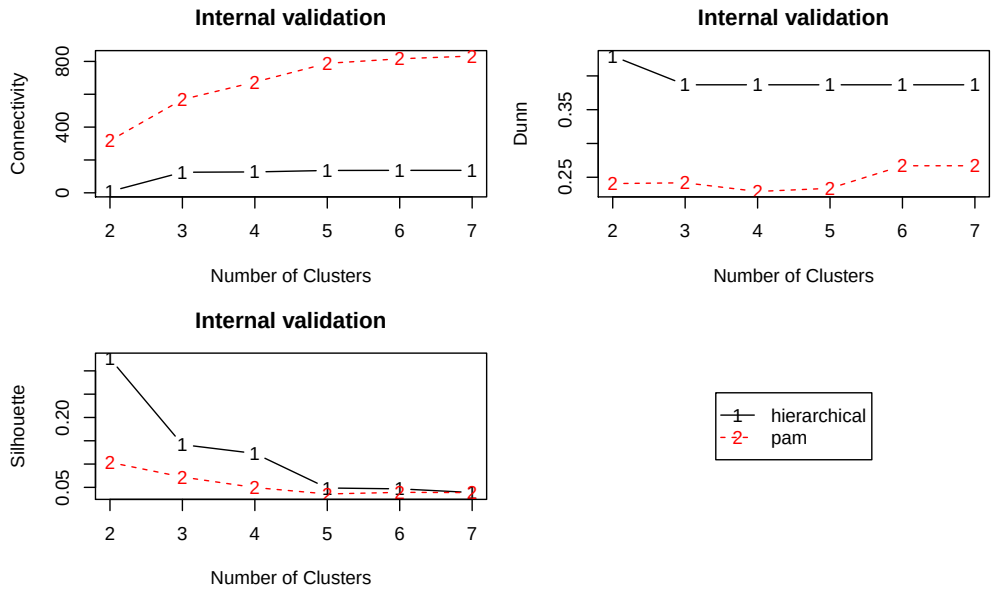


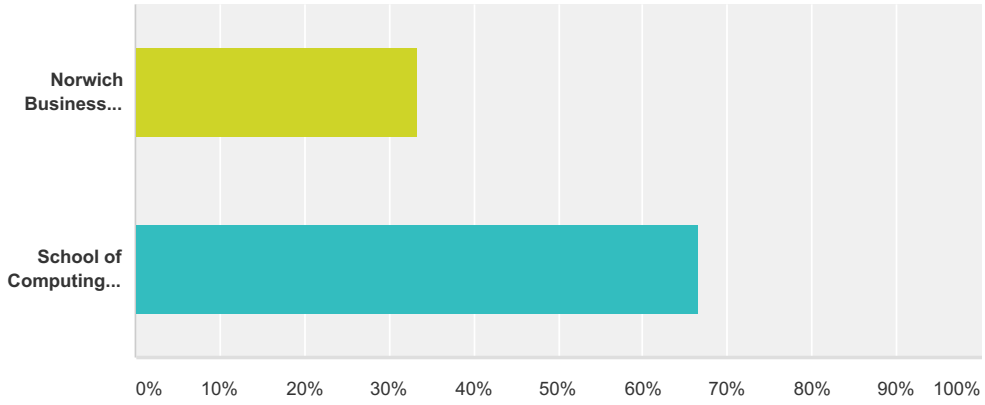
Figure I.12: Results of the internal validation (Connectivity, Dunn, Silhouette) of PAM and hierarchical clustering for the Portuguese subject dataset (*include 45% missing values*). The x-axis shows the number of clusters (from 2 to 7 clusters) while the y-axis shows the score of the validation test.

Appendix J

Questionnaire Survey Results

Q1 What is your school of study?

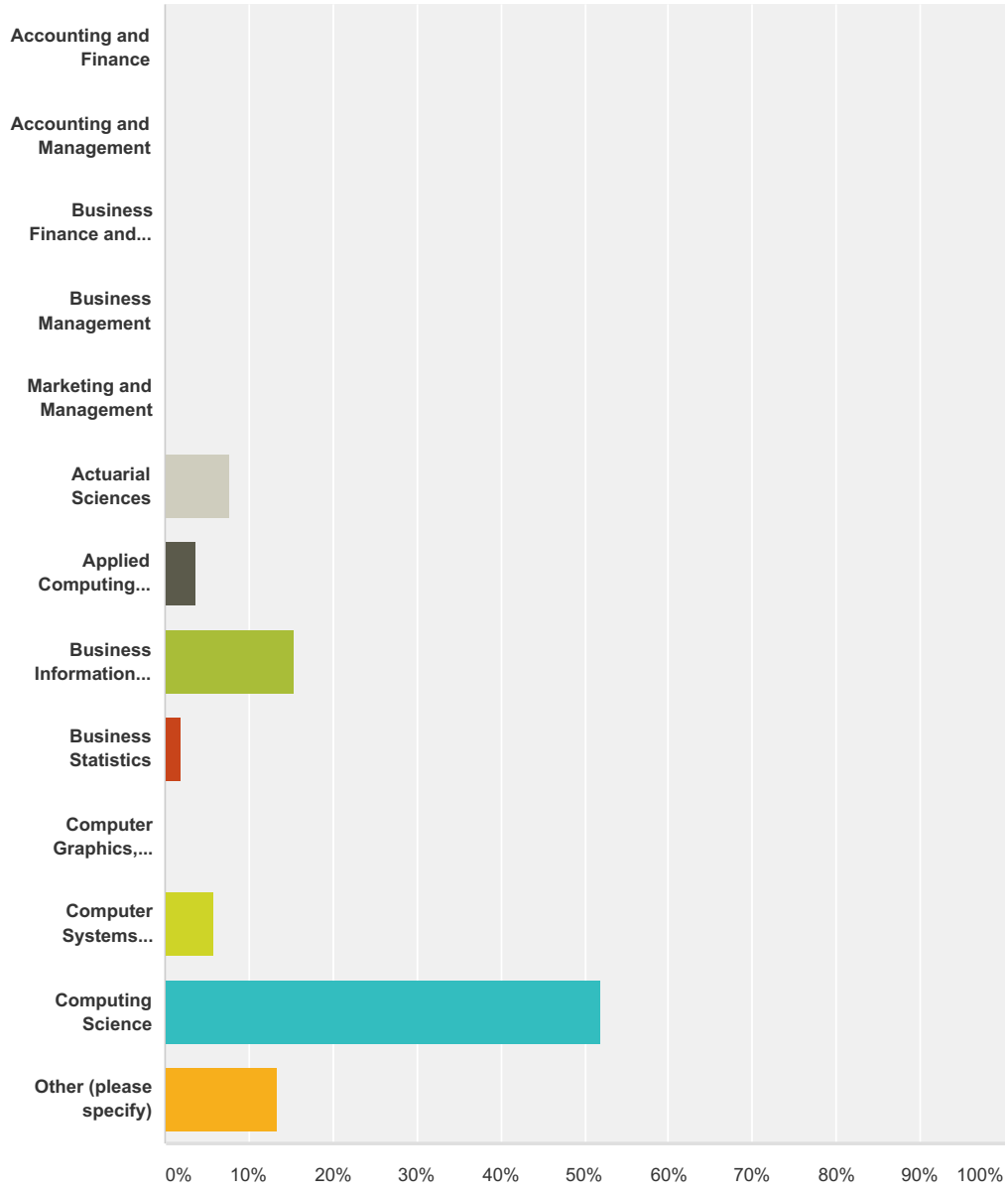
Answered: 81 Skipped: 0



Answer Choices	Responses	
Norwich Business School.	33.33%	27
School of Computing Sciences.	66.67%	54
Total		81

Q2 What is your course of study?

Answered: 52 Skipped: 29

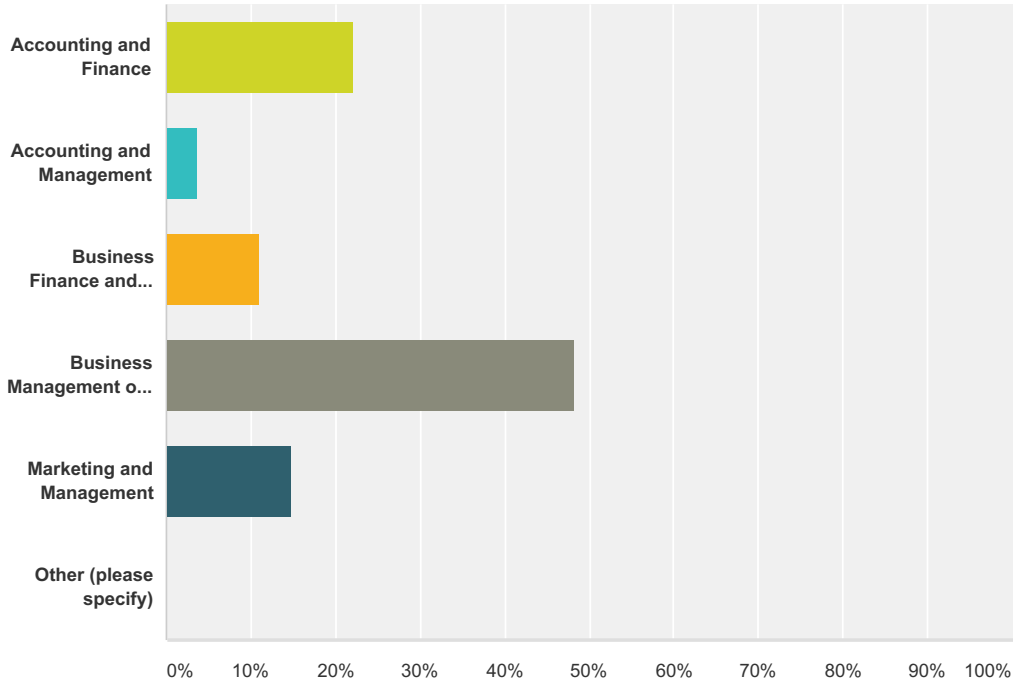


Answer Choices	Responses	
Accounting and Finance	0.00%	0
Accounting and Management	0.00%	0
Business Finance and Management	0.00%	0
Business Management	0.00%	0
Marketing and Management	0.00%	0
Actuarial Sciences	7.69%	4

Applied Computing Science	**3.85%**	2
Business Information Systems	**15.38%**	8
Business Statistics	**1.92%**	1
Computer Graphics, Imaging and Multimedia	**0.00%**	0
Computer Systems Engineering	**5.77%**	3
Computing Science	**51.92%**	27
Other (please specify)	**13.46%**	7
Total		**52**

Q3 What is your course of study?

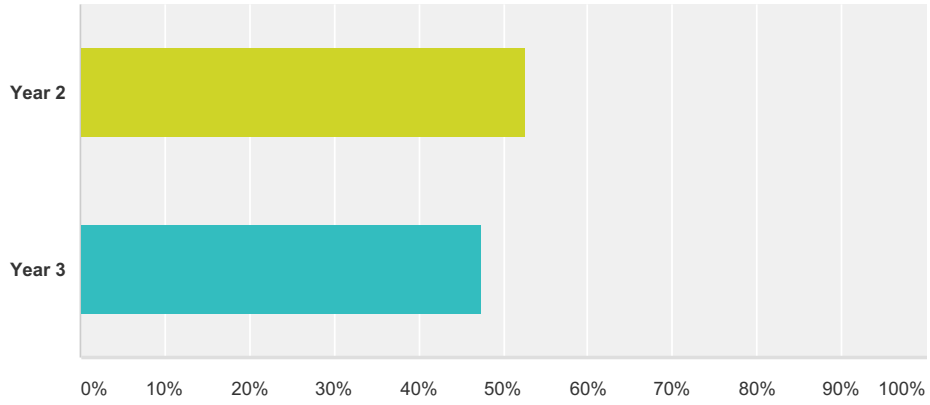
Answered: 27 Skipped: 54



Answer Choices	Responses	
Accounting and Finance	22.22%	6
Accounting and Management	3.70%	1
Business Finance and Management	11.11%	3
Business Management or Management	48.15%	13
Marketing and Management	14.81%	4
Other (please specify)	0.00%	0
Total		27

Q4 What is your year of study?

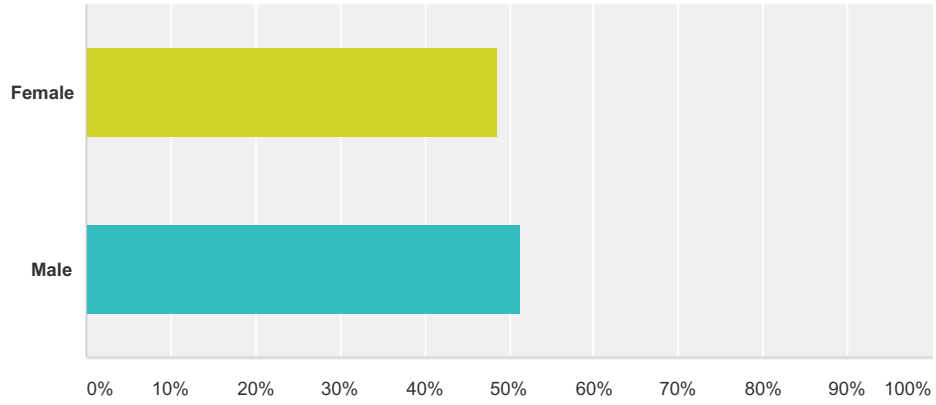
Answered: 76 Skipped: 5



Answer Choices	Responses	
Year 2	**52.63%**	40
Year 3	**47.37%**	36
Total		**76**

Q5 What is your gender?

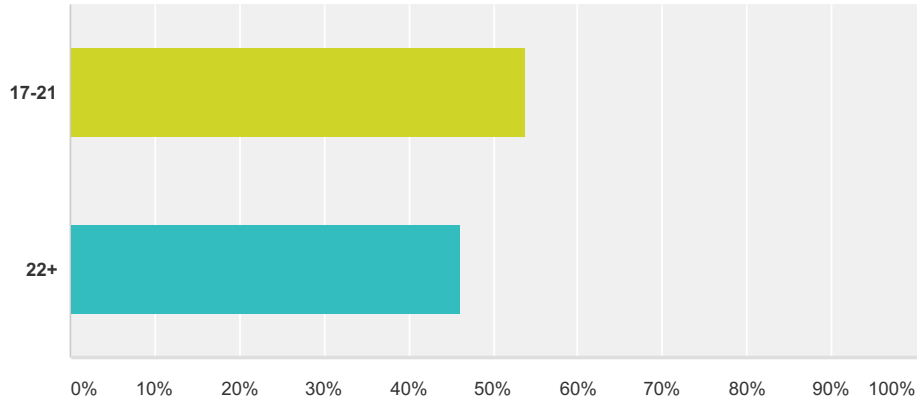
Answered: 76 Skipped: 5



Answer Choices	Responses	
Female	**48.68%**	37
Male	**51.32%**	39
Total		**76**

Q6 What is your age?

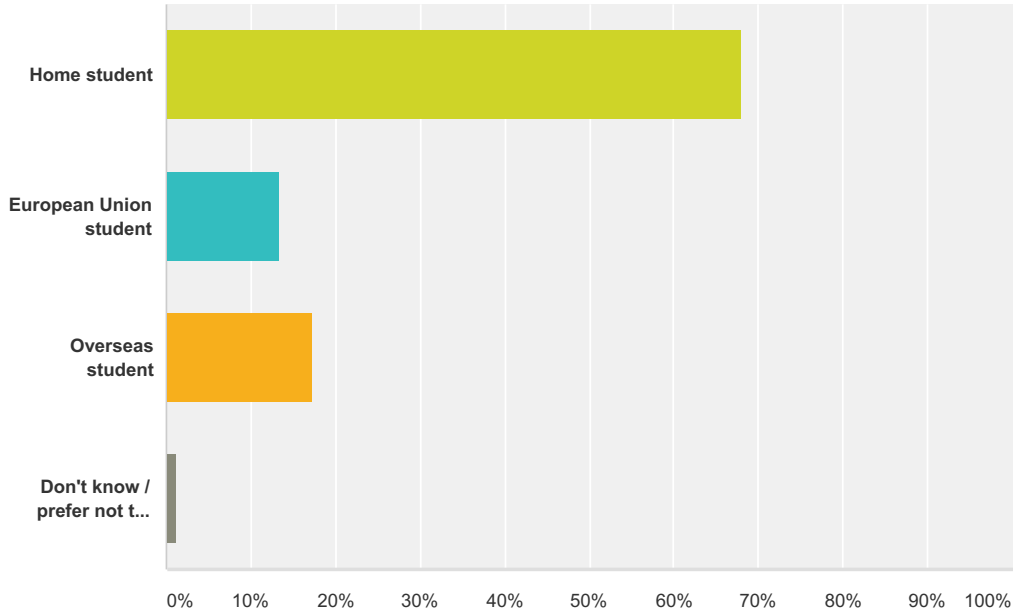
Answered: 76 Skipped: 5



Answer Choices	Responses	
17-21	**53.95%**	41
22+	**46.05%**	35
Total		**76**

Q7 What is your fee status?

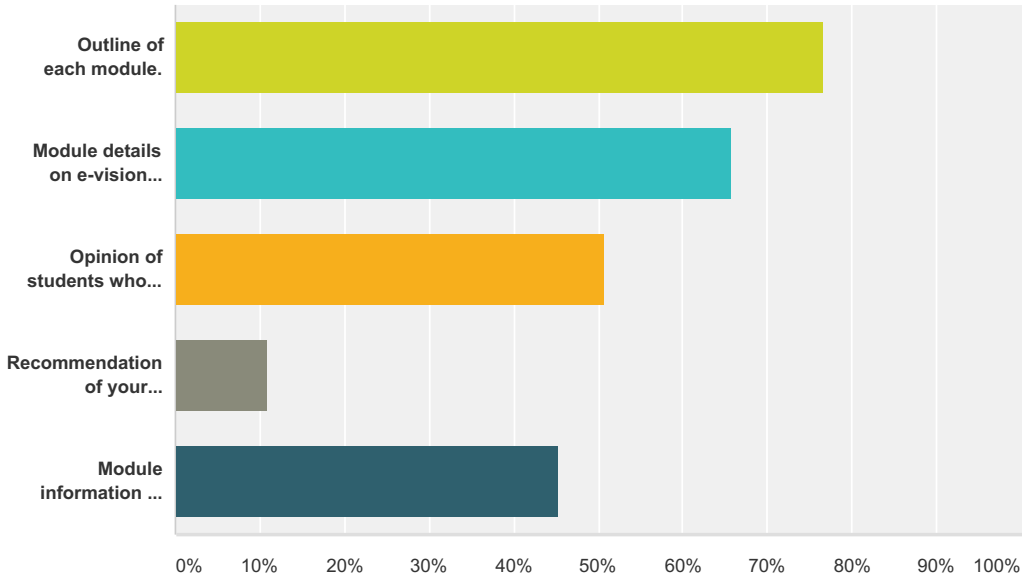
Answered: 75 Skipped: 6



Answer Choices	Responses	
Home student	**68.00%**	51
European Union student	**13.33%**	10
Overseas student	**17.33%**	13
Don't know / prefer not to say	**1.33%**	1
Total		**75**

Q8 What source/s of information did you consider when you chose your optional modules? (Check all that apply)

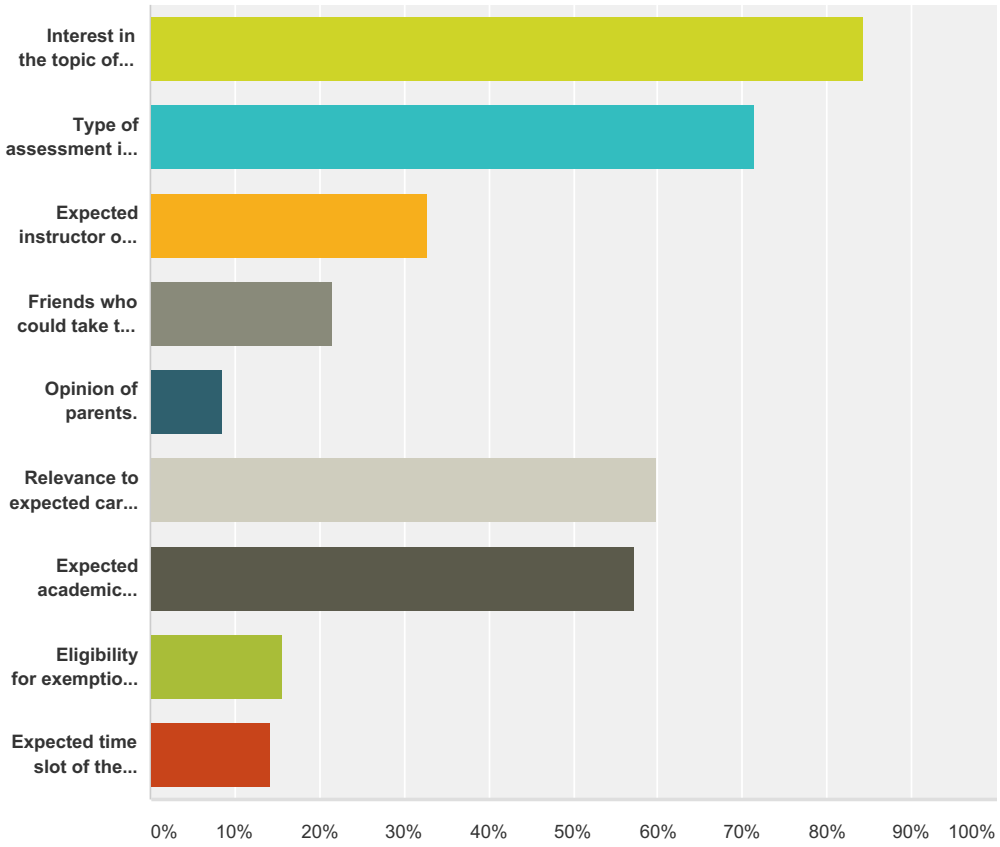
Answered: 73 Skipped: 8



Answer Choices	Responses	
Outline of each module.	**76.71%**	56
Module details on e-vision catalogue (https://evision.uea.ac.uk).	**65.75%**	48
Opinion of students who have experienced the module.	**50.68%**	37
Recommendation of your academic adviser.	**10.96%**	8
Module information day /fair.	**45.21%**	33
Total Respondents: 73		

Q9 What criteria did you consider when you chose your optional module(s)? (Check all that apply)

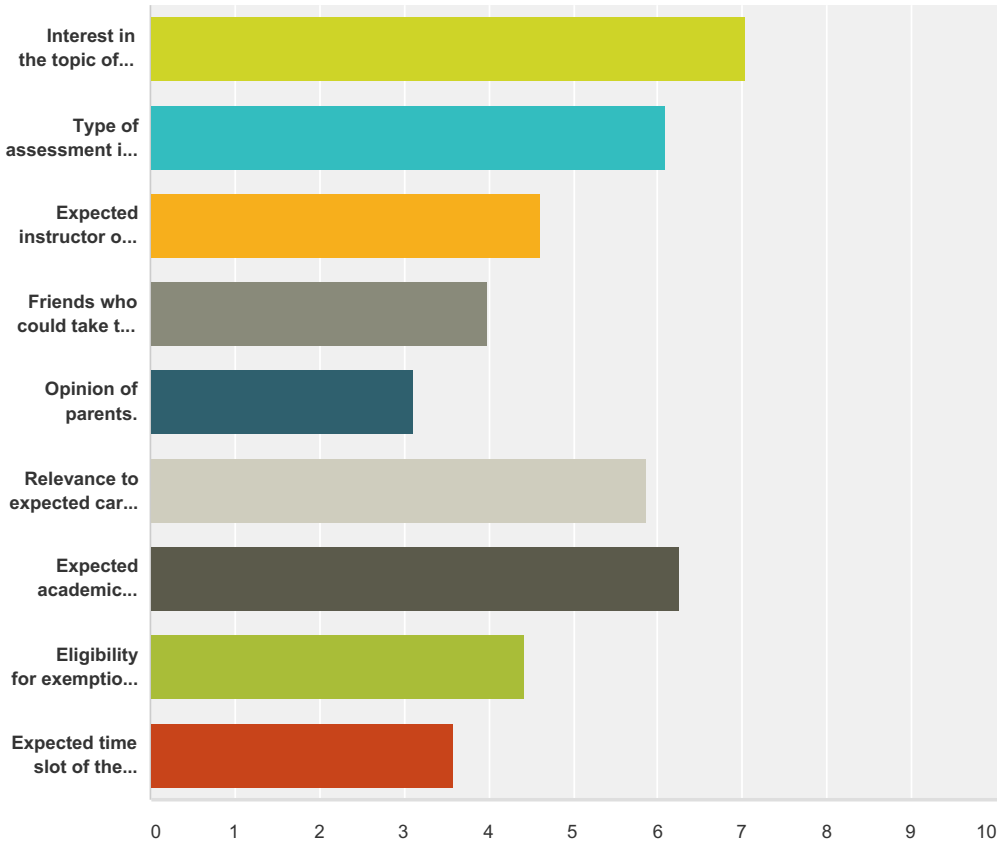
Answered: 70 Skipped: 11



Answer Choices	Responses	
Interest in the topic of the module.	**84.29%**	59
Type of assessment in each module.	**71.43%**	50
Expected instructor of the module.	**32.86%**	23
Friends who could take the same module.	**21.43%**	15
Opinion of parents.	**8.57%**	6
Relevance to expected career options.	**60.00%**	42
Expected academic performance.	**57.14%**	40
Eligibility for exemption from professional exams.	**15.71%**	11
Expected time slot of the lecture.	**14.29%**	10
Total Respondents: 70		

Q10 Please rank the following criteria based on your priority when choosing your optional module(s)? (1 being the highest priority and 9 being the lowest priority)Please note each rank number is allow to be chosen once.

Answered: 61 Skipped: 20



	1	2	3	4	5	6	7	8	9	Total	Score
Interest in the topic of the module.	**45.90%** 28	**13.11%** 8	**6.56%** 4	**9.84%** 6	**6.56%** 4	**9.84%** 6	**0.00%** 0	**0.00%** 0	**8.20%** 5	61	7.03
Type of assessment in each module.	**4.92%** 3	**29.51%** 18	**18.03%** 11	**13.11%** 8	**9.84%** 6	**11.48%** 7	**4.92%** 3	**6.56%** 4	**1.64%** 1	61	6.10
Expected instructor of the module.	**4.92%** 3	**9.84%** 6	**4.92%** 3	**14.75%** 9	**14.75%** 9	**14.75%** 9	**16.39%** 10	**14.75%** 9	**4.92%** 3	61	4.62
Friends who could take the same module.	**4.92%** 3	**3.28%** 2	**4.92%** 3	**16.39%** 10	**9.84%** 6	**11.48%** 7	**14.75%** 9	**21.31%** 13	**13.11%** 8	61	3.98
Opinion of parents.	**4.92%** 3	**6.56%** 4	**3.28%** 2	**6.56%** 4	**6.56%** 4	**1.64%** 1	**16.39%** 10	**9.84%** 6	**44.26%** 27	61	3.11
Relevance to expected career options.	**11.48%** 7	**16.39%** 10	**18.03%** 11	**16.39%** 10	**6.56%** 4	**16.39%** 10	**3.28%** 2	**8.20%** 5	**3.28%** 2	61	5.87

Expected academic performance.	**18.03%** 11	**14.75%** 9	**21.31%** 13	**9.84%** 6	**13.11%** 8	**11.48%** 7	**4.92%** 3	**4.92%** 3	**1.64%** 1	61	6.26
Eligibility for exemption from professional exams.	**4.92%** 3	**3.28%** 2	**18.03%** 11	**4.92%** 3	**19.67%** 12	**9.84%** 6	**13.11%** 8	**13.11%** 8	**13.11%** 8	61	4.43
Expected time slot of the lecture.	**0.00%** 0	**3.28%** 2	**4.92%** 3	**8.20%** 5	**13.11%** 8	**13.11%** 8	**26.23%** 16	**21.31%** 13	**9.84%** 6	61	3.59

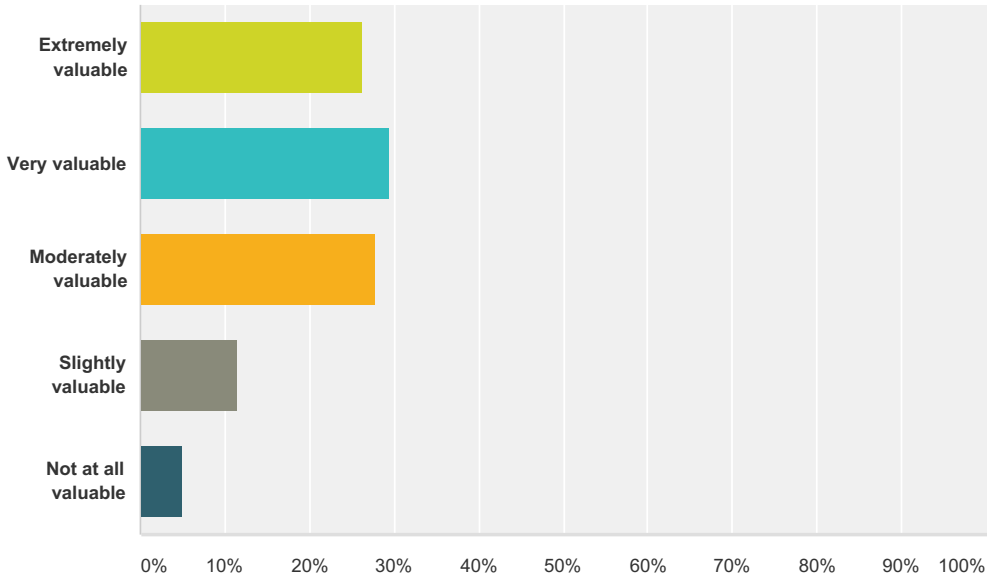
Q11 What optional module(s) have you chosen? (please enter at least 1 module)

Answered: 61 Skipped: 20

Answer Choices	Responses	
Module 1	100.00%	61
Module 2 (optional)	73.77%	45
Module 3 (optional)	44.26%	27
Module 4 (optional)	29.51%	18
Module 5 (optional)	16.39%	10
Module 6 (optional)	11.48%	7

Q12 Recently, universities have started to use advanced data mining techniques to provide personalised predictions of module performance. For example, in the process of your module choice, the university enrolment system could show you your predicted mark based on previous students with similar personal characteristics.How much would you value this personalised prediction of module performance in making your module choice(s)?

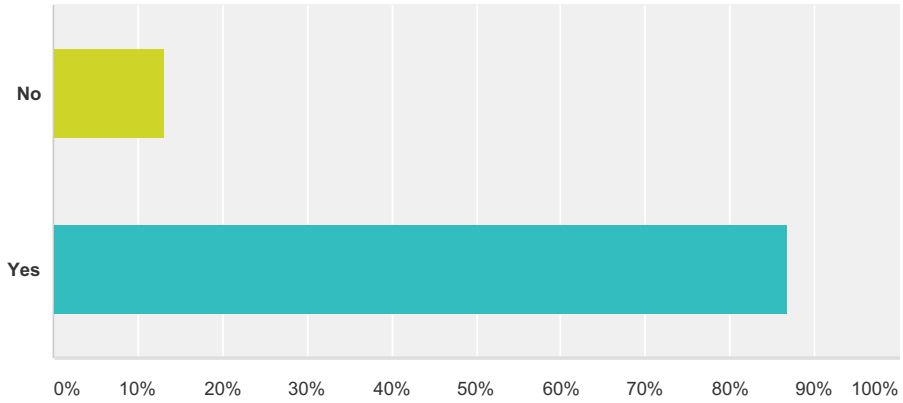
Answered: 61 Skipped: 20



Answer Choices	Responses	
Extremely valuable	26.23%	16
Very valuable	29.51%	18
Moderately valuable	27.87%	17
Slightly valuable	11.48%	7
Not at all valuable	4.92%	3
Total		61

Q13 Would you be interested to know your personal predicted marks for the modules you are currently studying?

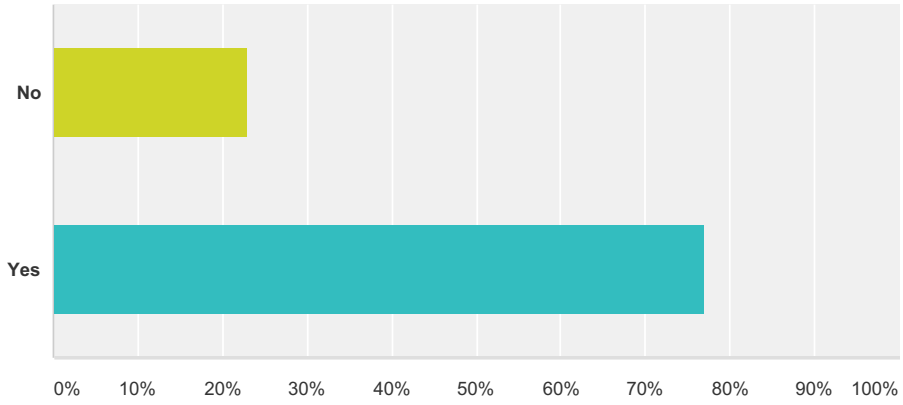
Answered: 61 Skipped: 20



Answer Choices	Responses	
No	**13.11%**	8
Yes	**86.89%**	53
Total		**61**

Q14 Do you think that knowing your personal predicted marks would have affected your decisions in choosing your optional modules?

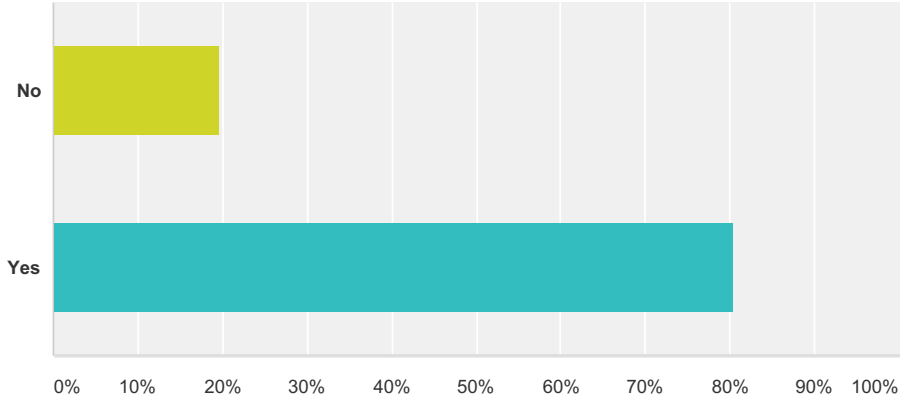
Answered: 61 Skipped: 20



Answer Choices	Responses	
No	**22.95%**	14
Yes	**77.05%**	47
Total		**61**

Q15 Currently, universities can provide personal enrolment recommender programmes that can suggest optional modules based on a student's expected performance, expected student satisfaction and his /her career choices. The recommendation would be personalised, that is based on previous students with similar personal characteristics. Would you have been interested to have had such a broadly-based programme during your module enrolment process?

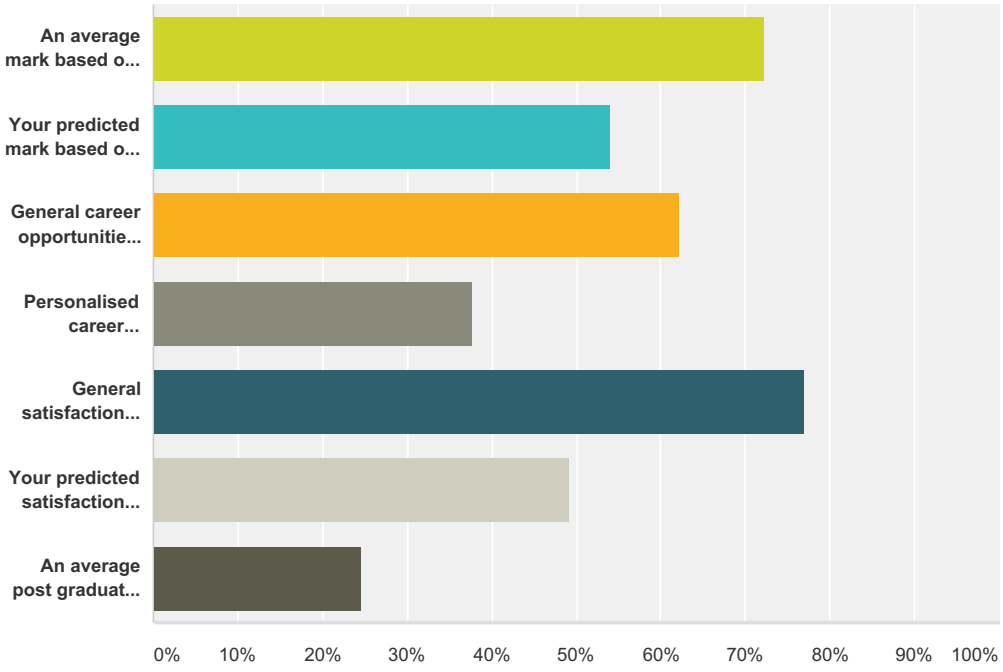
Answered: 61 Skipped: 20



Answer Choices	Responses	
No	19.67%	12
Yes	80.33%	49
Total		61

Q16 In your opinion, what additional information would be helpful in choosing your optional module(s)? (Check all that apply)

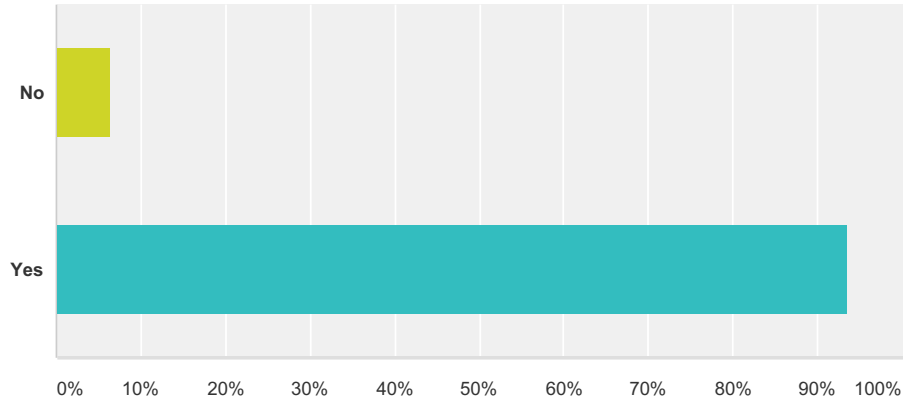
Answered: 61 Skipped: 20



Answer Choices	Responses	
An average mark based on the past few years of students marks.	**72.13%**	44
Your predicted mark based on previous students with similar personal characteristics.	**54.10%**	33
General career opportunities associated with the module.	**62.30%**	38
Personalised career opportunities based on previous students with similar personal characteristics who took the module.	**37.70%**	23
General satisfaction rate of students who took the same module in the past few years.	**77.05%**	47
Your predicted satisfaction rate based on students with similar personal characteristics.	**49.18%**	30
An average post graduation salary of students who took the module.	**24.59%**	15
Total Respondents: 61		

Q17 In your opinion, would it be helpful to know the previous students' evaluation of the module instructor during your module enrolment process?

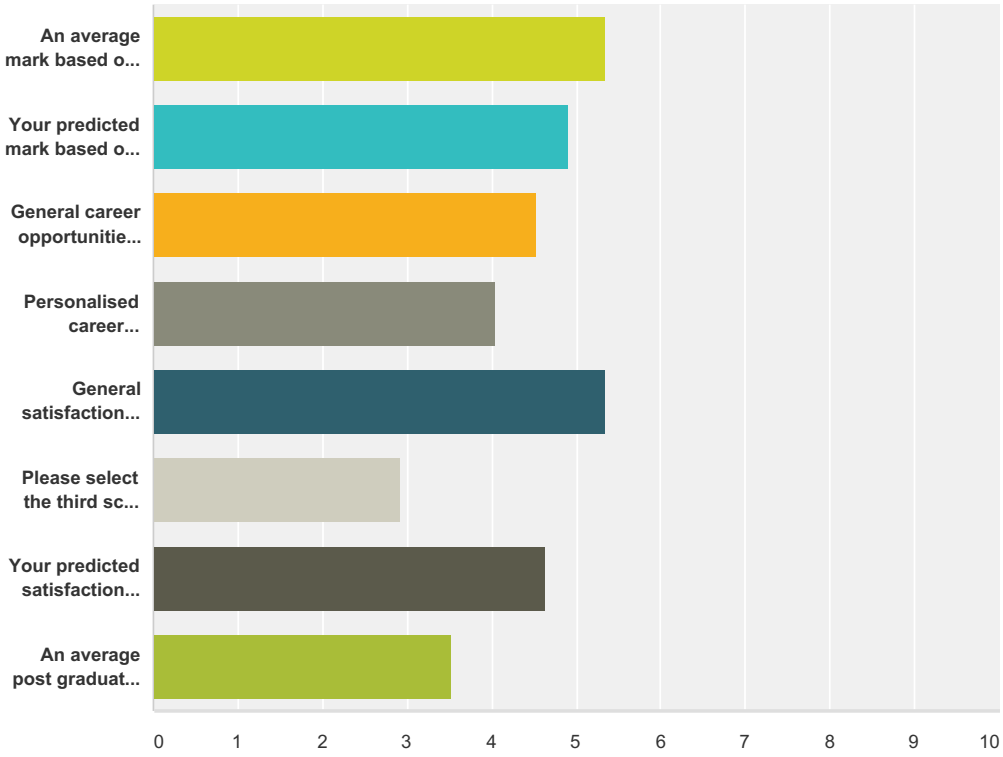
Answered: 61 Skipped: 20



Answer Choices	Responses	
No	**6.56%**	4
Yes	**93.44%**	57
Total		**61**

Q18 On a scale from 1 (being the lowest) to 7 (being the highest), how useful would the following information be in making your module choices?

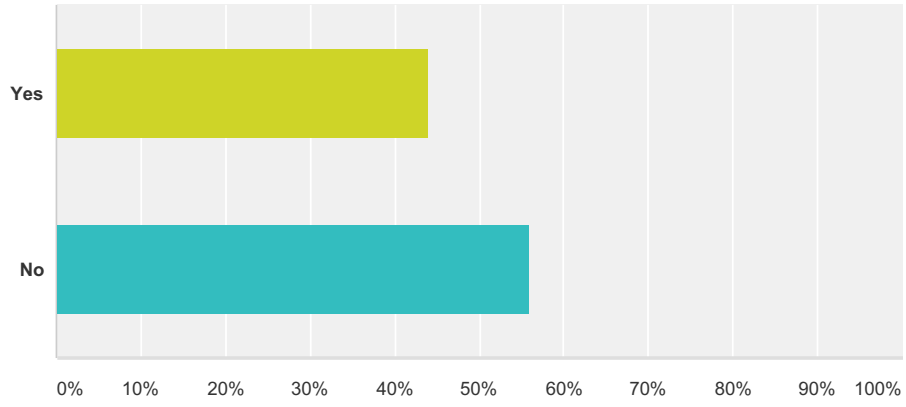
Answered: 59 Skipped: 22



	1 (lowest)	2	3	4	5	6	7 (highest)	Total	Weighted Average
An average mark based on the past few years of students marks.	3.39% 2	3.39% 2	6.78% 4	13.56% 8	16.95% 10	27.12% 16	28.81% 17	59	5.34
Your predicted mark based on previous students with similar personal characteristics.	3.39% 2	10.17% 6	10.17% 6	13.56% 8	15.25% 9	27.12% 16	20.34% 12	59	4.90
General career opportunities associated with the module.	1.69% 1	10.17% 6	8.47% 5	23.73% 14	28.81% 17	22.03% 13	5.08% 3	59	4.54
Personalised career opportunities based on previous students with similar personal characteristics who took the module.	1.69% 1	15.25% 9	15.25% 9	28.81% 17	25.42% 15	10.17% 6	3.39% 2	59	4.05
General satisfaction rate of students who took the same module in the past few years.	0.00% 0	3.39% 2	10.17% 6	10.17% 6	25.42% 15	27.12% 16	23.73% 14	59	5.34
Please select the third scale (circle).	5.08% 3	5.08% 3	83.05% 49	5.08% 3	1.69% 1	0.00% 0	0.00% 0	59	2.93
Your predicted satisfaction rate based on students with similar personal characteristics.	3.39% 2	5.08% 3	20.34% 12	10.17% 6	25.42% 15	27.12% 16	8.47% 5	59	4.64
An average post graduation salary of students who took the module.	10.17% 6	18.64% 11	20.34% 12	18.64% 11	25.42% 15	5.08% 3	1.69% 1	59	3.53

Q19 Would you be willing to be interviewed for 20 minutes to discuss the process of optional module enrolment and to find out what your predicted marks are?(1- You will be asked to give your informed consent before the interview. 2- We will not require your name and all your recorded data will be anonymous. 3- A £10 Amazon voucher will be emailed to the selected students after they complete their interview.)

Answered: 59 Skipped: 22



Answer Choices	Responses	
Yes	44.07%	26
No	55.93%	33
Total		59

Q20 If yes, at what email address would you like to be contacted? (optional)

Answered: 26 Skipped: 55

Q21 Do you have any comments, other questions you think we should add, or concerns?

Answered: 4 Skipped: 77

Bibliography

[1] Zahyah Alharbi, James Cornford, Liam Dolder, and Beatriz De La Iglesia. Using data mining techniques to predict students at risk of poor performance. In *SAI Computing Conference (SAI), 2016*, pages 523–531. IEEE, 2016.

[2] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[3] IBM. Big data and analytics. the four V's of big data. `http://www-01.ibm.com/software/data/bigdata/`, September 2015.

[4] David Kiron, Rebecca Shockley, Nina Kruschwitz, Glenn Finch, and Michael Haydock. Analytics: The widening divide. *MIT Sloan Management Review*, 53(3):1–22, 2011.

[5] Margaret H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2002.

[6] Alex Berson, Stephen Smith, and Kurt Thearling. An overview of data mining techniques. In *Building Data Mining Application for CRM*. McGraw-Hill Professional, 2004.

[7] Miguel APM Lejeune. Measuring the impact of data mining on churn management. *Internet Research*, 11(5):375–387, 2001.

[8] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[9] Barry Leventhal. An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice*, 12(2):137–153, 2010.

[10] Jing Luan. Data mining and knowledge management in higher education-potential applications. In *Workshop associate of institutional research international conference*, pages 1–18, Toronto,Canada, 2002.

[11] César Vialardi, Jorge Chue, Juan Pablo Peche, Gustavo Alvarado, Bruno Vinatea, Jhonny Estrella, and Álvaro Ortigosa. A data mining approach to guide students through the enrollment process based on academic performance. *User modeling and user-adapted interaction*, 21(1-2):217–248, 2011.

[12] Ya-Huei Wang and Hung-Chang Liao. Data mining for adaptive learning in a tesl-based e-learning system. *Expert Systems with Applications*, 38(6):6480–6485, 2011.

[13] Jiawei Han and Micheline Kamber. *Data mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[14] Jiawei Han. How can data mining help bio-data analysis? In *Proceedings of the 2nd International Conference on Data Mining in Bioinformatics*, pages 1–2. Springer-Verlag, 2002.

[15] Boris Kovalerchuk and Evgenii Vityaev. Data mining for financial applications. In *Data Mining and Knowledge Discovery Handbook*, pages 1153–1169. Springer, Boston, MA, 2009.

[16] Zhan-fan Liang and Lang Liu. The application of data mining techniques to programme decision making and information management. In *International Conference on Networks*

Security, Wireless Communications and Trusted Computing, 2009. NSWCTC'09., volume 2, pages 780–786. IEEE, 2009.

[17] Herb Edelstein. Building profitable customer relationships with data mining. In *Customer Relationship Management*, pages 339–351. Vieweg+Teubner Verlag, Wiesbaden, 2000.

[18] Bamshad Mobasher, Namit Jain, Eui-Hong Han, and Jaideep Srivastava. Web mining: Pattern discovery from world wide web transactions. Technical report, Technical Report TR96-050, Department of Computer Science, University of Minnesota, 1996.

[19] Jayanthi Ranjan and Kamna Malik. Effective educational process: a data-mining approach. *Vine*, 37(4):502–515, 2007.

[20] Ryan Shaun Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17, 2009.

[21] Richard A Huebner. A survey of educational data-mining research. *Research in Higher Education Journal*, 19, 2013.

[22] Hamid R Nemati and Christopher D Barko. Organizational data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 1057–1067. Springer, Boston, MA, 2010.

[23] Toon Calders and Mykola Pechenizkiy. Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter*, 13(2):3–6, 2012.

[24] Albert Corbett. Cognitive computer tutors: Solving the two-sigma problem. In *International Conference on User Modeling*, pages 137–147. Springer, Berlin, Heidelberg, 2001.

[25] Oliver Scheuer and Bruce M McLaren. Educational data mining. In *Encyclopedia of the Sciences of Learning*, pages 1075–1079. Springer US, 2012.

[26] Jeff Guan, William Nunez, and John F Welsh. Institutional strategy and information support: the role of data warehousing in higher education. *Campus-wide Information Systems*, 19(5):168–174, 2002.

[27] Osmar R Zaíane. Building a recommender agent for e-learning systems. In *Computers in Education, 2002. Proceedings. International Conference on*, pages 55–59. IEEE, 2002.

[28] Jui-Long Hung, Morgan C Wang, Shuyan Wang, Maha Abdelrasoul, Yaohang Li, and Wu He. Identifying at-risk students for early interventionsa time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, 5(1):45–55, 2017.

[29] Natthakan Iam-On and Tossapon Boongoen. Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 8(2):497–510, Apr 2017.

[30] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[31] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[32] Olga C Santos and Jesus G Boticario. Modeling recommendations for the educational domain. *Procedia Computer Science*, 1(2):2793–2800, 2010.

[33] Aleksandra Klašnja-Milićević, Mirjana Ivanović, and Alexandros Nanopoulos. Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, 44(4):571–604, 2015.

[34] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[35] Marija Blagojević and Živadin Micić. A web-based intelligent report e-learning system using data mining techniques. *Computers & Electrical Engineering*, 39(2):465–474, 2013.

[36] Cesar Vialardi, Javier Bravo, Leila Shafti, and Alvaro Ortigosa. Recommendation in higher education using data mining techniques. *Proceedings of the 2nd International Conference on Educational Data Mining,EDM 2009.*, pages 190–199, 2009.

[37] Nabila Bousbia and Idriss Belamri. Which contribution does edm provide to computer-based learning environments? In *Educational Data Mining*, pages 3–28. Springer, 2014.

[38] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, 97:320–324, 2013.

[39] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

[40] Cristobal Romero and Sebastian Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[41] Félix Castro, Alfredo Vellido, Àngela Nebot, and Francisco Mugica. Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pages 183–221. Springer, 2007.

[42] Ryan Shaun Baker. Data mining for education. *International encyclopedia of education*, 7(3):112–118, 2010.

[43] Philip J Piety. *Assessing the educational data movement*. Teachers College Press, 2013.

[44] Eduventures. Predictive analytics in higher education data-driven decision-making for the student life cycle. *Eduventures*, pages 1–13, 2013.

[45] SoLAR. Society for Learning Analytics research. `http://www.solaresearch.org/about/`, July 2015.

[46] George Siemens and Ryan Shaun Baker. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254. ACM, 2012.

[47] Marie Bienkowski, Mingyu Feng, and Barbara Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, pages 1–57, 2012.

[48] Keith E Maull, Manuel Gerardo Saldivar, and Tamara Sumner. Online curriculum planning behavior of teachers. In *EDM*, pages 121–130. ERIC, 2010.

[49] Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan SJd Baker. *Handbook of educational data mining*. CRC Press, 2010.

[50] Behrouz Minaei-Bidgoli, Deborah A Kashy, Gerd Kortemeyer, and William Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in education, 2003. FIE 2003 33rd annual*, volume 1, pages T2A–13. IEEE, 2003.

[51] Suhem Parack, Zain Zahid, and Fatima Merchant. Application of data mining in educational databases for predicting academic trends and patterns. In *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*, pages 1–4. IEEE, 2012.

[52] Alfredo Vellido, Félix Castro, and A Nebot. Clustering educational data. *In Romero C, Ventura S, Pechenizkiy M, Baker Ryan Shaun, eds. Handbook of educational data mining*, pages 75–92, 2011.

[53] Agathe Merceron and Kalina Yacef. Measuring correlation of strong symmetric association rules in educational data. *In Romero C, Ventura S, Pechenizkiy M, Baker Ryan Shaun, eds. Handbook of educational data mining*, pages 245–256, 2011.

[54] Riccardo Mazza and Christian Milani. Gismo: a graphical interactive student monitoring

tool for course management systems. In *International Conference on Technology Enhanced Learning, Milan*, pages 1–8, 2004.

[55] Maomi Ueno. Online outlier detection system for learning time data in e-learning and its evaluation. In *CATE*, pages 248–253, 2004.

[56] Reihaneh Rabbany, Mansoureh Takaffoli, and Osmar R Zaïane. Analyzing participation of students in online courses using social network analysis techniques. In *Proceedings of educational data mining*. Citeseer, 2011.

[57] Nikola Trcka, Mykola Pechenizkiy, and Wil van der Aalst. Process mining from educational data. *In Romero C, Ventura S, Pechenizkiy M, Baker Ryan Shaun, eds. Handbook of educational data mining*, pages 123–142, 2011.

[58] Wu He. Examining students online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1):90–102, 2013.

[59] Albert T Corbett and John R Anderson. Knowledge tracing - modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1995.

[60] Jenifer L Wheeler and J Wesley Regian. The use of a cognitive tutoring system in the improvement of the abstract reasoning component of word problem solving. *Computers in Human Behavior*, 15(2):243–254, 1999.

[61] Louise Corti, Annette Day, and Gill Backhouse. Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 1(3), 2000.

[62] Cristóbal Romero, José Raúl Romero, and Sebastián Ventura. A survey on pre-processing educational data. In *Educational Data Mining*, pages 29–64. Springer, 2014.

[63] Cristóbal Romero, José Raúl Romero, Jose María Luna, and Sebastián Ventura. Mining rare association rules from e-learning data. In *Educational Data Mining 2010*, 2010.

[64] R Barahate Sachin and M Shelake Vijay. A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100. IEEE, 2012.

[65] Roderick JA Little and D Rubin. *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1987.

[66] André Krüger, Agathe Merceron, and Benjamin Wolf. A data model to ease analysis and mining of educational data. In *Educational Data Mining 2010*, 2010.

[67] Richard A Huebner. Mining for knowledge in higher education. *ASBBSeJournal*, page 56, 2008.

[68] Marvin L Brown and John F Kros. Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8):611–621, 2003.

[69] J Ben Schafer. The application of data-mining to recommender systems. *Encyclopedia of data warehousing and mining*, 1:44–48, 2009.

[70] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[71] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[72] Hendrik Drachsler, Hans Hummel, and Rob Koper. Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning. In *Workshop on Social Information Retrieval for Technology-Enhanced Learning (SIRTEL07)*

at the EC-TEL conference, pages 18–26, Crete, Greece, 2007.

[73] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, and Lars Schmidt-Thieme. Recommender system for predicting student performance. *Procedia Computer Science*, 1(2):2811–2819, 2010.

[74] Zacharoula Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[75] CMA. Consumer law compliance review: Higher education undergraduate sector findings report. *Competition and Markets Authority*, pages 1–47, 2016.

[76] Arthur L Stinchcombe. *Information and organizations*. University of California Press, Berkeley CA, 1990.

[77] Cleopatra Veloutsou, John W Lewis, and Robert A Paton. University selection: information requirements and importance. *International Journal of Educational Management*, 18(3):160–171, 2004.

[78] Cleopatra Veloutsou, Robert A Paton, and John Lewis. Consultation and reliability of information sources pertaining to university selection: some questions answered? *International Journal of Educational Management*, 19(4):279–291, 2005.

[79] Tristram Hooley, Robin Mellors-Bourne, and Moira Sutton. Early evaluation of unistats: user experiences. Technical report, UK Higher Education Funding Bodies, 2013.

[80] Liz Browne and Steve Rayner. Managing leadership in university reform: data-led decision-making, the cost of learning and déjà vu? *Educational Management Administration & Leadership*, 43(2):290–307, 2015.

[81] Jane Hemsley-Brown and Izhar Oplatka. University choice: what do we know, what dont we know and what do we still need to find out? *International Journal of Educational Management*, 29(3):254–274, 2015.

[82] Simon Marginson. University rankings and social science. *European Journal of Education*, 49(1):45–59, 2014.

[83] Diane Reay, Jacqueline Davies, Miriam David, and Stephen J Ball. Choices of degree or degrees of choice? class,raceand the higher education choice process. *Sociology*, 35(4):855–874, 2001.

[84] Kim Slack, Jean Mangan, Amanda Hughes, and Peter Davies. hot,coldand warminformation and higher education decision-making. *British Journal of Sociology of Education*, 35(2):204–223, 2014.

[85] Mary R Hedges, Gail A Pacheco, and Don J Webber. What determines students choices of elective modules? *International Review of Economics Education*, 17:39–54, 2014.

[86] Louis Soares. The personalization of higher education: Using technology to enhance the college experience. *Washington, DC: Center for American Progress*, 2011.

[87] Yingxu Wang, Ying Wang, S. Patel, and D. Patel. A layered reference model of the brain (lrmb). *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(2):124–133, March 2006.

[88] Robert Andrew Wilson and Frank C Keil. *The MIT encyclopedia of the cognitive sciences*. MIT press, 2001.

[89] Yingxu Wang. The theoretical framework of cognitive informatics. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 1(1):1–27, 2007.

[90] Yingxu Wang and Guenther Ruhe. The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence*, 1(2):73–85, 2007.

[91] Richard H Thaler, Cass R Sunstein, and John P Balz. Choice architecture (december 10, 2014). *The Behavioral Foundations of Public Policy, Ch. 25, Eldar Shafir, ed. (2012). Available at SSRN: https://ssrn.com/abstract=2536504 or http://dx.doi.org/10.2139/ssrn.2536504*, 2014.

[92] Thomas C. Leonard. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy*, 19(4):356–360, Dec 2008.

[93] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, USA, 2011.

[94] Christopher P Chambers and Takashi Hayashi. Choice and individual welfare. *Journal of Economic Theory*, 147(5):1818–1849, 2012.

[95] Gary Browning, Abigail Halcli, and Frank Webster. *Understanding contemporary society: Theories of the present*. Sage, 1999.

[96] Anthony Heath. *Rational choice and social exchange: A critique of exchange theory*. CUP Archive, 1976.

[97] Michael J Shapiro, G Matthew Bonham, and Daniel Heradstveit. A discursive practices approach to collective decision-making. *International Studies Quarterly*, 32(4):397–419, 1988.

[98] Andreas Reckwitz. Toward a theory of social practices: A development in culturalist theorizing. *European journal of social theory*, 5(2):243–263, 2002.

[99] Luigina Mortari and Roberta Silva. Analyzing how discursive practices affect physicians decision-making processes: A phenomenological-based qualitative study in critical care contexts. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 54:0046958017731962, 2017.

[100] Manoj Bala and DB Ojha. Study of applications of data mining techniques in education. *International Journal of Research in Science and Technology,(IJRST)*, pages 2249–0604, 2012.

[101] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: a case study. In *Educational Data Mining 2009*, 2009.

[102] Serge Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131):17–33, 2006.

[103] Juan-Francisco Superby, JP Vandamme, and N Meskens. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining*, pages 37–44, 2006.

[104] Kash Barker, Theodore Trafalis, and Teri Reed Rhoads. Learning from student data. In *Systems and Information Engineering Design Symposium, 2004. Proceedings of the 2004 IEEE*, pages 79–86. IEEE, 2004.

[105] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. A comparative analysis of techniques for predicting academic performance. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages T2G–7. IEEE, 2007.

[106] Yiming Ma, Bing Liu, Ching Kian Wong, Philip S Yu, and Shuik Ming Lee. Targeting the right students using data mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–464. ACM, 2000.

[107] N. M. Norwawi, S. F. Abdusalam, C. F. Hibadullah, and B. M. Shuaibu. Classification of students' performance in computer programming course according to learning style. In *2009 2nd Conference on Data Mining and Optimization*, pages 37–41. IEEE, Oct 2009.

[108] Dorina Kabakchieva. Predicting student performance by using data mining methods for

classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

[109] Subitha Sivakumar and Rajalakshmi Selvaraj. Predictive modeling of students performance through the enhanced decision tree. In Akhtar Kalam, Swagatam Das, and Kalpana Sharma, editors, *Advances in Electronics, Communication and Computing: ETAEERE-2016*, pages 21–36. Springer Singapore, Singapore, 2018.

[110] S. Kotsiantis, K. Patriarcheas, and M. Xenos. A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education. *Knowledge-Based Systems*, 23(6):529–535, 2010.

[111] Zachary A Pardos, Sujith M. Gowda, Ryan S.J.d. Baker, and Neil T. Heffernan. The sum is greater than the parts: Ensembling models of student knowledge in educational software. *SIGKDD Explor. Newsl.*, 13(2):37–44, May 2012.

[112] Ryan S. J. d. Baker, Zachary A. Pardos, Sujith M. Gowda, Bahador B. Nooraei, and Neil T. Heffernan. Ensembling predictions of student knowledge within intelligent tutoring systems. In Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver, editors, *User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings*, pages 13–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[113] Vasile Paul Bresfelean, Mihaela Bresfelean, Nicolae Ghisoiu, and Calin-Adrian Comes. Determining students academic failure profile founded on data mining methods. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on*, pages 317–322. IEEE, 2008.

[114] W Hämäläinen, J Suhonen, E Sutinen, and H Toivonen. Data mining in personalizing distance education courses. In *Proceedings of the 21st ICDE World Conference on Open Learning and Distance Education*, pages 18–21, 2004.

[115] W Hämäläinen and M Vinni. Comparison of machine learning methods for intelligent tutoring systems. In *Intelligent Tutoring Systems*, pages 525–534. Springer, 2006.

[116] Wei Zang and Fuzong Lin. Investigation of web-based teaching and learning by boosting algorithms. In *Information Technology: Research and Education, 2003. Proceedings. ITRE2003. International Conference on*, pages 445–449. IEEE, 2003.

[117] Sotiris B Kotsiantis, CJ Pierrakeas, and Panayiotis E Pintelas. Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 267–274. Springer, 2003.

[118] Martin Muehlenbrock. Automatic action analysis in an interactive learning environment. In *the Workshop on Usage Analysis in Learning Systems*, pages 73–80, Amsterdam, the Netherland, 2005. AIED.

[119] Cristóbal Romero, Sebastián Ventura, Pedro G Espejo, and César Hervás. Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining*, pages 8–17, Montral, Qubec, Canada, June 2008.

[120] Michel C Desmarais and Xiaoming Pu. A bayesian student model without hidden nodes and its comparison with item response theory. *International Journal of Artificial Intelligence in Education (IJAIED)*, 15:291–323, 2005.

[121] Anders Jonsson, Jeff Johns, Hasmik Mehranian, Ivon Arroyo, Beverly Woolf, Andrew Barto, Donald Fisher, and Sridhar Mahadevan. Evaluating the feasibility of learning student models from data. In *Educational Data Mining: Papers from the AAAI Workshop*, pages 1–6, 2005.

[122] Jiri Vomlel. Bayesian networks in educational testing. *In Proceedings of First European Workshop on Probabilistic Graphical Models (PGM02*, 1:83–100, 2002.

[123] Chen-Chung Liu. *Knowledge discovery from web portfolios: tools for learning performance*

assessment. PhD thesis, Department of Computer Scince Information Engineering Yuan Ze University, Taiwan, 2000.

[124] Mihaela Cocea and Stephan Weibelzahl. Can log files analysis estimate learners' level of motivation? In *LWA 2006: Lernen - Wissensentdeckung - Adaptivitt*. University of Hildesheim, Institute of Computer Science, 2006.

[125] Mihaela Cocea and Stephan Weibelzahl. Cross-system validation of engagement prediction from log files. In *Creating new learning experiences on a global scale*, pages 14–25. Springer, 2007.

[126] Marc Damez, Thanh Ha Dang, Christophe Marsala, and Bernadette Bouchon-Meunier. Fuzzy decision tree for user modeling from human-computer interactions. In *Proceedings of the 5th International Conference on Human System Learning, ICHSL*, volume 5, pages 287–302, 2005.

[127] Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*, pages 531–540. Springer, 2004.

[128] Myung-Geun Lee. Profiling students' adaptation styles in web-based learning. *Computers & Education*, 36(2):121–132, 2001.

[129] Teresa Hurley and Stephan Weibelzahl. Eliciting adaptation knowledge from on-line tutors to increase motivation. In *User Modeling 2007*, pages 370–374. Springer, 2007.

[130] Hana Bydovska and Lubomír Popelínskỳ. Predicting student performance in higher education. In *Proceedings of the 2013 24th International Workshop on Database and Expert Systems Applications*, DEXA '13, pages 141–145. IEEE Computer Society, 2013.

[131] Kimberly E Arnold. Signals: Applying academic analytics. *EDUCAUSE Quarterly*, 33(1):n1, 2010.

[132] Pedro Strecht, Luís Cruz, Carlos Soares, João Mendes-Moreira, and Rui Abreu. A comparative study of classification and regression algorithms for modelling students' academic performance. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015.*, pages 392–395, Madrid, Spain, 2015. International Educational Data Mining Society (IEDMS).

[133] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[134] Hao Cen, Kenneth Koedinger, and Brian Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent tutoring systems*, pages 164–175. Springer, 2006.

[135] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3):243–266, 2009.

[136] Michael P. O'Mahony and Barry Smyth. A recommender system for on-line course enrolment: an initial study. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 133–136, New York, NY, USA, 2007. ACM.

[137] Hans Fredrik Unelsrød. Design and evaluation of a recommender system for course selection (published MSc thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Ålesund, Gjøvik, Norway. https://brage.bibsys.no/xmlui/handle/11250/252564, 2011.

[138] Jungwon Cho and Eui-young Kang. Personalized curriculum recommender system based on hybrid filtering. In *Advances in Web-Based Learning–ICWL*, pages 62–71. Springer, 2010.

[139] Mohamad Farhan Mohamad Mohsin, Mohd Helmy Abd Wahab, Norita Md Norwawi, Cik Fazilah Hibadullah, and Mohd Zaiyadi. An investigation into influence factor of student

programming grade using association rule mining. *Advances in Information Sciences and Service Sciences*, 2:19–27, 06 2010.

[140] Patrick D Schalk, David P Wick, Peter R Turner, and Michael W Ramsdell. Predictive assessment of student performance for early strategic guidance. In *Frontiers in Education Conference (FIE), 2011*, pages S2H–1. IEEE, 2011.

[141] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong. Predicting NDUM student's academic performance using data mining techniques. In *2009 Second International Conference on Computer and Electrical Engineering*, volume 2, pages 357–361. IEEE, Dec 2009.

[142] Vo Thi Ngoc Chau and Nguyen Hua Phung. On semi-supervised learning with sparse data handling for educational data classification. In Tran Khanh Dang, Roland Wagner, Josef Küng, Nam Thoai, Makoto Takizawa, and Erich J. Neuhold, editors, *Future Data and Security Engineering. FDSE 2017.*, pages 154–167. Springer International Publishing, Cham, 2017.

[143] Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *British Medical Journal Publishing Group (BMJ)*, 3(8):1–8, 2013.

[144] Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010.

[145] Alice M Arnold and Richard A Kronmal. Multiple imputation of baseline data in the cardiovascular health study. *American Journal of Epidemiology*, 157(1):74–84, 2003.

[146] Francesco Sambo, Andrea Facchinetti, Liisa Hakaste, Jasmina Kravic, Barbara Di Camillo, Giuseppe Fico, Jaakko Tuomilehto, Leif Groop, Rafael Gabriel, Tuomi Tiinamaija, and Claudio Cobelli. A bayesian network for probabilistic reasoning and imputation of missing risk factors in type 2 diabetes. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 172–176. Springer, 2015.

[147] Asil Oztekin, Dursun Delen, and Zhenyu (James) Kong. Predicting the graft survival for heartlung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics*, 78(12):e84 – e96, 2009. Mining of Clinical and Biomedical Text and Data Special Issue.

[148] Jau-Huei Lin and Peter J. Haug. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, 41(1):1–14, 2008.

[149] I.-N. Lee, S.-C. Liao, and M. Embrechts. Data mining techniques applied to medical information. *Medical Informatics and the Internet in Medicine*, 25(2):81–102, 2000.

[150] Paul Ellett. Understanding the undergraduate grading system in the UK. `http://www.hotcoursesabroad.com/study-in-the-uk/applying-to-university/understanding-undergraduate-grading-system-in-uk/`, February 2013.

[151] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pages 5–12, Porto, Portugal, April 2008. EUROSIS.

[152] Catherine Marshall and Gretchen B Rossman. *Designing qualitative research*. Thousand Oaks, CA :Sage publications, 2014.

[153] Amanda Coffey and Paul Atkinson. *Making sense of qualitative data: complementary research strategies*. Sage Publications, Inc, 1996.

[154] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded source-*

book. Thousand Oaks, CA: Sage, 1994.

[155] Juliet M. Corbin and Anselm Strauss. *The basics of qualitative research: Techniques and procedures for developing grounded theory.(2nd ed.)*. Thousand Oaks, CA: Sage Publications, 1988.

[156] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data: the management revolution. *Harvard business review*, 90(10):60–68, 2012.

[157] Jurij Jaklič et al. The deployment of data mining into operational business processes. In *Data mining and knowledge discovery in real life applications*. InTech, 2009.

[158] Dean Abbott. Three ways to get your predictive models deployed. http://abbottanalytics.blogspot.co.uk/2013/01/three-ways-to-get-your-predictive.html, 2013.

[159] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[160] IBM Knowledge Center. Feature selection node. `https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/featureselectionnode_general.htm`, 2012.

[161] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[162] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[163] TERRY M Therneau. A short introduction to recursive partitioning. *Orion Technical Report 21, Stanford University, Department of Statistics*, 1983.

[164] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. Chapman & Hall, New York, 1984.

[165] J. Ross Quinlan. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[166] J. Ross Quinlan. Improved use of continuous attributes in c4.5. *Journal of artificial intelligence research*, 4:77–90, 1996.

[167] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer New York, 2013.

[168] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[169] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[170] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[171] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[172] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[173] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

[174] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[175] David Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[176] S James Press and Sandra Wilson. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364):699–705, 1978.

[177] Krzysztof Grabczewski. Techniques of decision tree induction. In *Meta-Learning in Decision Tree Induction*, pages 11–117. Springer International Publishing, Cham, 2014.

[178] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[179] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[180] Ismail Bin Mohamad and Dauda Usman. Standardization and its effects on k-means clustering algorithm. *Res. J. Appl. Sci. Eng. Technol*, 6(17):3299–3303, 2013.

[181] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[182] Leonard Kaufman and Peter J. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1 Norm and Related Methods*, 26(4):405–416, 1987.

[183] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[184] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning – with Applications in R*, volume 103 of *Springer Texts in Statistics*. Springer, New York, 2013.

[185] Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[186] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[187] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[188] Julia Handl and Joshua Knowles. An evolutionary approach to multiobjective clustering. *IEEE transactions on Evolutionary Computation*, 11(1):56–76, 2007.

[189] Sinan Saraçli, Nurhan Doğan, and Ismet Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1):203, 2013.

[190] Marika Vezzoli and Paola Zuccolotto. Cragging measures of variable importance for data with hierarchical structure. In *New Perspectives in Statistical Modeling and Data Analysis*, pages 393–400. Springer-Verlag Berlin Heidelberg, 2011.

[191] Thomas G Dietterich et al. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000.

[192] Norman Levinson. The Wiener (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1):261–278, 1946.

[193] Charles E Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298, 1978.

[194] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine learning Technologies*, 2(1):37–63, 2011.

[195] Robert Kopal Goran Klepac, Klepac. *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. IGI Publishing, 2014.

[196] Magne Setnes and Uzay Kaymak. Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing. *Fuzzy Systems, IEEE Transactions*

on, 9(1):153–163, 2001.

[197] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[198] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, volume 2 of *IJCAI'95*, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[199] Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566, 2002.

[200] Hun Myoung Park. Comparing group means: T-tests and One-way ANOVA using Stata, SAS, R, and SPSS. *University Information Technology Services Centre for Statistical and Mathematical Computing, Indiana University*, pages 1–57, 2009.

[201] Robin L Plackett. Karl Pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pages 59–72, 1983.

[202] IBM Knowledge Center. Matrix node output browser. `https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/matrix_browser_matrixtab.htm`, 2012.

[203] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[204] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[205] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[206] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[207] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[208] Sean M McNee, Shyong K Lam, Catherine Guetzlaff, Joseph A Konstan, and John Riedl. Confidence displays and training in recommender systems. In *Proc. INTERACT*, volume 3, pages 176–183, 2003.

[209] Paul D Allison. Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55(1):193–196, 2002.

[210] Joseph F. Hair, Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black. *Multivariate Data Analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[211] A Dempster and D Rubin. Incomplete data in sample surveys. *Sample surveys*, 2:3–10, 1983.

[212] Clifford C Clogg, Donald B Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidman. Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of the American Statistical Association*, 86(413):68–78, 1991.

[213] John Barnard and Xiao-Li Meng. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical methods in medical research*, 8(1):17–36, 1999.

[214] Stef Van Buuren, Hendriek C Boshuizen, Dick L Knook, et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694, 1999.

[215] Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American*

Statistical Association, volume 1, pages 20–34. American Statistical Association, 1978.

[216] Marco Vriens and Eric Melton. Managing missing data. *Marketing Research*, 14(3):12, 2002.

[217] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[218] Craig K Enders. *Applied missing data analysis*. Methodology in the Social Sciences. The Guilford Press, 2010.

[219] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[220] Joseph L Schafer and Maren K Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571, 1998.

[221] Joseph L Schafer. *Analysis of incomplete multivariate data*. Monographs on statistics and applied probability ; 72. Chapman & Hall, London ; New York, 1997.

[222] Sandip Sinharay, Hal S Stern, and Daniel Russell. The use of multiple imputation for the analysis of missing data. *Psychological methods*, 6(4):317, 2001.

[223] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004.

[224] Nicholas J Horton and Stuart R Lipsitz. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254, 2001.

[225] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of statistical software*, 45(7):1–47, 2011.

[226] Nicholas J Horton and Ken P Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

[227] Data Protection Act. London: The stationery office. http://www. legislation. gov. uk/ukpga/1998/29/contents, 1998.

[228] Stephen D Lapan, MaryLynn T Quartaroli, and Frances J Riemer. *Qualitative research: An introduction to methods and designs*, volume 37. John Wiley & Sons, 2011.

[229] Michael Quinn Patton. *Qualitative evaluation and research methods*. Sage Publications, inc, 1990.

[230] Bernadette E Russek and Sharon L Weinberg. Mixed methods in a study of implementation of technology-based materials in the elementary classroom. *Evaluation and Program Planning*, 16(2):131–142, 1993.

[231] Meredith Damien Gall, Walter R Borg, and Joyce P Gall. *Educational research: An introduction*. Longman Publishing, 1996.

[232] Yvonna S Lincoln and Egon G Guba. *Naturalistic inquiry*, volume 75. Sage Publications, inc, 1985.

[233] Michael Quinn Patton. *Qualitative research*. Wiley Online Library, 2005.

[234] David Morgan. *The focus group guidebook*, volume 1. Sage publications, 1997.

[235] Paul Gill, Kate Stewart, Elizabeth Treasure, and Barbara Chadwick. Methods of data collection in qualitative research: interviews and focus groups. *British dental journal*, 204(6):291, 2008.

[236] Robert H Gault. A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, 14(3):366–383, 1907.

[237] Amitav Banerjee, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18(2):127, 2009.

[238] Jacob Cohen. Statistical power analysis for the behavioral sciences. 2nd, 1988.

[239] Norman K Denzin and Yvonna S Lincoln. *Handbook of qualitative research.* Thousand Oaks, CA: Sage publications, inc, 1994.

[240] Sharan B Merriam. *Case study research in education: A qualitative approach.* Jossey-Bass, 1988.

[241] Sara Lawrence-Lightfoot and Jessica Hoffmann Davis. *The art and science of portraiture.* Jossey-Bass Incorporated Pub, 1997.

[242] Graham R Gibbs. *Analysing qualitative data.* London: Sage Publications, inc, 2008.

[243] Robert Philip Weber. *Basic content analysis*, volume 49 of *Quantitative Applications in the Social Sciences.* Newbury Park, CA: Sage Publications, Inc, 1990.

[244] MNK Saunders, P Lewis, and A Thornhill. *Research methods for business students, 5/e.* Pearson Education, 2009.

[245] Wolfgang Greller and Hendrik Drachsler. Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, 15(3):42, 2012.

[246] Rebecca Ferguson. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

[247] Jacqueleen A Reyes. The skinny on big data in education: Learning analytics simplified. *TechTrends*, 59(2):75–80, 2015.

[248] Complete university league tables. http://www.thecompleteuniversityguide.co. uk/league-tables/rankings, February 2015.

[249] The guardian league table. http://www.theguardian.com/education/ universityguide, March 2015.

[250] Jeevan Vasagar. Most graduate recruiters now looking for at least 2:1. http://www.theguardian.com/money/2012/jul/04/graduate-recruiters-look-for-21-degree, July 2012.

[251] Tamara Thiele, Alexander Singleton, Daniel Pope, and Debbi Stanistreet. Predicting students' academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, 41(8):1424–1446, 2015.

[252] John TE Richardson. The under-attainment of ethnic minority students in UK higher education: what we know and what we don‘t know. *Journal of Further and Higher Education*, 39(2):278–291, 2015.

[253] Jo Morrison, Beatrice Merrick, Samantha Higgs, and Joanna Le Métais. Researching the performance of international students in the UK. *Studies in Higher Education*, 30(3):327–337, 2005.

[254] Emma Smith and Patrick White. What makes a successful undergraduate? the relationship between student characteristics, degree subject and academic success at university. *British Educational Research Journal*, 41(4):686–708, 2015.

[255] P Marshall and EHS Chilton. Singaporean students in British higher education: the statistics of success. *Engineering Science & Education Journal*, 4(4):155–160, 1995.

[256] William Annandale. Mid-ranking universities will feel squeeze when student numbers cap ends. http://www.theguardian.com/higher-education-network/blog/2014/jul/03/end-cap-student-numbers-universities-feel-squeeze, 2014.

[257] Richard Adams. University of Oxford rebuts Cameron's claims over student diversity. http://www.theguardian.com/education/2016/jan/31/university-of-oxford-rebuts-camerons-claims-over-student-diversity, 2016.

[258] Sally Weale. Universities told to raise numbers of working-class and black students. http://www.theguardian.com/education/2016/feb/11/universities-told-to-raise-numbers-of-working-class-and-black-students, February 2016.

[259] Richard Adams. Lower government funding will hit university teaching budgets in England. `http://www.theguardian.com/education/2014/mar/26/lower-government-funding-university-teaching-england`, 2014.

[260] Deborah Hermanns. We must resist the market forces destroying our universities. https://www.theguardian.com/commentisfree/2015/oct/30/market-forces-education-system-conservative-privatised-students-march, 2015.

[261] Simon Marginson. The impossibility of capitalist markets in higher education. *Journal of Education Policy*, 28(3):353–370, 2013.

[262] Ryan Shaun Baker and Paul Salvador Inventado. Educational data mining and learning analytics. In Johann Ari Larusson and Brandon White, editors, *Learning Analytics: From Research to Practice*, pages 61–75. Springer New York, New York, NY, 2014.

[263] Jeng-Fung Chen and Quang Hung Do. Training neural networks to predict student academic performance: a comparison of cuckoo search and gravitational search algorithms. *International Journal of Computational Intelligence and Applications*, 13(01):–1, 2014.

[264] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, and Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472, 2013.

[265] Leland Wilkinson. Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, 54(8):594, 1999.

[266] Stephen G West. New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods*, 6(4):315, 2001.

[267] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *American Psychological Association*, 7(2):147–177, 2002.

[268] Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

[269] Terry M Therneau, Elizabeth J Atkinson, et al. An introduction to recursive partitioning using the rpart routines. *Technical Report 61. URL http://www. mayo. edu/hsr/techrpt/61. pdf. Mayo Clinic, Rochester (MM)*, pages 1–59, 1997.

[270] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[271] RStudio's Team. Rstudio: integrated development for R. *RStudio, Inc., Boston, MA. URL https://www.rstudio.com/products/rstudio/release-notes/*, 2016.

[272] Rebecca Eynon. The quantified self for learning: critical questions for education. *Learning, Media and Technology*, 40(4):407–411, 2015.

[273] Ramez Elmasri and Shamkant Navathe. *Fundamentals of database systems*. Addison-Wesley Publishing Company, 2010.

[274] Jill Collis and Roger Hussey. *Business research: A practical guide for undergraduate and postgraduate students*. Palgrave Macmillan, 2013.

[275] Neil Selwyn. Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1):64–82, 2015.

[276] Ben Daniel. Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5):904–920, 2015.

[277] Mark Lycett. 'datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, 22(4):381–386, July 2013.

[278] Viktor Mayer-Schönberger and Kenneth Cukier. Big data: A revolution that will transform how we live, work, and think. *Houghton Mifflin Harcourt*, 2013.

[279] Kayode Ayankoya, Andre Calitz, and Jean Greyling. Intrinsic relations between data science, big data, business analytics and datafication. In *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology*, page 192. ACM, 2014.

[280] Imanol Ordorika and Marion Lloyd. International rankings and the contest for university hegemony. *Journal of Education Policy*, 30(3):385–405, 2015.

[281] Marie-Laure Bougnol and Jose H Dulá. Technical pitfalls in university rankings. *Higher Education*, 69(5):859–866, 2015.

[282] Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. Mit Press, 2008.

[283] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[284] Will Leggett. The politics of behaviour change: Nudge, neoliberalism and the state. *Policy & Politics*, 42(1):3–19, 2014.

[285] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM, 2002.

[286] Brent Smith and Greg Linden. Two decades of recommender systems at amazon. com. *IEEE Internet Computing*, 21(3):12–18, 2017.

[287] M Benjamin Dias, Dominique Locher, Ming Li, Wael El-Deredy, and Paulo JG Lisboa. The value of personalised recommender systems to e-business: a case study. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 291–294. ACM, 2008.

[288] Cass R Sunstein. Nudging and choice architecture: Ethical considerations. *Yale Journal on Regulation, Forthcoming*, 2015.

[289] Joanna Jones, Ruth Gaffney-Rhys, and Edward Jones. Handle with care! an exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (set) results. *Journal of Further and Higher Education*, 38(1):37–56, 2014.

[290] Peter Davies, Jean Mangan, Amanda Hughes, and Kim Slack. Labour market motivation and undergraduates' choice of degree subject. *British Educational Research Journal*, 39(2):361–382, 2013.

[291] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

[292] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[293] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[294] Avi Goldfarb and Catherine E Tucker. Privacy regulation and online advertising. *Management science*, 57(1):57–71, 2011.

[295] Sue Newell and Marco Marabelli. Strategic opportunities (and challenges) of algorithmic

decision-making: A call for action on the long-term societal effects of datification. *The Journal of Strategic Information Systems*, 24(1):3–14, 2015.