UNIVERSITY OF HUDDERSFIELD

DOCTORAL THESIS

Data Analytics and Machine Learning to Understand and Predict Student Performance

Author:
Azhan RASHID

Supervisor:
Dr. George BARGIANNIS

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

in the

Centre for Mathematics and Data Science
Department of Computer Science

May, 2023

Declaration

I, Azhan RASHID, declare that this thesis titled, "Data Analytics and Machine Learning to Understand and Predict Student Performance" and the work presented in it are my own. I confirm that:

- The author of this thesis (including any appendices and/ or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching.
- Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Details of these regulations may be obtained from the Librarian. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.
- The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

Signed: Azhan Rashid

Date: Wednesday 17th May, 2023

Abstract

Predicting student performance has attracted significant research interest in recent years, owing primarily to its potential benefits to both students, in terms of improving outcomes and post-graduation prospects, and educational institutions, in terms of addressing issues such as differential attainment and targeted proactive support of students at risk of lower performance. Substantial research effort has been devoted to exploring data analysis and machine learning techniques in this context. One of the main challenges is the availability of large and high-quality datasets and associated issues such as data imbalance and limited scope of data analysis. Additionally, most researchers focus on predicting performance in the form of a single predicted score, as opposed to a range of potential outcomes.

In this thesis, the aforementioned research gaps are addressed through a computational framework to predict student performance ranges using data analysis and machine learning. The framework contains a unique combination of layers ranging from data pre-processing to statistical analysis and learning prediction models, with each layer carefully positioned to avoid any biased outcomes. This increases confidence in the produced outcomes.

The proposed framework is validated using a rich, anonymised dataset provided by the University of Huddersfield that contains significantly more samples and relevant variables than what is commonly observed in the literature. Experiments focus on predicting the performance of students based on data available at the point of enrolment. This includes students that are completing their pre-qualifications for entrance (e.g. A Levels) and allows exploring the widest possible group of students available in the dataset. The predictions produced from the conducted experiments represent a range of overall grade achievement (boundaries) at the end of their course.

Results show an accuracy of 84%/86% (worst/common case scenario). Baseline comparison shows an improvement of 3%/5% (worst/common case scenario) compared to existing literature. In most cases, improvement is seen in both the best and the worst performing models. This robustness of the framework can be partly attributed to including means of tackling data imbalance, as well as exploring a wide range of data analysis and machine learning models.

The main contributions of this thesis and the included framework involve: predicting students' performance in the form of a range; integrating approaches to tackle imbalanced data; performing in-depth data analysis using a range of statistical methods; and considering both supervised and unsupervised learning algorithms. It is envisioned that the framework can be integrated into existing student performance dashboard systems, allowing academics and administrators to harness its predictive capabilities and drive decision-making to improve outcomes across the student body or targeted efforts, such as reducing differential attainment.

Acknowledgements

The research is supported in collaboration with the University of Huddersfield. The institution has provided a commercial standard dataset. The dataset is rich with many details to explore, it is suitable to evaluate reliable evaluations. They kindly funded the entire project from tuition fees to living expenses.

It is thanks to all the supervisors that contributed to the research project. The research could not be accomplished without their support. The background knowledge and resources they provided are worthwhile. The supervisors include:

- Dr George Bargiannis
- Dr Sofya Titarenko
- Dr Jarek Bryk
- Prof Andrew Crampton
- Prof Fionn Murtagh

Finally, credit should also go to the family.

Contents

List of Figures

List of Tables

List of Abbreviations

D	The dataset applied to the work.
D_{train}	The training set split from D.
$D_{testing}$	The testing set split from D.
C_{min}	The student's minimum performance from the Consistency-Scale.
C_{max}	The student's maximum performance from the Consistency-Scale.
C_{avg}	The student's average/typical performance from the Consistency-Scale.
PSO	Particle Swarm Optimisation.
RFE	Recursive Feature Elimination.
GA	Genetic Algorithm.
RFC	Random Forest (Classification).
ADAC	AdaBoost (Classification).
KNNC	K-Nearest Neighbours (Classification).
GBC	Gradient Boost (Classification).
GPC	Gaussian Process (Classification).
PAC	Passive Aggressive (Classification).
ETC	Extra Trees (Classification).
XGBC	Extreme Gradient Boost (Classification).
NN18	(Feed Forward) Neural Network 1 Hidden Layer with 8 nodes (Classification).
NN38	(Feed Forward) Neural Network 3 Hidden Layers with 8 nodes (Classification).
NN58	(Feed Forward) Neural Network 5 Hidden Layers with 8 nodes (Classification).
NN108	(Feed Forward) Neural Network 10 Hidden Layers with 8 nodes (Classification).
NN14	(Feed Forward) Neural Network 1 Hidden Layer with 4 nodes (Classification).
NN34	(Feed Forward) Neural Network 3 Hidden Layers with 4 nodes (Classification).
NN54	(Feed Forward) Neural Network 5 Hidden Layers with 4 nodes (Classification).
NN104	(Feed Forward) Neural Network 10 Hidden Layers with 4 nodes (Classification).
KMU	K-Modes Clustering.
KPU	K-Prototype Clustering.
UK	United Kingdom.
D.O.B	Date of Birth.
PCA	Principal Component Analysis.
MCA	Multiple Correspondence Analysis.
FAMD	Factor Analysis of Mixed Data.
SVM	Support Vector Machine.
KM	Kilometers.
MSE	Mean Squared Error.
MAE	Mean Absolute Error.
AI	Artificial Intelligence.
ML	Machine Learning.

Chapter 1

Introduction

1.1 Background

The focus of this thesis is on predicting student performances through the use of data analytics and Machine Learning algorithms. This is an active interdisciplinary research field in AI & Education. The major challenge in terms of predicting performance is producing outcomes with the best accuracy. High prediction accuracy results in precise early warning signs for those who are on the verge of performing poorly in their academic studies. The typical input data in this prediction process includes student records collected by educational institutions (or equivalent). The output is an evaluation of a student's academic capabilities, often looking at particular year groups or levels of study (e.g. undergraduates).

The purpose of this research is primarily to support institutions in identifying the best & poor performing students. In the latter case, follow-up actions involve evaluating solutions to tackle concerns and avoid students at risk of poor performance (in terms of their grades). Tackling at-risk students benefit both institutions and the students themselves. The institution's reputation increases and students have wider opportunities after graduation.

Research in the use of AI in student performance prediction has been active since the early 1990s. An early example is using classification analysis to predict retention using doctoral and master's students (Pyke and Sheridan [68]). Since then a wider range of algorithms has been embedded in this topic. After the first phase, additional types of Machine Learning models from the Supervised Learning family have been applied (Adekitan and Noma-Osaghae [2], Li, Lynch, and Barnes [55]) such as Random Forest, Support Vector Machines, & Naive Bayes. In this second phase, the models delivered many positive outcomes compared to the first phase. Such models continue to be explored in the context of student performance prediction in more recently published research (Mengash [60], Francis and Babu [31]). It is also worth noting that Classification (especially Multi-Classification) is more common than Regression with regard to predicting student performance.

Feed Forward Neural Networks have also attracted considerable interest, especially in recent years, in relation to predicting student performance, achieving improved prediction accuracy compared to other Supervised Learning approaches (Waheed et al. [80], Francis and Babu [31]). The benefits of using Feed Forward Neural Networks are primarily due to their ability to detect complex nonlinear relationships. The common approach by researchers seems to involve applying one Feed Forward Neural Network with 2 or fewer hidden layers. In rare cases, more than 2 hidden layers are included (Waheed et al. [80]). The answer to the question of whether more hidden layers improve prediction accuracy is still unclear.

Other Supervised Learning techniques such as Gradient Boosting have also been explored, but less commonly. A systematic literature review led to only a few papers

using these recent algorithms (Alwarthan, Aslam, and Khan [8]). These papers are focused on detecting student dropouts (Niyogisubizo et al. [62], Kiss et al. [50], Adnan et al. [3]). This can be related to predicting student performance by associating dropout incidence with poor student performance.

Researchers have also comparatively explored different supervised learning algorithms in their experiments (Bujang, Selamat, and Krejcar [20], Masangu, Jadhav, and Ajoodha [59]). In cases where a mixture of Supervised Learning algorithms are applied, established ones are more likely to be included than relatively more recent ones (Rodríguez-Hernández et al. [69]). Moreover, the common Supervised Learning model amount is $3-4$ in one experiment setting. Finally, it should be noted that Unsupervised Learning is very uncommon in the context of student performance prediction, with only one recent study identified in a recent systematic literature review (Alwarthan, Aslam, and Khan [8]).

Commonly used datasets in the existing literature involve student records that contain qualitative and quantitative variables such as date of birth, ethnicity, gender, & grades. Time series of academic cycles such as semesters are also used (Zhang and Rangwala [88]). The experiment process involves collecting educational data, applying pre-processing and feeding them into a selection of decision-making models, and returning and evaluating prediction results. In some cases, there are additional steps applied. These include feature selection, which chooses the best combination of variables (Greatorex-Voith and Anand [34], Xu, Moon, and Schaar [85]) and cross-validation (Helal et al. [39], Yang et al. [86]).

1.2 Research Gaps

In one literature review, 357 papers on predicting student performance from $2010-2018$ are evaluated (Hellas et al. [40]). The common decision-making model contributes 18% (Linear Modeling) to the total number of papers collected. The common features to derive student performance include course / pre-course performance (e.g. quizzes) contributing to 13%. Even though the rise of (Feed Forward) Neural Networks is occurring more recently, Machine Learning is popular in all years (even in the latest year).

The same applies to another literature review, 44 papers since 2022 are evaluated (Alwarthan, Aslam, and Khan [8]) and analysis shows that the best predictors are ensemble learning models like Random Forest. The popular Supervised Learning models applied are those used back in the 1990s (e.g. Logistic Regression). And the common variables include course / pre-course performance contributing to 72%. Again, Feed Forward Neural Networks have been shown to be applied but not so much compared to other Supervised Learning methods. The common sample amount of a dataset ranges from $30-300$.

In addition, other literature reviews (Chaudhry and Kazim [24], Dignum [26]) show more current research challenges in AI for education. Many scopes of areas were explored but one interesting challenge is the ethical side of using such data. This is due to the enormous amount of personal data. Ethical issues also relate to many legal barriers in studied countries. They also relate to collecting a suitable dataset that contains sufficient samples & variables. This is the other challenge with predicting student performance accurately as a rich high-quality dataset is required to become reliable. Other forms of challenges (unrelated to this thesis) are interactions with robotic teachers & online studying (e.g. mental health).

Even though there has been extensive relevant research, some important issues that deserve further exploration have been identified.

First, imbalanced data exist in many datasets in this field. The effect of exploring potential methods is lacking. Since the issue cannot be fully enclosed, one can introduce methods to reduce the chances. Second, student performance evaluation has always been with performance values but tends to ignore outcomes with boundary limits (maximum/minimum performance ranges). In real life, this can be better suited as students' academic performance fluctuates.

A third common issue is the lack of in-depth data analysis. The analysis explored is insufficient and does not fully explore the strengths and weaknesses of approaches from a statistical perspective. A fourth and final research gap involves the usage of less explored approaches in the area, such as integrating Unsupervised Learning and applying several Feed Forward Neural Network architectures with different hidden layers at once. Since Unsupervised Learning is not designed for predictions, a model pipeline process needs to be developed. This means it is one of the pre-processing steps within it. Due to its clustering nature, a suggested duty can be grouping similar student data before producing predictions.

The aforementioned gaps are interrelated, and this thesis aims to address all of them to an extent in the next section.

1.3 Aim & Objectives

The main aim of this thesis is to design, develop and validate a computational framework for predicting student performance in higher education that focuses on maximum/minimum performance ranges and is capable of achieving good results while working with different Machine Learning approaches and imbalanced datasets.

To achieve this aim, the following objectives need to be met:

- To predict a performance range for students that describes the boundaries of academic ability. The current trend is predicting a performance value. This is to address the fact that a performance value ignores many circumstances that are likely to occur in their studies, such as finance, personal family, mental, health, etc.
- To develop an algorithm to reduce the probability of imbalanced data, overfitting, and underfitting. This involves exploring for which variables this is possible and for which it may not be (such as ethnicity), as well as investigating different machine learning models to achieve this, including both supervised and unsupervised approaches.
- To validate the developed computational framework through in-depth data analysis with a given dataset, alongside determining optimal configuration and outcomes in each layer of the framework.

1.4 Contributions

This thesis makes contributions to research at the confluence of AI and Data Analytics and Education. It delivers novel methods for evaluating student performances that can be better suited commercially, and practically. The results can be directly useful either to Education researchers that are exploring the use of AI and Data Analytics or to AI and Data Analytics researchers that are either looking to explore

Education as a domain or are exploring domains with similar characteristics to Education.

With respect to the stated aim & objectives, the contributions of this thesis are:

- Performance Range: The concept of expressing student performances with a range output is unexplored. The contribution here is the output format, which provides an insight into their minimum/maximum academic capabilities. A performance range can be useful to identify hidden patterns such as grouping student ranges and mapping them with the necessary resources.
- Imbalanced Data: There has been limited focus on addressing imbalanced data in this application area. The contribution here is the algorithm process that takes into account imbalanced data, which can be a potential solution for resolving the imbalanced data problem. Reducing the probability can deliver preciseness & reliability to the prediction results (from the decision-making models). The algorithm is designed to balance the ratio of data as much as possible.
- In-Depth Data Analysis: Basic data analysis is applied in existing research which includes data specification and generic statistics (e.g. correlation). Other useful analyses that this thesis contributes to include dimensional reduction, hypothesis tests, quantile, skew, and so on. This is important because one can evaluate how the data should be applied to an experiment that returns the most reliable evaluations.
- Different Feed Forward Neural Networks Architectures: Even though (Feed Forward) Neural Networks have been applied previously, architectures commonly consist of a single network with $1-2$ hidden layers. The contribution here is introducing several architectures with different hidden layers in one setting. It is an opportunity to determine if hidden layers provide improvement to prediction accuracy.
- Unsupervised Learning Integration: Unsupervised Learning in this research niche has been limited due to its clustering nature (not compatible with student performance predictions). This thesis contributes to existing research by introducing a model pipeline process that includes Unsupervised Learning as one of the pre-processing steps. Its purpose involves grouping similar student data before producing predictions.

1.5 Thesis Layout

The remainder of this document is organised as follows:

- Chapter 2 - Literature Review: This chapter explores past relevant works and identifies missing gaps existing in the literature, as well as challenges, trends, & motivations. This includes showcasing the broadness of the field.
- Chapter 3 - Computational Framework: This chapter explains in-depth each layer of the proposed computational framework providing the reader with an understanding of its purpose. It also provides a specification of the input and output requirements of each layer, such as data cleaning requirements and the model process to return predictions.

- Chapter 4 - Experiment: In this chapter, the validation of the computational framework presented in Chapter 3 is provided using a dataset explained at the beginning of the chapter. It provides the configurations, comparison with past works, and results.
- Chapter 5 - Experiment Evaluation: An explanation of the results produced in Chapter 4 (Section 4.7) is provided here in full detail. This includes baseline comparison with similar past works (application and benchmarks results), determining the best and worst outcomes, and so on. The chapter also evaluates whether the aim and objectives have been met and what learning has been achieved.
- Chapter 6 - Conclusion: The chapter summarises the work in this thesis and identifies opportunities for further research.

Chapter 2

Literature Review

2.1 Introduction

This chapter expands on the literature overview in Chapter 1 by presenting the details of the systematic literature review conducted. In this introductory section, the setup of the systematic literature review is provided. The papers summarised in this chapter relate to AI and Education with the main focus being on predicting student performance using decision-making algorithms (e.g. Supervised Learning). For readers unfamiliar with machine learning and data analysis, Appendix A provides the necessary background knowledge.

2.1.1 Research Questions

Before exploring past works, it is important to set the main questions that are to be addressed by the systematic literature review. Having a set of questions delivers the expectation of what information is worth exploring and not, it is also time efficient and helps exclude irrelevant content. The main questions are the following:

- What are the challenges, state-of-the-art research, and trends in predicting student performance?
- What are the most common and least common AI and data analytics techniques employed for predicting student performance?

The main questions contribute to the core of the thesis & this chapter provides the answers. There is a chance that this chapter provides answers to additional questions, even though this may be the case, they are not a compulsory requirement. It provides answers to the latest missing gaps that can deliver usefulness to the research community. It is important to avoid duplicate work as it doesn't contribute new knowledge to the research field. So, reading papers published in recent years should be considered.

2.1.2 Search Strategy

The literature review relies on several search techniques. This includes using keywords, reading the abstract, and/or filtering based on the published year. Collecting recent papers provide an understanding of the current trend, challenges, and state-of-the-art research. The referencing style first cites the author(s) alongside the reference number that is listed in the bibliography (an example: Helal et al. [39]).

To find papers, a search strategy is required. An automated keyword search is used, with two levels of keywords, conjunctively combined with AND. The first

level includes the term *predict student performance*. The second level includes a disjunction of different terms related to the topic: *Supervised Learning*, *Neural Networks*, *Regression*, *Classification*, *Academic*, *Education*, *Grades*, *Student Data*, *Dropout*. These search levels help narrow down to specific papers.

Several different platforms were used, primarily Google Scholar, but also IEEE, ScienceDirect, Arxiv & ResearchGate. Papers mapped with highly reputable platforms are considered as it ensures authenticity, trustworthiness, and usefulness to the research community. The common filtering options applied here include countries, keywords, & years.

With regard to year filtering, it is important to know the state-of-the-art research. As a result, most papers must be from recent years. Although the work can have similar methodologies (between older and more recent papers), it may include unique details that can be expanded to introduce new knowledge to the community. An example can be a recent decision-making algorithm. Older papers should not be ignored completely, as those papers are more likely to set trends and be well-cited.

Furthermore, authors and papers should have a track record of citations, especially in recent years. Not having citations reduces trust in the value of the work. A good record of citations provides confidence that papers are of high quality and contain useful research results. All platforms employed provide features related to discovering citation-related information.

2.1.3 Search Scope

The papers applied for this literature review must follow specific rules. These rules are applied to deliver high-quality content to the reader. Not having high-quality criteria can lead to uncertainty about past relevant research.

Firstly, the language must be English. This is because communication must be there in order to avoid wrong summarising (e.g. wrong aims & findings). One may make the argument of using translation software, but there can be occasions that the word may be interpreted in the wrong manner. This can lead to misunderstanding and can make the outcomes of the literature review unreliable.

Secondly, papers must mention sufficient details for reliability purposes which are using relevant data, methodologies, and results (with discussion). Exploring & understanding papers is important to evaluate the similarities and differences in this research. It provides details of the current work done and which work is yet to be done. If a paper does not provide sufficient information on its scope then it can lead to not contributing new knowledge to the field.

Thirdly, papers must relate to AI & Education. They should provide benefits toward predicting student performances with AI technologies. This means the scope, data, and methods should contribute to this thesis in some way. Examples include data analysis, pre-processing, and machine learning algorithm application.

In addition, it is important to state exclusion criteria. This systematic literature review excludes aspects of AI & Education that go beyond predicting student performance such as robotic teachers, & tools for online course platforms. Also, papers with no citations have been excluded to ensure that included content has been considered worth referring to at least once.

2.1.4 Related Surveys

There are four literature review papers that investigate many past papers on the same topic. Table 2.1 provides details for these surveys. The *Time Period Explored*

defines the year range of papers covered within a case study, *Reviewed Studies* defines the number of papers covered within a case study, and *Focus* mentions the topics each case study covers. These literature review papers are analysed in more detail in Section 2.3.4.

TABLE 2.1: Related surveys on AI & Education (predicting student performances).

Paper	Time Period Explored	Reviewed Studies	Focus
Alwarthan, Aslam, and Khan [8]	2010 – 2022	44	Challenges, state-of-the-art research, trends.
Hellas et al. [40]	2010 – 2018	357	Challenges, state-of-the-art research, trends.
Chaudhry and Kazim [24]	2011 – 2021	85	Challenges, state-of-the-art research, trends.
Dignum [26]	2000 – 2020	35	Challenges.

2.2 Classification Of Reviewed Studies

2.2.1 Year And Country Of Publication

This research field has been quite prolific in terms of publishing work in relation to predicting student performance with AI techniques. Since the early 1990s, work in this area has been continuously published by the research community. Although many researchers may use the broad term AI, this does not necessarily mean all forms of AI are applied. In most cases, the papers generally refer to a specific family of approaches such as Supervised Learning. This is because AI is a very broad field and using a broad title in papers can be confusing to readers. Also, different AI techniques have different requirements which may not be compatible with the given problem.

Despite the extensive publication record, methodologies employed tend to be similar with minor differences. Furthermore, the statistics from the collected past papers should be discussed. Figures 2.1 and 2.2 provide line charts of the papers identified in this literature review based on year and papers filtered by country.

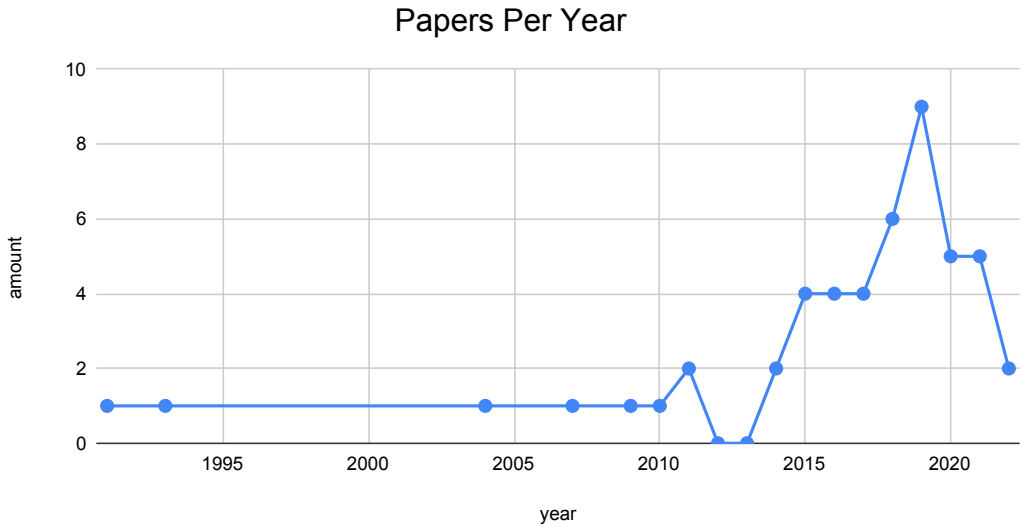

FIGURE 2.1: Distribution of reviewed papers per year.

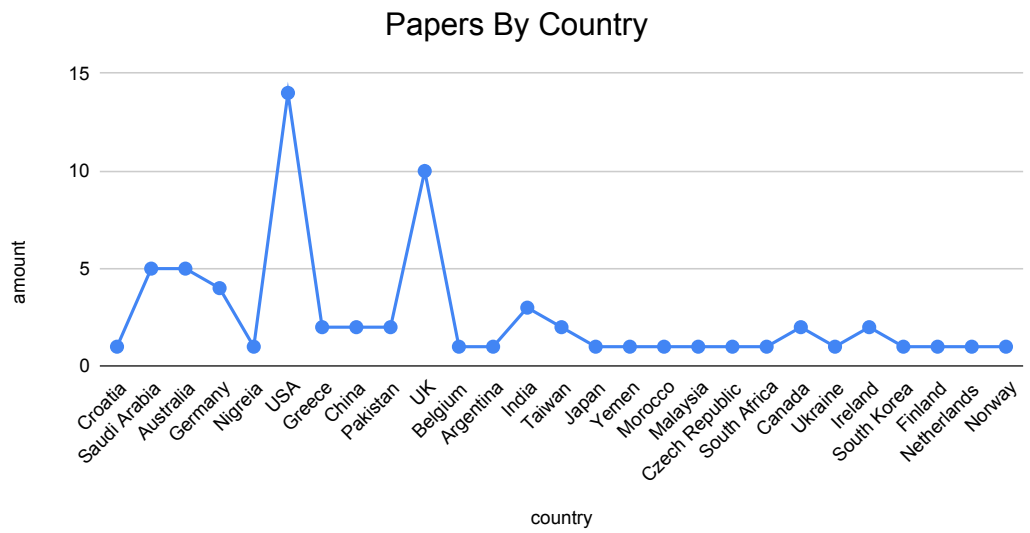

FIGURE 2.2: Distribution of reviewed papers per country.

2.2.2 Citations & Keywords

Occasionally, the common keywords that exist in past papers include *Education*, *Machine learning*, *Supervised Learning*, & *Neural Networks*. In terms of titles and abstracts, the phrase *predicting student performance* or similar is mentioned. Sometimes, titles may include the decision-making family (e.g. Supervised Learning). The common details that are not mentioned in abstracts include the data specification (e.g. samples, variables), & further work. Table 2.2 provides the number of occasions a paper includes the aforementioned terms.

TABLE 2.2: Occurrence of common terms associated with papers.

Keyword	Amount	%
Supervised Learning	42	86%
Neural Networks	12	24%
Education	49	100%
Random Forest	26	53%
Regression	36	73%
Predicting Student Performance	42	86%
Data	49	100%
Machine Learning	42	86%

In terms of paper citations, the highest citation found in a paper is above 70. It does depend on the time of publication as older papers are more likely to be referenced. Most papers (reviewed in this thesis) are from recent years, so a substantial citation amount is less likely. Also, the research field is another factor, a small research niche is likely to have a lower number of citations than a more commonly researched area.

2.2.3 Authors & Institutions

As shown in Figure 2.2, authors are from many countries, and most of them (as expected) are from English-speaking (in terms of first language) countries such as the

USA, & UK. Generally speaking, most papers include several authors. The range of authors in a paper is from $1-10$. The most common case is $2-4$ authors. Authors are commonly from the same institution and/or country, but this is not the case across all reviewed studies. It is more likely that papers originate from the same country than institutions due to the wider amount of authors. There are 14 papers originating from the same country with one or several authors. A potential reason for several authors in one paper could be to show authenticity among the communities and readers. The citation amount from authors generally does deliver authenticity and sometimes (from the papers in this chapter), some authors have over 200 citations on a yearly basis. Figures 2.3 and 2.4 provide the number of authors and institutions associated with a paper.

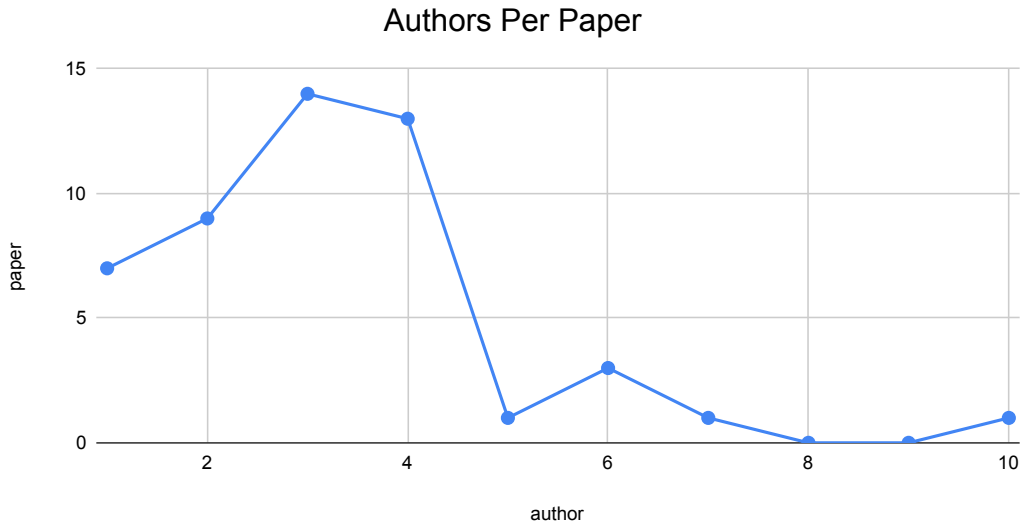

FIGURE 2.3: Distribution of reviewed papers per number of authors.

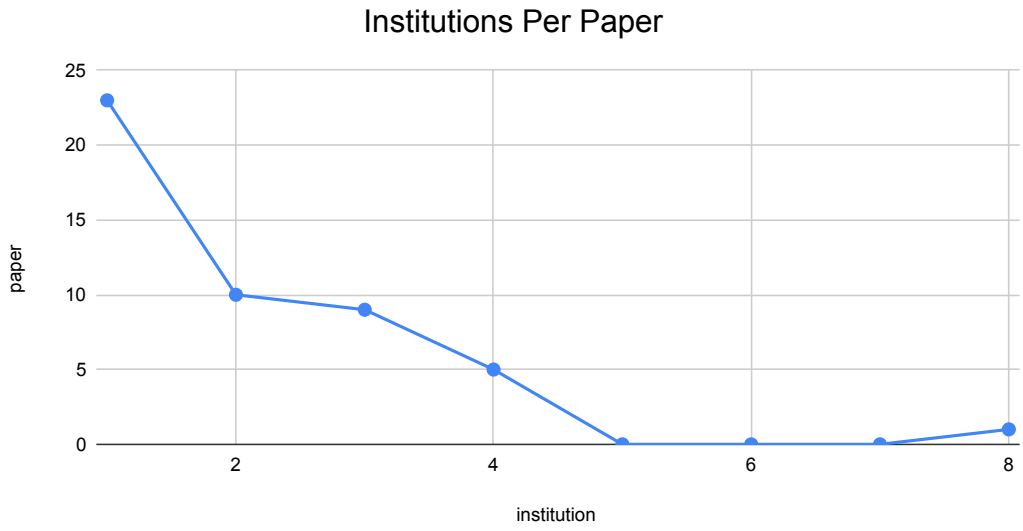

FIGURE 2.4: Distribution of reviewed papers per number of institutions of authors.

2.3 Analysis Of Reviewed Studies

2.3.1 Chapter Layout

The remainder of the chapter summarises past studies based on the most to least relevant works (to this thesis). The judgment on which research papers are relevant is investigating the similarities against this thesis. This includes exploring the aim of the research (predicting dropouts/grades), data (e.g. owner, variables, target audience, category), predictive methods (e.g. Supervised Learning), and metrics (e.g. AUC, & PRAUC).

Papers are subdivided into the following groups:

- Single Supervised Learning Models
- Multiple Supervised Learning Models
- Insights from Existing Surveys
- Other Related Literature

In each group (section), the papers are sorted based on most to least relevancy. The analysis first showcases papers directly related to the scope of this thesis and then provides other papers that still have similarities (e.g. using student data). The groups are a mixture of machine learning algorithms and educational research goals. Each section includes an introduction, papers, and then a conclusion. Afterward, a summary of the past works alongside the missing gaps is discussed.

2.3.2 Single Supervised Learning Models

This section provides papers that apply Supervised Learning. The papers are in relation to AI & Education (specifically student performance predictions). The research is subdivided into three categories: *Linear/Logistic Regression*, *Feed Forward Neural Networks*, and *Others*. The *Others* category includes SVM, Naive Bayes, and K-Nearest Neighbours (KNN).

Linear/Logistic Regression

Utzman, Riddle, and Jewell [78] predict student performance for entry using specifically enrollment data. The aim is to evaluate if admissions data can be used to predict the risk of poor performance from a therapy exam (also known as the National Physical Therapy Examination or NPTE). According to the paper, the possible grade outcome is between $1-4$, where grade 1 is the only pass. The data in the study contributes to 3365 students given by twenty physical therapist education programs. 2941 of students passed the first attempt with a minor number of students never passing (46). The grades observed are uGPA, vGRE, qGRE. The study included exploring the relationship between student admission data using a Logistic Regression. The details of the data are not fully enclosed but it does mention some variables such as age and ethnicity. The coefficient scores & odds ratio (95%) between admission data, year cycle, and past grades (e.g. uGPA) are explored against NPTE (as target). It also demonstrated the relations using the AUC/ROC metric. They found that students with high-level qualifications (e.g. PhD), race, ethnicity, and previous grades (undergraduate GPA or uGPA, verbal GRE or vGRE, quantitative GRE or qGRE) can be useful for evaluating a student's chances of failing the

NPTE test. But there are variations with no evidence of a direct relationship. Furthermore, It delivered motivation for future NPTEs exams to record race & ethnicity as it isn't common practice used in previous years.

Yang et al. [86] compare multiple Linear Regression with and without Principal Component Analysis (PCA) to evaluate the academic performance of students at an early stage. The PCA is used to improve the prediction accuracy. The data is a blended Calculus course from Open edX & Maple T.A. Open edX is the video viewing data and Maple T.A. are exercises (home completions, and quiz grades). The data contains 58 (33 males & 25 females) freshman students from Northern Taiwan, between $2015 - 2016$. The data is first pre-processed (e.g. imputing missing data), and PCA is applied to reduce the number of independent variables and extra the new (condensed) variables which are then trained with the multiple Linear Regression. This builds the student academic performance prediction by factor scores extracted by PCA. The benchmarks first apply cross-validation (10 folds) and then the result is returned. The findings show using the PCA alongside multiple Linear Regression significantly improved by almost 50% than not using PCA (according to pMSE). When the data is reduced to 6 components (using PCA), the first variable contributes 81%, showing intense accuracy in lower dimensions.

Pereira [64] applies 2 Linear Regression models to predict student grade scores based on FAMD outcomes. The paper contains much more analysis compared to the majority of papers such as hypothesis tests, FAMD, histograms, and several statistics (e.g. median & mean). The data contains 1044 students and 334 variables (e.g. age, study time, paid extra classes) from a school in Portugal. Two groups are explored: 1) the features that impact the final score; 2) exploring family-related variables to student performance (e.g. parent careers).

The FAMD is used to select the best-suited variables to predict their grades (between these groups). Their variance of best accuracy is 4 dimensions (55% accuracy) for the first group and 7 dimensions (64% accuracy) for the second group. The first group shows 6 variables (e.g. age & health) contributing to the first dimension. The second group shows 8 variables (e.g. absences & guardian) contributing to the first dimension. The results of linear regression show that the accuracy is 78% (for the first group, RMSE: 2.38) and 76.28% (for the second group, RMSE: 2.61). The paper demonstrates that variables from different groups (e.g. family, school) impact their performance (not just one). A suggestion for future work in this paper is to perform more tests with more dimensions.

Pyke and Sheridan [68] used a Classification analysis to predict the retention of 477 master candidates and 177 doctoral candidates. They conducted the outcomes using a list of characteristics such as gender, age, degree completion, funding type, etc. The analysis included exploring the coefficient scores between the student's data using a Logistics Regression. They also explored patterns between characteristics (e.g. gender) and student performance. The findings show that high-performing master's students have a better source of funding, support, and course duration type (full-time or part-time). Doctoral students on the other hand show funding and support associated with their success. The judgment is based on Odds-Ratio and prediction accuracy benchmark which is $81\% - 88\%$ for doctoral and $77\% - 81\%$ for master students. They also included an analysis of characteristics of the percentage of graduated and withdrawn (e.g. gender).

Jiang et al. [49] uses MOOC (Massive Open Online Courses) data to evaluate student behaviour in their first week. The data is derived from a Biology course offered by the University of California, Irvine on the Coursera platform. The course is four weeks with three units. Each unit contains assessments, quizzes & videos. The data

is collected from different sources including Coursera and the university's registry. A total of 232 from the University of California (Irvine) and 172 from Coursera. Four types of predictors are used: average first-week quiz score, number of peer assessments in week one, social network engagement, and confirmation if the student is from the university or the online course. The evaluation is judged with two Logistic Regression models and observation is based on the Odds Ratio and metrics such as Accuracy. The first model applies to the certificate the learner gets (Distinction or Normal certificate). The second model applies whether students get the Normal certificate or not. The authors found that the number of peer assessments is the biggest factor in achieving a Distinction certificate. Each peer assessment completed makes it seven times more likely to obtain a Distinction certificate than a Normal certificate. In the second model, the major factor is the quiz scores. It shows that students from the university registry are more likely to perform better than distant learners. The evaluation shows that the first model is more able to predict (93%) than the second model (80%). The results show that assessment performance and being social are important factors to judge student behaviour in week one.

Jaber et al. [47] explores the correlation between the National Board of Medical Examiners Surgery Shelf Exam (NBME/SSE) and weekly quizzes. They applied a dataset of 156 third-year students that completed their studies in $2015 - 2017$. Several kinds of correlation, Linear, and Logistic Regression are analysed to produce a conclusion. Two quizzes occur on a weekly basis (for 12 weeks) and the grading is between $0 - 5$ for each quiz (0 is the worst). Students are given quizzes to test their medical knowledge.

Three models are used in each analysis. One analysis of models explores the independent associations between each weekly quiz and NBME scores using Beta coefficients. The other analysis of models explores the independent associations between the weekly quizzes and the probability of achieving below 70 points on the NMBE exam. Each analysis contains both Linear Regression and Logistic Regression.

The first analysis applies two Linear Regression models and one Logistic Regression model. Model 1 reports the univariate analysis with Linear Regression. Model 2 reports the coefficient Beta scores using several Linear Regressions that tweak the variables in Model 2. Model 3 uses the same variables as Model 2 but it contains further tweaking on the academic block and year variables. These models are tested with three topics (Trauma/Burns, Esophagus/Anorectal, Wound/ICU).

The second analysis applies one Linear Regression model and two Logistic Regression models. Model 1 reports the univariate analysis with Linear Regression. Model 2 explores the Odds-Ratio (95% CI) using several Logistic Regression models to tweak the variables in Model 2. Model 3 uses the same variables as Model 2 but it contains further tweaking on the academic block and year variables. These models are tested with two topics (Cardiac/Vascular, Wound/ICU).

The findings discovered from the first analysis show an increase in the NBME shelf surgical exam performance from all three topics. The second analysis shows positive predictive outcomes for students who achieve less than 70 points on the SSE. It also shows later weeks of the course deliver higher performance for both weekly quizzes and the SSE. The outcome is similar to the correlation analysis such as Kruskal Wallis.

Feed Forward Neural Networks

Bilal et al. [18] predicted student performances in Croatia using enrollment data. The enrollment data includes gender, the parent's education, and monthly income. The total number of samples is 76 students and 17 variables of characteristics. The grading variable is an ordinal data type that consists of 5 outcomes (1 is the lowest and 5 is the best). The model applied is a Neural Network of one hidden layer. The input neurons are the total number of inputs, the only hidden layers contain 28 neurons and an output layer of one neuron. The evaluation shows an accuracy of over 90%. The findings show that admission data can be used to predict student performance before enrollment.

Rodríguez-Hernández et al. [69] published a paper that applies procedures to a *Systematic* Neural Network to predict academic performance in higher education. Also, compare the performance with several well-known predictors such as Logistic Regression and Random Forest. The dataset applied consists of 162030 students and 4 variables (e.g. socioeconomic) from an institution in Columbia. The *Systematic* Neural Network involves tuning the momentum & learning rate hyperparameters until it reaches a condition. The finding shows the iterative tuning in the hyperparameters managed to improve accuracy, ranging from $71\% - 82\%$. The Neural Network outperformed the other models in the evaluations (e.g. F1 scores & PRAUC).

Zacharis [87] uses an Artificial Neural Network (ANN) to predict academic performance in blended learning. The data applied originated from a Moodle server (LMS) and is used to predict a student's success in a course. The data contains 265 students with 6 variables (e.g. grades & quiz efforts). Correlation (Pearson) and statistics (e.g. median and standard deviation) are explored to evaluate relationships (many of them have strong relationships to grades). 4 variables (messages, content contribution, file viewed, and quiz effects) are used with the ANN and the data is randomly assigned to training (60%), testing (20%), and holdout (20%) subsets. The ANN contains 4 input nodes, 1 hidden layer (4 nodes) & 2 output nodes. The finding shows their ANN to produce accuracy scores of $95\% - 100\%$.

Lau, Sun, and Yang [53] applies a Neural Network that consists of 11 input variables, 2 hidden layers (30 neurons), and 1 output layer (applies the backpropagation training rule). The model is used to predict academic performance (known as CGPA) for students in a university. A conventional statistical analysis (Pearson, single t-test ANOVA) is applied first to identify factors that have positive relationships to student performance. Afterward, the Neural Network is then applied. The data contains about 1000 undergraduate students (275 females, 801 males), from $2011 - 2013$ and contains variables of their socio-economic background (e.g. gender, location). The findings show (using AUC), the accuracy of the Neural Network returned 84%. The analysis explores the error loss with PCA (5 components)/ histogram (per iteration) and a graph of the ROC curve. The paper visualises future the use of statistical analysis and then applying (selected) variables to a Neural Network.

Waheed et al. [80] explored a Deep Learning model with Virtual Learning Environments (VLE) big data to predict student academic performance. The study showcases the dropouts with unique handcrafted features (interaction behaviour like clicks, and access to previous lectures). The dataset contains around 32000 students of 9 months from $2014 - 2015$. The number of outcomes is destination, pass, and fail. The Deep Learning model contains 3 hidden layers ($20 - 30$ neurons/nodes) and a batch size of $32 - 64$ depending on the combination of outcomes (e.g. distinction & fail, pass & fail). The findings discover the Neural Network performed the best accuracy at around 84% compared to 2 other Supervised Learning models (SVM

& Logistic Regression) using cross-validation. Students accessing previous lectures tend to perform better than those who do not. Then they repeat the experiment on a quarterly basis (e.g. Q1) with just the Deep Learning model and show the prediction accuracy remained relatively similar.

Ding et al. [27] explore a feature learning method using the Unsupervised Learning approach to help improve the predictor performance in Massive Open Online Courses (MOOC). Unsupervised Learning supports the decision for selecting variables from a dataset. The feature selection approach learns the compact representation of each variable with a high level of redundancy. The learning patterns are derived from a modified auto-encoder (AE) and merged with the long short-term memory (LSTM) network to evaluate the feature list. They call it an LSTM-AE.

The data is collected from an online course of twelve chapters called *Introduction to Computing with Java*, which lead by the Hong Kong University of Science and Technology, from June 2014 to September 2014. The data contains 4991 students that completed the course ($\frac{1}{9}$ of the total number enrolled). 20 variables are used for the experiment and each student is given a label per chapter.

They first compared their novel auto-encoder against several models such as Logistic Regression and a Neural Network. The Mean Squared Error scores are lower with the novel encoder in all chapters. Two further auto-encoders are created called Symmetric VAE and Asymmetric VAE for comparison.

The PCA and t-SNE analysis show relation to chapters 3, 7, and 11 chapters as k. The analysis shows as k is increased, the grades are closely together compared to low chapters.

They then compared the LSTM-AE against a Neural Network and models using Symmetric VAE, and Asymmetric VAE as auto-encoders. The findings show in all tests (Cross-Validation and Mean Squared Errors) the LSTM-AE performs up to 17% better compared to the supervised Neural Network models. The LSTM-AE outperforms its competitors in all chapters.

Others

Helal et al. [39] explores Classification Supervised Learning models to predict the academic performance of their first-year domestic undergraduate students, from an Australian university. The data is from a learning management system (LMS) and enrollment data. In total, there are over 25 variables and around 2600 (7000 LMS in total and 2600 in enrollment data in total) students (that exist in both datasets). Enrollment data includes age, gender & parental education. LMS data includes forum posts, viewing lesson activities, and visiting the course home page. The experiment involves using the data combined, separately, cross-validation with the enrollment dataset, and cross-validation with activity data (from LMS). Results show that the accuracy performance (and other benchmarks) showed a range of $78\% - 86\%$. Also, the results don't show a huge difference between the data filtering (but the data combined performed the best), the best predictor is the Naive Bayes.

Xu, Moon, and Schaar [85] predicts student performances (college students in the USA) that tackle 3 new challenges: 1) student differs in terms of background; 2) courses are not equally informative for accurate predictions; 3) student progression should be incorporated into the prediction. It presents a novel machine-learning method for predicting student performances. The method contains 2 unique features: 1) a bi-layered structure that contains multiple base predictors; 2) using latent factor models and matrix factorization to discover course relevance. The uniqueness is tied to how weights are updated each quarterly, where the weights are comprised

of an ensemble progressive predictor. The data is from UCLA and it contains 1169 enrolled undergraduates from the Aerospace & Mechanical Engineering niche. The paper explores correlation statistics between high-school GPA & final GPA (shown positive relationships). The findings show error loss (using Mean Squared Error) to be lower per quarterly compared to other Supervised Learning models (e.g. Logistic Regression & KNN).

Hussain and Khan [46] applies Cross-Validation (10 folds) methods alongside data mining to predict student academic performance. It applies Supervised Learning models to derive their outcomes. The data originally comes from an academic management system. That data consists of intermediate and secondary students, between $9 - 12$ classes. The data contains many variables including demographic data (e.g. religion), and 90000 students. It first applies a Genetic Algorithm (GA) feature selection optimisation algorithm (a unique method). Decision Tree & KNN (Regression & Classification versions) are applied with 2 different variables as targets (quantitative marks and qualitative grade). For both models, an experiment with and without the feature selection is applied. For the Classification, the findings show the accuracy to be 94%/96% for Decision Tree, & 86%/90% for KNN (with/without feature selection). For Regression, the findings show the Root MSE to be 5.34/8.23 for Decision Tree, & 24.31/27.66 for KNN (with/without feature selection). Applying the feature selection helped improve the benchmark performance in all cases. Decision Tree performs better than KNN in all cases. The future work from here is to apply the data with Deep Learning methods and aims to produce a framework that returns the highest accuracy percentage possible.

Badr et al. [10] involves the development of a Classification Based on Association (CBA) model to help predict the possibility of a student dropping out of a programming course using rule-based metrics. The data is originally from the Mathematics Department in the College of Sciences, from 2008–2014. The original data needed to be translated and it ended with 203 rows and 57 variables. They explore feature selection to choose the best variables using a coefficient correlation. In total, 4 variables are chosen that are related to English and Mathematics. The model is called CARM which is computer software. Two experiments are conducted based on subjects (1: English & Mathematics; 2: English). Once the data is inputted, the software evaluates the benchmark based on Classification Based on Association (CBA) rules. The output returns the prediction of a student. 17 students are applied to the results and the finding showed a prediction accuracy of around 52% in both experiments.

Xenos [84] applies a Bayesian network to determine student behaviour in distance education using computers (logs). The aim of the study is to develop a method of modelling the educational experience of its designer using past data. The data is given by the Informatics course of the School of Sciences and Technology of the Hellenic Open University (HOU). With a first-year student record of 800 and their computer behaviour is observed with the Bayesian networks. The findings show that modelling with Bayesian Network delivers the direction of decision-making (e.g. what factors to consider). It shows many positive and negative relationships between their behaviour. The advantage of using the Bayesian networks is the modularity which motivates readers to use this model.

In summary, the first discussed decision-making model shows positive uses with student performance predictions, which may be due to its straightforward specification. There is a balanced ratio here but this is not the case when the restriction is lifted (more heavily toward Classification problems). As expected, Regression models can be used for statistical analysis (correlation).

Applying only Feed Forward Neural Networks is popular but less so compared to statistical model families. The common architecture includes applying $1-2$ hidden layers with an input & output layer, the number of neurons/nodes varies. The datasets are mainly student records from institutions or an educational platform (e.g. Moodle). The methodologies tend to be very similar and the Accuracy benchmark metric is used the most to define the best & worst predictors. The papers are heavily lean toward Classification problems.

Even though Regression models (Linear/Logistic) are used, there is a wide range of alternative options available such as Random Forest, Naive Bayes & SVM. Of course, the model choices are dependent on the circumstance(s) and/or problem(s). This includes the data specification (e.g. data type, variables), and benchmark metric.

2.3.3 Multiple Supervised Learning Models

This section provides papers that contain several Supervised Learning models in one experiment setting for the purpose of predicting student performance. Whenever Neural Network is mentioned, it is referring to Feed Forward Neural Networks.

Adekitan and Noma-Osaghae [2] used admission data to predict students in the first year. The data contains many factors that are academic and non-academic of 1445 students and 5 variables from a university in Nigeria. Numerous statistical analyses are conducted including quantitative analyses (e.g. mean, max, skew), commutative probability, density function & box-plot representation of test scores. Multiple forms of test scores are explored with the same analysis. A number of Classification Supervised Learning models are considered (e.g. Random Forest, Logistics Regression) to evaluate their performance. The data split is 70% for the training set and 30% for the testing set. Results show that Logistic Regression performs the best accuracy of 50%. It also expresses in their findings that cognitive entry requirements are not adequate for first-year students.

Mengash [60] applied admission data to predict student performance of first-year students from a dataset given by a female-only university in Saudi Arabia. The dataset contains 2039 of enrolled students in a Computer Science and Information College from $2016-2019$. 1569 students are from $2016-2017$ and 902 students are from $2017-2018$. There are three types of grades HSGA, SAAT, and GAT which must be passed in order enter to university. The analysis conducted included showing the average performance between the grades given by previous schools with a range of $70\%-95\%$. The analysis also shows a strong correlation between SAAT to other admission data. The models applied are Artificial Neural Network and Classification models (Decision Trees, SVM, Naive Bayes) with an accuracy, PRAUC & F1 score of 80%. The best performer is the Decision Tree, while the worst is the Naive Bayes. They found that prediction for first-year students is possible. The weights of models are compared before and after enrollment with a significant difference between them (around 20%).

Zhao et al. [89] applied prediction methods to students in the Master of Data Science course from Fordham University. The dataset applied is admission data that consist of 17 variables and 826 students. However, the paper does mention that 826 students applied but only 132 enrolled Around 60% got accepted on the course but around 70% did not enrol. The paper experiments only apply 132 students in conclusion. The dataset mainly contains demographic data (e.g. age, marital status, gender) but contains several types of grades such as writing, quantitative, school GPA, etc. The goal of their research is to detect any poor performance occurrence

from the students using analysis and Supervised Learning (Classification) & a Neural Network, as well as evaluate which measures the university should focus on. The analysis includes exploring the variation of each attribute with the categorical variables. This is represented as pie charts. Also, they discovered many findings from their data. For example, students are mainly males, most students are foreign nationals and they are from older generations. In addition, other forms of analysis include exploring the average grades between top 20%, mediocre 60%, and bottom 20% performers. The difference in grades shows a slight increase between the groups. The method of conducting experiments with the decision-making models includes splitting the data into 10 folds (Cross Validation) and exploring the predictions between the two groups. One is the bottom 20% against the rest and the top 20% against the rest. The Random Forest is the best performer (in general). The range of accuracy is between $60\% - 100\%$. The results also present the best predictors (attributes & variables) from Classification models. The findings of the research helped the institution to focus on rubric measures and financial aid.

Masangu, Jadhav, and Ajoodha [59] applies Classification Supervised Learning models (Decision Tree and Perceptron Classification and SVM, Logistic Regression and Random Forest) to predict student (grade) performance and evaluate the benchmark accuracy. The data is from KAGGLE which consists of demographic data (e.g. gender, nationality), it contains 480 students (e.g. 305 males and 175 females), from two academic semesters (semester 1: 245, semester 2: 235), and 16 variables (e.g. gender, nationality). The work contains prediction with several classes $(2 - 4)$. The grade is label-encoded which is the target. The findings show that SVM produced the best prediction accuracy of 70.8%, and the range of accuracy is $47\% - 71\%$ (rounded to the whole number). When repeating the same experiment with just class variables, the absences have a strong relationship to class variables. The paper discusses that future work involves collecting more samples & variables to predict more student grades.

Bujang, Selamat, and Krejcar [20] provides a predictive method for final-year students using Supervised Learning models. The data contains 12 qualitative variables, 489 students, between $2016 - 2019$ academic cycle, and originates from a Computer System Architecture course in an institution. The pre-processing phase involves collecting a set of variables and applying a feature selection technique built-in WEKA (Best-First search method) to collect the feature combination. It became 5 variables in total for the experiment after pre-processing. The Supervised Learning models are J48, Random Forest, SVM, & Linear Regression. The findings show with Root MSE converted to prediction accuracy is around $85.9\% - 99.8\%$. The J48 performed the best and the SVM performed the worst. They provide several predictive analytics of the benchmarks which are represented in bar charts and histograms. The future of this research shows collecting more data and producing predictions to improve student performance.

Mueen, Zafar, and Manzoor [61] applies data mining techniques to predict student performance from two undergraduate courses. The authors do not mention the number of samples and variables but it does mention the students are from $2014 - 2015$, the variables are limited and they use WEKA for data mining. The chosen decision-making models include Naive Bayes, Neural Network, and Decision Tree. The results show a performance of 86% (maximum) accuracy with a range of $80\% - 86\%$. They perform 2 tests with different combinations of variables: 1) with all variables; 2) best-suited variables. The future work they hope to accomplish is to explore larger datasets. They also recommend instructors interact with students via forums.

Aggarwal, Mittal, and Bali [5] uses non-academic parameters for predicting student performances. The paper compares the performance between two model groups that use academic data (e.g. program, age, entrance year test) and all data (e.g. gender, household income, year of birth). The dataset contains 6807 students with 20 variables. Each model group applies 8 classification models (e.g. Logistic Regression, Bagging, AdaBoost, Neural Network). The findings show around 78% F1 Score with just academic data and around 92% with all data. The outcome shows the motivation for using student data, not just academic data (to evaluate student performance).

Li, Lynch, and Barnes [55] perform feature selection & Regression decision-making models (e.g. SVM, Decision Trees) to evaluate student's final performance using log data in 2013 (e.g. test scores, response) from the first 6 weeks. It's worth noting that Cross-Validation Leave-One-Out is applied. The feature selection used is the Feature Variance which is specifically designed for numerical variables. The data consisted of 249 students from the Department of Computer Science at North Carolina State University. The students are allocated to one module: Discrete Mathematics for Computer Scientists. There are analyses done against the grade showing not an even distribution (imbalanced data). The findings show the best accuracy is 51% (SVM) and the worst is 24% (Naive Bayes). The paper also compared the benchmark with and without normalised inputs (Z-Score) which showed relatively similar results.

Francis and Babu [31] explores the potential of predicting academic performance for students using hybrid data mining approaches. It exposes a new framework that uses both classification & clustering techniques. The data originates from an institution that is not named for reasons of confidentiality. The process includes data preprocessing, collecting features, training the data with classification models (SVM, Naive Bayes, Decision Tree, and Neural Network classifiers) to collect the features that returned the best prediction accuracy, and then training those variables with a clustering technique (K-Means) and collects the common grade (from the clustered group), this is the academic prediction. The results show the framework produced an accuracy of $40\% - 75\%$. It also explores the benchmarks between many combinations of features and shows the data with academic, behaviour & parental details produced the best accuracy (around 75%). The paper claims it can be expanded in the future using a larger quantity of features.

Hussain et al. [45] apply data mining techniques to identify academic performance using the WEKA machine learning software. The data contains 300 students with 24 variables (e.g. family size, gender, study hours) from 3 institutions. Feature Selection is first applied to collect good combinations of features using correlation-based and rank-based methods (built-in WEKA). Using Classification models (Decision Tree, BayesNet, Random Forest, PART), the findings show the accuracy range is $65\% - 99\%$, where the Random Forest performed the best (99%). The paper claims the outcome may find the kind of courses adapted to every cluster that holds similar student characteristics. It can also deliver multiple summary reports & teaching routes.

Dronyuk, Verhun, and Benova [28] explores non-academic factors to evaluate a student's academic potential of being qualified to conduct a software engineering test. The data is a total of 101 from past student CVs with 5 variables. A total of 3 Classification Supervised Learning models are explored (Logistic Regression, Naive Bayes & Random Forest). Using cross-validation of 10 folds, the accuracy shows around 76%. Their finding discovered that past student CVs can be used to determine whether a student is qualified for the software engineering test.

Zhang and Rangwala [88] aim is to use a time-series approach to detect early signs of dropouts using an Iterative Logistic Regression and other Classification models in future semesters. It is reported that low retention rates and a high number of dropouts. With 41% of dropouts occurring in the United States. The dataset is given by George Mason University from Fall 2009 to Spring 2016 focussing on first-time entry students. The data contains 13643 of students and 11 variables such as cohort, age, high school GPA, etc. The experiment includes using a set of Classification models (e.g. Naive Bayes, KNN), and one additional model called Iterative Logistic Regression. The key difference between the Iterative Logistic Regression to a standard Logistics Regression is as semesters are being processed, all the previous semester's dropout predictions are appended as a new feature for the next semester. The time series is known as Anchored Walk Forward. The analysis applied includes comparing the dropout rates from Fall $2009 - 2013$ and benchmark scores. Evaluation using metrics such as Accuracy and PRAUC show using an Iterative Logistic Regression improved as semesters processed (most cases). The benchmark scores have shown a range of $52\% - 98\%$. The Iterative Logistic Regression showed higher rates of True Positives. It provides an indication of previous dropouts do influence dropouts in the following semester.

He et al. [36] explores predictive models to identify at-risk students of not complete their online course (MOOC) on weekly basis. 6 Classification models are applied (regularised Logistic Regression, SVM (LibSVM), Random Forest, Decision Tree (J48), Naive Bayes, and BayesNet). In addition, two novel variants of Logistic Regression are used: 1) Sequentially Smoothed LR (LR-SEQ) that minimises the regularisation; 2) Simultaneously Smoothed LR (LR-SIM) that correlates prediction (early and later weeks) that influences each other.

The dataset contains 1117 students that completed the course. It includes 778 completed assignments, over 100000 students enrolled, 110 recorded videos for a period of 9 months, and 7 variables that exist (e.g. week, percentage of lectures viewed) in the dataset. The findings show the LR-SIM Logistic Regression managed to produce the best performance using AUC of around 80% in the first week. The ranges of values are $78\% - 80\%$. In other weeks $(2 - 9)$ the ranges are $86\% - 99.5\%$. The novel algorithm is the best performer and the additional clause correlates all weeks making it perform better than its competitors. The paper expresses future work including course instructors (MOOC) applying the novel model.

Trakunphutthirak, Cheung, and Lee [76] predict at-risk of failing (below 60% GPA) in their studies with log data (web-browsing & internet access activity). The log files contain 294 students (live-off campus). The data contains 24 categories of web browsing (e.g. games and streaming) and 147 internet access activities (e.g. Google-base). The benchmark performance between these datasets is compared and it collects the top 24 high correlated attributes from both datasets. The benchmark weight ratio (e.g. $10 : 90\%$, $40\% : 60\%$) is compared between both datasets to evaluate which type is more predictable.

The experiment includes initialising 4 Supervised Learning models (Logistic Regression, Naive Bayes, Neural Network, and Random Forest) and producing benchmarks with the F1-Score & PRAUC. The data is explored with many combinations of weights to evaluate the best correlation (using Pearson). Afterward, the best benchmark from the models is used to compare the 9 different weights ratio. The findings discovered the Random Forest performed the best (77%). The ranges of prediction accuracy are $40\% - 78\%$. Also, data from internet access manages to detect at-risk students better than web browsing. The future work here is to repeat the same experiment with richer log data over a longer time period.

Greatorex-Voith and Anand [34] uses a Data-Driven Framework to detect the risk of students not completing on time. Their framework is a portable solution that is applicable to institutions. This means other methods are on an institutional basis which can provide the same or different results. The aim is to detect high school dropouts and promises flexible pipelines that do not influence biased outcomes on an institutional basis (each institution). The framework includes a feature grouping method, where they group features together that have good relations to each other using Factor Analysis which is then applied to decision-making models. The data applied originated from many schools across the USA. The data contains many variables of each student such as enrollment, grades, and absences. The models are four Classification models: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machines (SVM) as their specifications are suited to the problem. Only students from the top 10% of the dataset are separated by each grade. The Precision score shown as grades increased $(6 - 11)$, the score became better. The comparison is evaluated with a baseline, a rate computed for not-on-time graduation per grade. Their findings evaluated a portable framework with a range benchmark score of $55\% - 80\%$.

Aguiar et al. [6] is a large study of identifying patterns between students that dropped out and not in high school. The data is given by an institution of 11000 students that are expected to graduate in 2013. The students are from 6^{th} to 12^{th} grade. The data contains 15 variables with a mixture of demographic and grades including D.O.B and gender. Most variables are quantitative (numerical) data types with minor variables as qualitative (categorical). The study is categorised into three sections: who, when, and why. In each section, numerous decision-making models and evaluated with metrics.

In the *who* section, they explored the type of demographic likely to drop out. The Logistic Regression, Random Forest, and unique model called Partner's model are used with a Cross-Validation of 10 folds. As more data is fed in, the benchmarks improved for each grade year except for the Partner's model. With accuracy ranges of $40\% - 80\%$. Also, the research explores the relationship between student mobility vs high-risk rates, and student GPAs (binned) vs high-risk rates. The analysis shows that students with mobility and/or low GPA grades are more likely to drop out.

In the *when* section, they explored the time occurrence of dropping out. The investigation included exploring a variable known as *time to off-track* which determines if the student retained a dropout. Using several models such as Cox Regression & Logistic Regression, they explored the Pearson correlation relationship between time to off-track vs risk scores and academic performance with time to off-track (as target). The correlation shows not a high correlation between time to off-track and risk scores from grade 6 to grade 10. Repeating the experiment to grade 11 shows a higher correlation with the risk scores. The outcomes show that the method is not fully reliable. The benchmarks from the models do show that grades 10^{th} and above are more accurate than lower grade years. The Ordinal Regression Tree is the best predictor for each grade year, ranging from $35\% - 80\%$ accuracy.

In the *why* section, they explored the possible reasons why one drops out. It is shown in the other sections that several variables contribute to dropouts and hard to find one precise answer. As a result, they created a dashboard to report student risk scores such as absences and past grades. The dashboard (web-based) uses these details to help learn the risk of dropping on future occasions. This provides them with more understanding of the topic.

Ahadi et al. [7] explores Machine Learning methods to identify which students

need assistance automatically. This is done by detecting occasions of high and low-performing students. The data is from two semesters of an introductory programming course at the University of Helsinki. Support is available in the computer labs and 20-30 hours from the instructor and teaching assistants. The type of students is from low and high educators and finance. The assignment performance is recorded in software called *Test My Score* where all behaviour is recorded when accessing the work on the computer. Examples include the pressed buttons and duration of completion. A total of 296 students are conducted in this experiment. The exploration includes comparing two questions.

The first question is using a method proposed by Jadud [48]. Jadud quantify a student's tendency to make errors, also known as *error quotient*. The error quotient correlation and average score from programming assignments are low. But the correlation between error quotient & grade is high. Also another correlation Watson, Li, and Godwin [82] is used and compared with Jadud. Watson, Li, and Godwin is similar to Jadud but they proposed an *error quotient* improvement that deliver better correlation. The improvement is considering the amount of time taking to complete a task. The evaluation of the question showed the outcome to be complicated as the factors are different. For example, the data in this study are shorter periods and the assignment tasks in this study are more compared to Watson, Li, and Godwin and Jadud.

The second question is about Machine Learning. The variables are first extracted from student records such as age and gender and many source code snapshot attributes which results in 52 variables. Feature selection (using Information Gain) is applied to reduce the number of samples and improve performance. The reduced version contains 13 variables. The models applied are Classifications such as Bayesian Classifier & Random Forest. The data applies Cross-Validation of 10 folds with 66% as the training set and 33% as the testing set. The accuracy benchmark shows the Random Forest as the best predictor, ranging from $86\% - 90\%$. The average accuracy is around 80% from all models.

In summary, the normality of applying several models for predicting student performance in one experiment setting is common. The purpose is to identify the best and worst benchmark performance. Generally, collecting a chuck of papers provides an idea of which model is best suited to the problem. The common Supervised Learning is Random Forest and the common Feed Forward Neural Network architecture is 1 hidden layer (node amount is random). It also shows that only one (Feed Forward) Neural Network is applied to each experiment.

2.3.4 Insights from Existing Surveys

In addition to the investigation of relevant AI technologies to predict student performance, it is always important to examine literature review papers. Doing so results in more exploration of papers to determine the challenges, trends, and state-of-the-art research. These collect sufficient papers related to AI & Education in a timespan (e.g. 10 years) and then explore those commenced work. These papers can provide guidance on which search strategies are worthwhile. Whenever Neural Network is mentioned, it is referring to Feed Forward Neural Networks.

Alwarthan, Aslam, and Khan [8] focuses on exploring the recently published studies (occurred in 2022) for predicting student academic performance and dropouts in a systematic review manner. The total number of papers explored is 44. The exploration involves discovering the latest work of Machine Learning algorithms and/or popular/unpopular variables.

The paper aims to provide the following answers: 1) The most common techniques to predict student performance; 2) The most common group of features (variables); 3) The most common subject in datasets.

In relation to the first aim, the findings show the following: recent (and latest) machine learning models include Gradient Boost & Neural Networks; ensemble learning models are the best predictors (e.g. Random Forest); most popular Supervised Learning models include Logistic Regression and Naive Bayes; clustering techniques (K-Means) are used to group high/low performing students (but not very popular research in general); most Classification outcomes are binary, and most research applies Classification models (clustering models are the least common).

In relation to the second aim, the findings show the following: past researchers show no concrete evidence that admission characteristics have a strong relationship to academic performance (more research is required according to the paper); the common total sample of datasets is between $30 - 300$ (18/44 papers); the common variables (72%) in datasets are university details (e.g. course, attendance); the least common variables (12%) in datasets are social and economic; the overall number of features in datasets are around 25 variables; the data type of academic performance are qualitative (e.g. degree classifications, pass/fail).

Finally, for the third aim, the findings show that most data are not from arts and humanities subjects (5%) and most data are from STEM courses (41%).

Hellas et al. [40] is a literature review study that explores the current work for predicting academic performance. It explores the type of research done in the past and verifies if the research is increasing. There are two main focuses: 1) the current state-of-art in predicting student performances; 2) the quality of work used to evaluate prediction performance. The first question is divided into the following parts: a) determine relatable variables to student performance; b) the method of performance; c) how performance is classed; d) the type of variables and method combinations used to predict academic performance. The number of papers reviewed is 357, from $2010 - 2018$ (the majority are from 2017), and the collection occurred in 2018.

To answer question $1a$, the most popular variables for student performance is course grade value (e.g. GPA) or grades that are retrieved from a list of grades (e.g. A-F). Some papers (around 12%) they reviewed are unable to identify the precise student performance outcome. To answer question $1b$, the common factors include course performance and pre-course performance (contributes to 25%) such as GPA or high school grade. To answer question $1c$, the common method of identifying performance is using Machine Learning, specifically Linear Modelling (Statistical), Probabilistic Graphical Model (Classification), and Decision Trees (Classification). These 3 models contribute to 40% usage in the collected research papers. To answer question $1d$, the common combination of features and performance are course performance / pre-course performance as features, and course grade/score as the target, which contributes to 21%. Followed by course performance / pre-course performance as features, and dropout as the target, which contributes to 17%.

To answer question 2, the current research requires improvements. The major issue is the outcome, it is not expressed clearly. It also expresses the little reuse of the same data on methods. It is generally related to the ethical side. The paper expresses the ethical concerns (e.g. lack of consent, lack of inclusion) that leads to unethical practices. The answers to the main questions are the findings of this research.

Chaudhry and Kazim [24] published a literature review on AI & Education in the past decades. The paper provides a high-level industrial and academic overview of the topic. The paper mainly talks about the usefulness of AI to assist teachers (reducing workloads), ethical issues, and the impact of COVID-19 in the future. It

is reported the common 4 main subdomains include reducing teacher workloads, contextualised learning for students, changing assessment methods (producing & marking), & intelligent tutoring systems (ITS) to assist students.

Reducing teacher workload has been a massive problem and effective measures need to be applied. AI can be a solution. The pandemic (COVID-19) has shown its uses with online teaching and demanded more useful features. But changing teacher's habit from traditional habits to newer habits is a challenge. New skills need to be taught to use AI. Online teaching has provided some motivation for its benefits.

Contextualised learning for students is mainly down to the learning methods. Different students generally learn in different ways (e.g. audio & practically). Teachers do struggle to teach specific students due to their learning practices. AI on the other hand can be used to adapt the learning methods and then teach them knowledge. AI can help find the learning gaps and discover solutions to overcome teaching barriers. Open Learner Models [21] have been a useful AI tool to facilitate learners, teachers, and parents to enhance learning.

Changing assessment methods is a must to teach children the latest knowledge. Normally, a lot of work uses very traditional materials to assess their intelligence. More modern assessments are required to evaluate each student's capability. They should consider all aspects to judge one's ability. AIAssess is an AI example developed by UCL [58] to assess maths and science using a knowledge model, analytics model, and student model. The knowledge model stores knowledge of the topic, the analytics model learns student interactions, and the student model tracks student progression.

An intelligent tutoring system mimics teachers to provide personalised learning to students (e.g. teacher's voice). Unfortunately, the niche has been struggling to tackle many issues in the past decades. The recent discovery has shown a strong correlation between emotions and learning [11], currently, this has been the focus. Although there are existing services like ASSISTments [38] that show potential, it's not innovative compared to other subdomains.

In addition, there has been an increase in technology businesses that embraces AI in education. Examples of companies include EDUCATE by UCL Institute of Education and European Regional Development Fund, Google & Pearson. In fact, during the pandemic, the CEO of Google delivered motivation to re-image education and released 50 tools to facilitate remote learning.

The ethical is mainly on the trust of AI technologies and showing the drawbacks and legal barriers it can introduce. There should be measurements to support AI being embedded in institutions such as collecting data, its usage with Machine Learning, and outcomes. From an engineering perspective, it can be performance and robustness, bias and discrimination, interpretability and explainability [9, 63], and algorithmic privacy. Abusing this type of power can lead to many legal issues. The situation can be worse when there is a lack of awareness. So, humans would need to monitor the behaviour of these AI systems to avoid wrong outcomes.

It is worth noting that AI outcomes can be wrong, the paper expresses the consequences of wrong outcomes such as directing students to the wrong pathways (e.g. wrong grades). Also, it can deliver a huge psychological impact on learners (e.g. being lonely), discrimination against certain groups (e.g. skin colour), and teaching the wrong syllabus to students (e.g. knowledge level).

The paper justifies the motivation for research on AI in education due to the popularity of online platforms. The pandemic (COVID-19) shows the practice on a

large scale (due to lockdowns). However, there are still ethical barriers for members of education to adopt AI technologies.

Dignum [26] explores the challenges of AI in education on an ethical and legal basis. A well-detailed introduction is given to express AI and its capabilities. The paper expresses AI as autonomy (a system accomplishing its objectives without humans), adaptability (understanding the environment and learning behaviour), and interactivity (behaving as an individual with its own choices). The ART (accountability, responsibility, transparency) principle is used to describe how responsible and trustworthy AI is.

The major concerns are in relation to the privacy and security of personal data. It provides many barriers against the laws of the country. The papers express concerns about students not being self-dependent on technologies and lacking social skills which can result in mental health issues. Self-dependence is associated with students being unlikely to process work individually. The lack of social skills is associated with students communicating with robotic teachers rather than human teachers. The paper then expresses the issue of student data analysis and Machine Learning purposes (conclusions between student groups like ethnicity, and household income). The issue is increasingly getting worse as educational technologies advance.

Finally, the paper provides regulatory framework suggestions for AI and its challenges. The government (or equivalent) must be part of the evolution of AI. They should provide trust and reduction in incidents with many safety layers. Moreover, the human responsibility for incidents with AI should be negotiated to avoid injustice practices. A quote from Organisation for Economic Co-operation and Development OECD [29] *Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programs along the working life, support for those affected by displacement, and access to new opportunities in the labour market*.

The challenges on the other hand are primarily focused on changing people's mindsets and society. Currently, people either adapt or reject AI, and the public should move to the forefront of innovation (depending on the type). Also, technologists need to support the people in the public and the government (or equivalent) to understand the potential. The digital age is currently growing, it is a perfect time for reinvention and creativity.

In summary, literature review papers show a tremendous amount of analysis done with past papers in relation to predicting student performances. This includes exploring ethical and legal issues, data characteristics (e.g. sample, variables, common/uncommon variables), methodologies, scope, machine learning algorithms (e.g. Supervised Learning), and more. There are unrelated topics (e.g. robotic teachers) mentioned but these fall outside the scope of this thesis. These surveys provide additional confirmation on challenges, trends, and state-of-the-art research. Which is great as it increases the chances of contributing to new knowledge.

2.3.5 Other Related Literature

This section provides additional supporting research that is related to AI & Education and, to an extent, student performance. The research is subdivided into three categories: *Recommender System & NLP* and *Variables For Academic Predictions*. As such, this group of papers may not be directly related to the methodologies used in this thesis.

The discussion around important variables for academic performance is strongly relevant, as it may inform processes that are used in this thesis, such as feature selection. The discussion around other AI technologies (e.g. NLP) showcases additional research papers related to AI & Education and, to an extent, student performance.

Recommender System & NLP

Thai-Nghe et al. [74] proposes a novel approach using a recommender system to predict student performance. The recommender system is generally applied to domains such as books and movies. They validate their novel approach by comparing the approach with Regression models. The dataset is collected by Knowledge Discovery and Data Mining (KDD). Two mathematical topics (Algebra 2008-2009 and Bridge to Algebra 2008-2009) are retrieved from KDD. In total, they have up to 23 attributes with millions of instances. The size of data is over 8 Gigabytes. The data is log files of student interactions with computer-aided tutoring systems. The log records their activities, task success, and progress indicators.

The recommender system's user is the student. The recommender system's item is a combination of many factors: *problem hierarchy*, *problem name*, *step name*, *problem view*, and *knowledge components*. the paper claims all options have drawbacks such as imbalanced data. As a result, many combinations are explored to provide an unbiased overview.

The decision-making models are four Logistic Regression models. The pre-processing phase includes deriving the average (set of logged transactions per student) on the target variables. Again, many models with different factors are used: *A* as (Student-Average, Step-Average), *B* as (Student-PV-Average, Step-PV-Average), and *C* as (PG-Average, PN-Average, Student-PG-PV-Average).

The conclusion of performance is decided using a Root Mean Squared Error (RMSE). The findings show the novel recommender systems approach can be useful to predict student performances. The average score is 0.3 for the recommended system and 0.31 for the Logistic Regression.

Lee, Kuo, and Lin [54] applies Collaborative Filtering (CF) to provide course recommendations for students. The aim is to build a CF-model system that overcomes the issues of past CF models. For example, the imbalance distribution of course registration. The data are student records from the National Taiwan University (NTU), from $2008 - 2013$. It contains only students that have 4 years of registration records. There are 13977 students and over 800000 registration records.

Their CF model includes two stages: training and course dependency regularisation. The first is the Bayesian Personal Ranking Matrix Factorization (BPR-MF). The BPR-MF is used to find the two matrices P (feature matrix of students) & Q (feature matrix for courses). Then determine the multiplication that can best recover the matrix R (minimise the error using the AUC). The outcome returns the pairs for students and courses. BPR-MF is chosen as it is suitable for One Class Collaborative Filtering (OCCF).

Afterward, the distribution imbalance issue is next solved. By dividing the data records into Graduates Students and Current Students (horizontally, separated by undergraduate and postgraduates) and Fundamental Courses & Advance Courses (vertically, separated by good/poor performing students). Finally, regularisation is applied to help strengthen the node connections between dependent courses. This helps the recommendation return accurate outcomes.

The experiment involves comparing and executing several types of CF models using the AUC. The finding shows their CF model version tends to be more accurate

compared to past CF models. The score shows a 93% without regularisation and 94% with regularisation.

Petersen and Ostendorf [65] applies n-gram language models, parses, and traditional reading levels with SVM to produce a better method of assessing reading level (Natural Language Processing) for students. The data is a mixture of n-gram LMs, an automatic parser, and traditional methods of readability assessment. It presents the outcome when a teacher or student is searching on the website (or equivalent) for articles at a particular grade level. This is a binary-based problem, so Classification SVM is used. The experiment is conducted on many students from grades $2-5$ (USA). An additional experiment is to handle the problem of negative training data from classes that are not seen in training data. The findings have shown an accuracy of $38\% - 87\%$, unfortunately, the output is mainly showing low accuracy.

Other findings include only a small effect on the overall performance of the detectors, SVM performs better than humans when tested with the Weekly Test plan (with and without the plus), and both Classification and Regression compare favorably to other existing methods. It shows substantial variability in the human annotation of reading levels. Improvements can include more feature extraction & model exploration and using different sizes of out-of-vocabulary (not just 6 as used in this paper).

Variables For Academic Predictions

Kotter et al. [51] applies a Linear Regression model to determine the relationship between the academic performance (from exam called M1) of medical students with their background (age and gender) and pre-test scores. The dataset contains 456 students in freshman year. The experiment is conducted at the start & end of sophomore year. The exploration includes the coefficient correlation between PMSS (T1 & T2) scores to M1 grades, PMSS T1 scores (without age and gender) & PMSS T2 scores (with age and gender). The findings show that 2 and 14 months before M1 provide a positive relationship to M1 grades. Also, age & gender are useful variables to predict grades. Females that are older with stress are more likely to achieve poor academic performance. Future work is claimed to focus more on qualitative approaches to explore the influence of potential confounders.

Thiele et al. [75] perform a study to predict academic performance by examining a student's school and socio-demographic details. The data originates from an institution in the UK (University of Liverpool) and contains many characteristics (e.g. socio-economic, & deprivation). The data contains 5369 students which are derived from UCAS. Students on 4/5 year programs are ignored to avoid biased outcomes. The study applies a Logistic Regression to determine the relationship between their variables and academic performance. The results in general claims no evidence of a direct association between their background factors and academic performance. However, they managed to discover other factors. They discovered students who are Whites are more likely to achieve higher grades than other ethnicities (e.g. Asian, Black). Students who come from an academic background (A Levels), and live in the least deprived areas are more likely to achieve higher grades. The paper claims the data is mainly students who are Whites and below the age of 21. This does mean there is imbalanced data which may alter the outcomes. Further research is required to verify the institution's policies with firm evidence to avoid any discrimination.

Kweon et al. [52] explores the relationship between school environments and student performance. Specifically comparing students that attend campuses with high/low numbers of trees and how they affect student academic performance.

The data contains 219 public schools that contain environment measures (e.g. tree amount), demographic (e.g. ethnicity, number of students, student-teacher ratio), and performance data (e.g. maths grades). The students are grades $2-10$ (USA). The evaluation is based on correlations and regression models. Two aims are explored: 1) academic performance to amount of trees; 2) the relationship between ethnicity/race to trees amount. The findings show there is a relationship between the number of trees and the level of success academically (in schools with more trees). But this is not all the case, in large land, it can provide a negative impact on a student's academics. Students who are Whites generally have strong ties to campuses with higher amounts of trees. The paper expresses it can provide guidance to future projects.

Schwanz et al. [70] explored the relationship between parent relations to college student grades (GPA). The dataset is from two southeastern liberal arts universities that consist of 466 students in total. The dataset contains 281 females and 185 males with an age range of $18-29$ (average age is 20). The data is collected in two batches, each one from a university (313 and 153). There are different variations in the demographics of both universities. The judgment of outcomes is based on the T-score correlation. In their findings, both universities showed a positive correction between parental relations and a student's GPA (grade). They also discovered a negative correlation between parental relations and dropouts.

Pollio, Humphreys, and Eison [67] developed a questionnaire and collected responses of 6000 individuals in Milton, Pollio, and Eison. The individuals are from 23 different institutions in the USA. 4365 are students, 854 are faculty and 584 are parents. The remaining 362 is for business officials. The question included their reaction to high and low grades. It is discovered that parents have strong reactions to both high and low grades. With low grades, parents demanded an explanation, and with high grades, they are proud of the achievement.

Housing [41] is a report conducted in 2019 by the United Kingdom government that showcases the relationship between IMD ranks and several niches including education. According to the reports, education relates 13.5% to the overall IMD rank judgment. In the reports, they discovered individuals from highly deprived areas are less likely to have higher education (undergraduate level) compared to individuals from low-deprived areas. This is because of the lack of support from the community. Unfortunately, the data specification is not mentioned.

Strøm, Falch, and Lujala [72] conducted a large case study on the question of the travel distance relationship between each student's home location and the school location. The research occurred in Norway and is used to evaluate if distance affects graduation propensity. The data is given by National Educational Database in Statistics Norway in the spring of 2002. The data contains their location when they enrolled and matched their details using parental details. It contains over 35 variables of their demographic, distance, and performance. The commuting depends on the student's home. Students living close either walk or cycle to school. Students living far travel by car or public transport (e.g. bus). Variables include travel time, parent marital statuses, GPA, parental highest education, and gender. The analysis included comparing the T-score of distance variables with student performance (grades). It shows distance is a strong relationship between the student's home to the school of registration. Therefore, students living close by are more advantaged to graduating on time with better grades.

Hayward and Hoelscher [35] performed a study on whether entry qualifications have any influence on the success of degree completion (not dropping out). The comparison is between students from an academic background to students from a

vocational background. The data is a combination of third parties such as HESA and UCAS. The UCAS data contains full-time undergraduate students, from $1995-2004$ with many enrollment characteristics such as socioeconomic, entry qualifications, age, gender, etc. The HESA data contains enrolled students showing the rates of completion/non-completion alongside socioeconomics and demographic data. The HESA can also be linked with UCAS using the UCAS identity code. The study showed an increase in vocational courses from $1995-2004$ but the general academic remains relatively similar. The evaluation is conducted using (mainly) three Logistic Regression models, where the difference is the number of variables used for model training. The exploration is based on the intercept scores showing mainly no direct relationship between entry qualification and dropouts. The findings show that vocational courses have a higher risk of dropouts. But it also shows students from a vocational background risk level is based on the institution. Students that have high UCAS tariff scores are the least likely to drop out.

In summary, the papers related to variables demonstrate several relevant factors to student performance. The most popular variables include age, entry qualification, gender, and parental influence. The least popular variables include school, socioeconomics, travel distance, IMD, and POLAR 3/4. Some variables are difficult to collect. It could be restricted to a demographic (e.g. country) or they simply do not store it. For example, IMD, and POLAR 3/4 are variables recognizable in the UK (to the author's knowledge).

Even though several variables are explored in past research, there are additional (or new) variables that can deliver equivalent purposes. Typical examples include health data (e.g. disability) and (more) parental data (e.g. household income). These factors also can affect a student's academic performance. These data can be problematic to collect due to the ethics and law barriers within a country. Having said that, the observation from past papers shows motivation for repeat variable usage in future student performance predictions.

Furthermore, the papers related to other AI technologies show that AI & Education research on academic performance does not exclusively focus on prediction. Despite its importance in the real world, this is not the only important problem. The scope may differ but there are some similarities such as using student data to evaluate their outcomes and applying some of the AI technologies that are commonly used for academic predictions.

2.4 Summary & Comparison

Applying decision-making models to predict student performance has been occurring in the past decades. Since the 1990s, a wide range of AI-related approaches has gradually revolutionised the research field (Masangu, Jadhav, and Ajoodha [59], Aggarwal, Mittal, and Bali [5]). These algorithms apply processes differently to evaluate predictions. The earliest algorithms applied included Classification & Regression models that apply the Sigmoid or line-of-best-fit functions. Examples of more algorithms include AdaBoost, Random Forest, & KNN. In addition, (Feed Forward) Neural Networks also delivered usefulness in the research field. Of course, each model's suitability depends on the given problem. Nowadays, state-of-the-art research generally involves applying Supervised Learning models, including Feed Forward Neural Networks (Zhao et al. [89], Trakunphutthirak, Cheung, and Lee [76]).

In no particular order, the popular decision-making algorithms (models) include Random Forest, Logistic/Linear Regression, Naive Bayes, KNN, Decision Tree, SVM, & Feed Forward Neural Networks. Many papers produce evaluations that compare several of these models. Feed Forward Neural Network architecture applications, in particular, have shown capability in accurate predictions and/or execution speed. There are some cases where a mixture of recent & old Supervised Learning models is applied in one experiment but it tends to be heavily biased to one side. Normally, it is heavily biased toward older models than recent ones (e.g. Extra Trees) (Rodríguez-Hernández et al. [69]).

The datasets used are in the form of student records and contain mainly academic & demographic data (Mengash [60], Hayward and Hoelscher [35]). Typical examples include attendance, D.O.B, ethnicity, gender, and grades. The most important factor is academic performance as a student's career opportunities afterward are dependent on their studies. Also, most variables are more likely to be related to qualitative than quantitative data types. Demographic data is likely to be qualitative while academic data can be both qualitative and quantitative. Variables in these datasets are more likely to contain qualitative variables than quantitative ones. It is worth noting that student records contain personal data.

The process of predicting student performance includes collecting a dataset with student details (Greatorex-Voith and Anand [34], Li, Lynch, and Barnes [55]) which are characteristics and academic performance variables. Data is pre-processed such as adding/removing variables and label encoding. Pre-processed data is then fed into decision-making algorithms (models) to predict academic performance.

Sometimes, there are additional steps applied to the process such as feature selection, time series, and/or Cross-Validation. These additional steps could be useful to identify hidden patterns. Motivation for feature selection is due to the findings of variables that have a strong correlation to each other (Greatorex-Voith and Anand [34], Xu, Moon, and Schaar [85]). With these combinations, it is likely to improve the prediction accuracy than without (Hussain and Khan [46]). The algorithms used in past works are either applying built-in functions or applying a novel algorithm/approach (Hussain et al. [45]).

Time series are useful if the problem involves evaluating performance in a time cycle (e.g. semester, yearly). The common cycle shows to be on a semester basis, this may be because the full course/module is semester-long. To conduct a time series, the data must include the correct variable(s). Most data (that mention the variables) do not provide this type of detail or do not include sufficient samples for unbiased outcomes. The Supervised Learning models seem to be used with time series problems (Zhang and Rangwala [88]). Cross-Validation is great to support unbiased outcomes but it is not common (Zhao et al. [89], Yang et al. [86]).

Furthermore, sometimes predicting student performance involves identifying the chances of not completing their studies (dropping out) (Aguiar et al. [6], He et al. [36]). If a student dropped out (for whatever reason) that student is technically failed (and vice versa). These predictions are binary Classification problems (a pass/fail) academic outcome.

Classification metrics identify the number of correct/incorrect occasions between predicted and truth. Regression metrics compare the error losses between predicted and truth. There are numerous types of metrics one can explore which have their uniqueness. However, the purpose is the same, they are used to evaluate how well decision-making algorithms perform. For Classification, these include AUC, Accuracy, F1, & PRAUC (Helal et al. [39]). For Regression, these include MAE, & MSE

(Thai-Nghe et al. [74]). Fortunately, the development of predicting student performance and metrics are now simpler thanks to programming languages like Python (Scikit-Learn [71]).

The Classification problems are more commonly explored compared to Regression problems. This is because most academic performances are presented as qualitative rather than quantitative (Alwarthan, Aslam, and Khan [8], Chaudhry and Kazim [24]). Theoretically, Classification models definitely perform better prediction due to the number of options available. This results in a lesser chance of overfitting/underfitting and better optimisation. Classification problems can be binary-oriented (e.g. pass/fail) or multi-oriented (e.g. A, B, C). The multi-oriented case is an instance of a Multi-Classification problem and it is compatible with nearly all decision-making models (e.g. Logistic Regression). Going even deeper, the analysis shows that Multi-Classification problems are more common (Lau, Sun, and Yang [53]).

In addition, there are other past works that do not necessarily focus on predicting student performance. These include exploring the coefficient/correlation relationship between factors and academic performance via machine learning algorithms (Kotter et al. [51], Thiele et al. [75]), predicting student behaviour on MOOC (Jiang et al. [49]), and recommending courses with a Recommender system (Lee, Kuo, and Lin [54]). Although the research scope differs, they do share some similarities such as the data category (educational data) and target audience (students).

The trend of predicting student performance accurately with Machine Learning algorithms is rising. This includes exploring the potential of usage & presenting new methods (if possible) to evolve the research field. Most Machine Learning based research applies the same algorithms such as Random Forest, Naive Bayes, & Logistic Regression. (Feed Forward) Neural Networks have been applied to the problem in recent years mainly (Bilal et al. [18], He et al. [36]). Although there are more algorithms proposed recently, they have not gained the same traction. One reason may be the lack of significant improvement in results compared to established approaches. Unsupervised Learning has been used for predicting performance in rare cases. It has also been used as a process step, typically in the pre-processing stages (Ding et al. [27]).

The most commonly encountered data sample amount is 300 or below with most variables related to course/pre-course performance variables such as reading score (Alwarthan, Aslam, and Khan [8]). Data generally originates from educational institutions (e.g. universities) that focus on providing knowledge to individuals. The vast majority of variables are more characteristics than anything else.

The challenges are focused on predicting accurately, in order to avoid leading students in the wrong direction during their education life (e.g. not assigning poor student performance to a high-achieving student, not using insufficient quality data). Other challenges include the ethics and law of collecting data as it includes several personal information. This barrier leads to chances of unreliable outcomes due to the poor quality of data collected. The data must be of sufficient quality to deliver usefulness to institutions and the research field. These challenges have been reported in several studies that review AI & Education in the past decades (Hellas et al. [40], Chaudhry and Kazim [24], Dignum [26]).

Table 2.3 shows a comparison of all papers reviewed in this chapter. The *Paper* column showcases the papers, *AI* (*NN*: Neural Network, *SL*: Supervised Learning, *RS*: Recommender System, *NLP*: Natural Language Processing) provides the type of technologies in each paper, *Model Amount* records the number of models in a paper

(used in their experiments), *Hidden Layers* determines the number of hidden layers, *Rich Data Analysis* determines whether the data is explored with many forms of analysis (and not just generic data specification and/or statistics), *Performance Value* determines the output format produced in their methodologies, *Predict/Explore* (*P*: Predict, *E*: Explore) determines whether the paper is focused on predicting student performance (with AI) or exploring the niche (e.g. analysing case studies), *Imbalanced Data Effect* verifies if a paper introduces a method to reduce the probability of imbalanced data (which is represented as a novel algorithm). The answers can be *Y* (Yes), *N* (No), or - (not applicable).

TABLE 2.3: Display a comparison using all papers.

Paper	AI	Model Amount	Hidden Layers	Rich Data Analysis	Performance Value	Predict/Explore	Imbalanced Data Effect
Utzman, Riddle, and Jewell [78]	SL	1	-	N	Y	P	N
Yang et al. [86]	SL	1	-	N	Y	P	N
Pereira [64]	SL	1	-	N	Y	P	N
Pyke and Sheridan [68]	SL	1	-	N	Y	P	N
Jiang et al. [49]	SL	2	-	N	Y	P	N
Jaber et al. [47]	SL	2	-	N	Y	P	N
Bilal et al. [18]	NN	1	1	N	Y	P	N
Rodríguez-Hernández et al. [69]	SL; NN	7	1	N	Y	P	N
Zacharis [87]	NN	1	1	N	Y	P	N
Lau, Sun, and Yang [53]	NN	1	2	N	Y	P	N
Waheed et al. [80]	NN	1	3	N	Y	P	N
Ding et al. [27]	SL; NN	5	1	N	Y	P	N
Helal et al. [39]	SL	4	-	N	Y	P	N
Xu, Moon, and Schaar [85]	SL	5	-	N	Y	P	N
Hussain and Khan [46]	SL	2	-	N	Y	P	N
Badr et al. [10]	SL	1	-	N	Y	P	N
Xenos [84]	SL	1	-	N	N	P	N
Adekitan and Noma-Osaghae [2]	SL	6	-	N	Y	P	N
Mengash [60]	SL; NN	4	1	N	Y	P	N
Zhao et al. [89]	SL; NN	8	1	N	Y	P	N
Masangu, Jadhav, and Ajoodha [59]	SL	5	-	N	Y	P	N
Bujang, Selamat, and Krejcar [20]	SL	4	-	N	Y	P	N
Mueen, Zafar, and Manzoor [61]	SL; NN	3	1	N	Y	P	N
Aggarwal, Mittal, and Bali [5]	SL; NN	8	-	N	Y	P	N
Li, Lynch, and Barnes [55]	SL	7	-	N	Y	P	N
Francis and Babu [31]	SL; NN	4	1	N	Y	P	N
Hussain et al. [45]	SL	4	-	N	Y	P	N
Dronyuk, Verhun, and Benova [28]	SL	3	-	N	N	E	-
Zhang and Rangwala [88]	SL	6	-	N	Y	P	N
He et al. [36]	SL	6	-	N	Y	P	N
Trakunphutthirak, Cheung, and Lee [76]	SL; NN	4	1	N	Y	P	N
Greatorex-Voith and Anand [34]	SL	4	-	N	Y	P	N
Aguiar et al. [6]	SL	4	-	N	Y	P	N
Ahadi et al. [7]	SL	9	-	N	Y	P	N
Thai-Nghe et al. [74]	RS; SL	2	-	N	Y	P	N
Lee, Kuo, and Lin [54]	RS	1	-	Y	Y	P	-
Petersen and Ostendorf [65]	SL; NLP	1	-	N	Y	P	-
Alwarthan, Aslam, and Khan [8]	-	-	-	-	N	E	-
Hellas et al. [40]	-	-	-	-	N	E	-
Chaudhry and Kazim [24]	-	-	-	-	N	E	-
Dignum [26]	-	-	-	-	N	E	-
Kotter et al. [51]	SL	1	-	N	N	E	-
Thiele et al. [75]	SL	1	-	N	N	E	-
Kweon et al. [52]	SL	1	-	N	N	E	-
Schwanz et al. [70]	-	-	-	N	N	E	-
Pollio, Humphreys, and Eison [67]	-	-	-	N	N	E	-
Housing [41]	-	-	-	N	N	E	-
Strøm, Falch, and Lujala [72]	-	-	-	N	N	E	-
Hayward and Hoelscher [35]	SL	1	-	N	N	E	-

The focus of the novel framework for predicting student performance that is the focus of this thesis is to address the following research gaps that were identified through the systematic literature review presented in this chapter:

- Performance range: the current work uses a performance value as output to evaluate student performance. Even though it may resemble the preciseness to student performance, it does not explore the performance range where there is a high/low boundary which can be described as a tolerance. With the given outcome, it can improve the prediction accuracy, it also can be more suitable to predict performance as it considers the struggles of being a student. In this thesis, the student performance output is a performance range that provides their minimum & maximum academic performance.
- Lack of effect on reducing imbalanced data: It is clear from past works, not much effect is applied to help reduce imbalanced data. As imbalanced data exist, it can deliver reliability concerns of the outcome. In addition, imbalanced

data also relates to overfitting/underfitting. This thesis explores a potential method of reducing imbalanced data which can deliver confidence in the outcome.

- Limited use of Unsupervised Learning: Unsupervised Learning has rarely been used due to its incompatibility within this research niche. This thesis introduces a model pipeline process that integrates Unsupervised Learning as a pre-processing step. Its purpose involves grouping similar student data before producing predictions.

- Lack of in-depth data analysis: Many papers show a lack of in-depth data analysis. For instance, few researchers explore dimensionality reduction, hypothesis tests (e.g. distribution tests), additional quantitative analysis (e.g. skew, kurtosis), and exploring student groups between performance and other factors. This thesis provides an in-depth analysis of the data and evaluates its strengths and weaknesses.

- Lack of usage of several Feed Forward Neural Networks at once: Research that adopts (Feed Forward) Neural Networks commonly applies one model that generally consists of $1-2$ hidden layers. This raised the question of whether there can be an improvement in results with more hidden layers. To conduct a fair comparison, exploration needs to be done in one setting (experiment with the same data with the same neurons). This thesis explores the use of more than 1 Feed Forward Neural Network with different hidden layers to verify any differences in prediction accuracy (if any).

Chapter 3

Computational Framework

In this chapter, a novel computational framework for higher education is proposed, aiming to predict the academic performance of students. The framework design follows a pipeline process that consists of a unique combination of layers. Each layer must be processed in chronological order. Each layer is placed in a certain position to achieve optimal results. The identified research gaps in existing literature (as discussed in Chapter 2, in Section 2.4) are addressed through these layers. In the remainder of the thesis, the proposed framework is referred to as the *computational framework* or simply *framework*.

Figure 3.1 illustrates the layered architecture of the framework. The chronological steps include:

- Data & Global Pre-Processing
- Data Analysis
- Predict Pre-Processing
- Feature Selection
- Training/Testing Split
- Anchored Training Data
- Model
- Benchmark

As mentioned earlier, the computational framework is designed to address identified research gaps. In particular, the following list associates particular layers with research gaps that are addressed:

- Data Analysis - This produces an in-depth analysis of a given dataset.
- Anchored Training Data - This reduces the probability of imbalanced data before being used with decision-making models.
- Model - This applies a model pipeline process that integrates Unsupervised Learning to assist student performance predictions (as a pre-processing step which is grouping relevant student data before producing predictions), predict student performance with ranges, and uses several Feed Forward Neural Networks with different hidden layers in one setting (experiment).
- Benchmark - This partially fulfills predicting performance range but in this case, it validates if the prediction is correct or not.

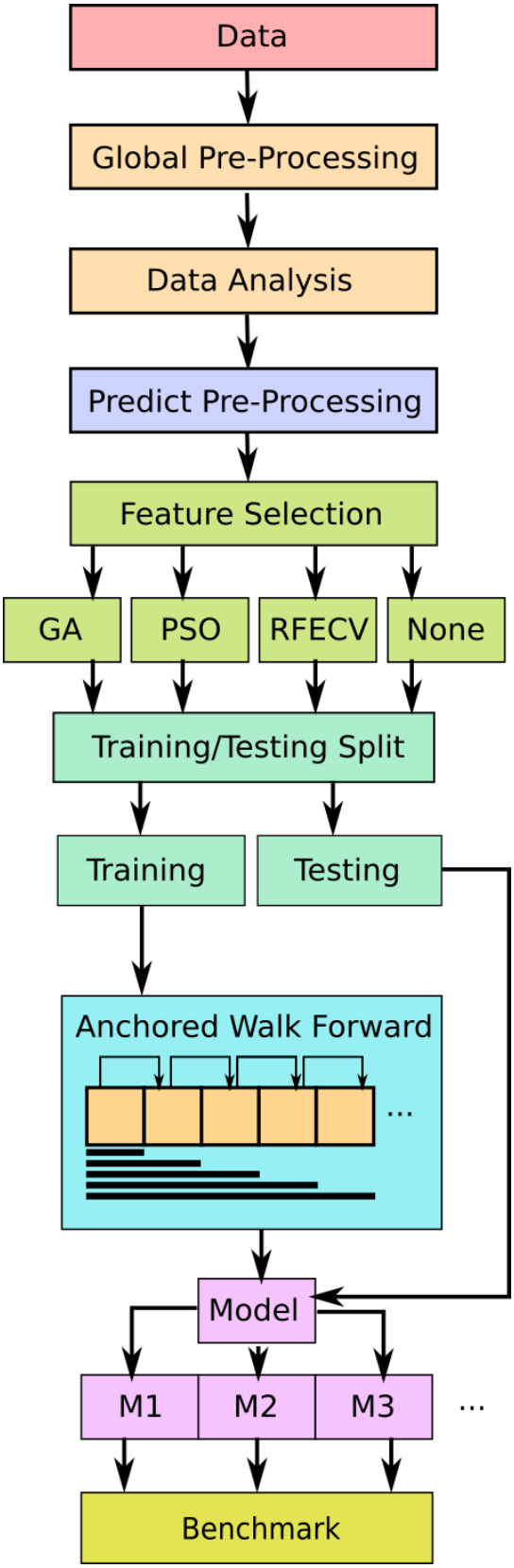

FIGURE 3.1: The computational framework in the form of a layered pipeline. Each block represents a layer that must be executed (top to bottom).

3.1 Data & Global Pre-Processing

Data must be collected that represents student performance alongside their details. In this layer, a generic understanding of the data must be achieved (e.g. attributes, definition, data type). Global Pre-Processing is the initial stage of data cleaning and must be done before going to the next layer.

The following list includes a series of actions that are considered for the Global Pre-Processing (if required):

- Adding Variables (optional): Adding variables refers to producing new variables using algorithms alongside existing variables in the original dataset. Adding variables may also be a replacement for variables that do not provide any purpose (e.g. the variable is discontinued).
- Removing Variables (optional): Removing variables are those that do not provide any purpose to the dataset. For example, an original dataset may represent a completed module which is not ideal for student analysis. One would need to group the data by students (ID); doing so may return some variables not delivering any purpose (because of the change). Therefore, removing them is a suitable option.
- Group data by students: The data must represent a single student per row; this is important given that this framework is aiming to evaluate each particular student's performance. If a dataset does not fulfill this format then one must represent each row as each student by grouping the data with their ID number or equivalent. The quantitative variables are computed by the average. The qualitative variables are collected from the first row. In most cases, qualitative variables refer to characteristics, whereas quantitative variables refers to performance (e.g. grades). If this is not the case, it is very likely the data does not deliver any important purpose to a student.
- Remove any rows that contain incomplete data: Oftentimes, data may be unavailable for particular variables and/or students, which doesn't provide any use. Adding an estimation can be an option but there is no confidence that this estimation may be appropriate. To avoid any complications resulting from incomplete or suboptimal data imputation, rows that contain empty data are removed. Also, this framework applies decision-making models which only work if data is not incomplete. It removes rows that contain at least one empty cell.

After this layer, the Data Analysis layer is executed. From this point, a given dataset is collected, and the Global Pre-Processing has been completed (if needed).

3.2 Data Analysis

In this layer, an in-depth analysis of all variables is conducted (when necessary).

This layer addresses the research gap in the existing literature that points towards the lack of in-depth data analysis that includes applying several different analysis methods. This allows identifying particular strengths and weaknesses of individual methods and can provide a more well-rounded analysis of the dataset in question.

Numerous statistical analyses and pattern explorations between (student) groups are conducted to understand the strengths and weaknesses of a given dataset. It is

more complicated and requires specific qualifications (statistics). Examples indicatively include correlation, dimensional reduction, hypothesis tests, and exploring good/bad performing individuals.

Of course, the dataset must be relevant with sufficient samples and variables. The degree of *sufficience* can be subjective and differs depending on circumstances. The analysis is not somewhat required for predicting student performance but it does provide insight into which variables to use and/or which student group to explore.

The following list of tasks is recommended for the layer of in-depth data analysis, provided that the previous step of global pre-processing has been completed as needed:

- Frequencies: Presents the number of attributes of each qualitative variable and statistics (e.g. standard deviation, median) of each quantitative variable. For qualitative variables, the percentage of the popularity of each attribute is also included.
- Quantitative: Presents the correlation, hypothesis tests, and statistical analysis (e.g. quantile 25%, skew) when necessary.
- Dimensionality Reduction: Explores the predictive capabilities of the dataset when reduced in lower dimensions. If performance is not severely affected then the dataset can be simplified (this depends on the available data).
- Above/Below 60% grade (or other): Explore patterns between student performance with several variables (e.g. attendance, travel type).
- Grade Classification (or equivalent): Explore patterns between student performance with several variables (e.g. ethnicity, socioeconomic).
- D.O.B: Explore patterns between students with date/time variables (e.g. weekday, year) and student performance.

While the abovementioned task list is recommended, there is the possibility of replacing some with equivalent ones, provided that they provide similar functionality. The amount and type of analysis is a matter of preference, the only condition here is the analysis should be sufficient to give one a level of understanding of the dataset; the more analysis is conducted the better the understanding of the dataset that is achieved. Also, the results of the data analysis phase are directly dependent on the quality of the dataset.

After this layer, the Predict Pre-Processing layer is executed. From this point, a given dataset's strengths, weaknesses, and hidden patterns are understood using data analysis.

3.3 Predict Pre-Processing

This layer is the starting point to predict student performance. In addition to the Global Pre-Processing applied to the data, a second series of Pre-Processing tasks described in this layer needs to be considered.

The following list describes the measurements for the Predict Pre-Processing (if required):

- Collect a student group niche (optional): One may choose a specific student group depending on the problem. Examples of groups include first-year students, graduates, or perhaps first-time enrolling. Occasionally, a particular student group is applied as it may provide deeper hidden patterns (e.g. one group may be performing better than another). There are no restrictions on a niche but it should have sufficient samples for reliable outcomes. If applied, this must be the first or second step.

- Collect specific variables (optional): This is associated with the student niche, some niches may not be eligible for all variables. For example, current students are not eligible to have a final grade variable. So, it is required to collect the variables that suit the current circumstance of a student niche (group). If applied, this must be the first/second step.

- Label encodes any qualitative variables: Data variables that are nominal or ordinal represent in language format for human communication (e.g. English). However, in decision-making models, this data type does not provide any use. As a result, the data must be converted to a numerical format. Label encoding the data convert each unique attribute to a unique identity number. In this way, the decision-making models can use those factors in their evaluation.

- Scale-down & normalise: To avoid any biased outcomes, the data is scaled down with qualitative variables and normalised with quantitative variables (Z-Score). In decision-making models, applying these transformations helps improve performance and computation speed.

After this layer, the Feature Selection layer is executed. From this point, a given dataset is in the right format for producing outcomes.

3.4 Feature Selection

In normal circumstances, the Feature Selection layer depends on the data type. In this framework, the selected Feature Selection does not involve any barriers to data types and therefore does not provide any issues in checking eligibility. Furthermore, the configuration of each Feature Selection is a matter of preference. There are no prerequisite settings for hyperparameters.

The following Feature Selections are applied:

- Genetic Algorithm (GA)

- Particle Swarm Optimisation (PSO)

- Recursive Feature Elimination (RFE)

After this layer, the Training/Testing Split layer is executed. From this point, a particular combination of features from a given dataset has been selected (or the original feature set is retained if no Feature Selection is applied).

3.4.1 Genetic Algorithm (GA)

The Genetic Algorithm follows the theory of evolution where the next generation's performance is improved by learning from its predecessors. The method is converted to a Feature Selection by mapping a list of features to each individual. The fitness score is calculated with each individual and the one with the highest fitness score is selected. The selected individual's assigned features are chosen. Equation 3.1 is the fitness score for the Genetic Algorithm. Where N is the total population, p is a specific population, M is the total number of folds, f is a specific fold, and then average the benchmark scores from each f.

$$Fitness_{ga} = \frac{1}{M}\sum_{f=1}^{M}\sum_{p=1}^{N}\left(\frac{TP_{f,p} + TN_{f,p}}{TP_{f,p} + TN_{f,p} + FP_{f,p} + FN_{f,p}}\right) \tag{3.1}$$

3.4.2 Particle Swarm Optimisation (PSO)

Particle Swarm Optimisation relies on a group of particles assisting each other by moving toward the solution of a given problem inside a sub-space. The method is converted to a Feature Selection by assigning a set of features to each particle. Each particle's assigned features compute the benchmark performance using a Supervised Learning machine learning algorithm.

The best predictor is positioned in the middle of a given sub-space. The remaining particle collects attributes from the particles in the middle and starts to move toward them. This means incoming particles can become better predictors than the original. If so, then it becomes the best predictor. Each particle continues to compete until the iterations are completed. The best particle and the list of features assigned to that particle is the final decision.

3.4.3 Recursive Feature Elimination (RFE)

Recursive Feature Elimination requires one decision-making Supervised Learning model, this chosen model fits the features to itself, ranks the features, and eliminates the lowest ranking feature. The process repeats until the execution is stopped. The execution is stopped by notifying the minimum number of features. For example, if the minimum is 3 features to an RFE, this means it keeps eliminating the worst features until it reaches 3.

It is crucial the correct model is selected as it can return the wrong list of features to a problem. For example, Classification problems should use Classification Supervised Learning models such as Logistic Regression. It is also worth noting that not all models can be used with this approach. Only models that compute the coefficients or feature importance are compatible as these metrics are used to find the weakest features.

For the proposed computation framework, the Cross-Validation version (5 folds) is applied which ranks each variable performance in all folds and removes one feature per iteration. The variables that are ranked first are the selected features.

3.4.4 Skipping Feature Selection

It is worth noting that there is always the option to skip feature selection in cases where the feature set is already optimised, or there is a limited number of features available. Retaining the original set of features may actually result in better performance and save computation time, so it may be preferable.

3.5 Training/Testing Split

This segment splits the data into two subsets. The D_{train} are samples used to train each model to predict the performance range for each student in D_{test}. The D_{train} collects 67% and the D_{test} collects 33% of the original dataset D. The data must be shuffled before splitting to avoid biased results.

Again, the split is a matter of opinion, the split percentage above is the default. It is worth mentioning that leaving a larger split for D_{train} can result in a longer execution time.

After this layer, the Anchored Training Data layer is executed. From this point, the dataset (D) is divided into training D_{train} & testing D_{test} sets with a ratio of 67% : 33% (or other if specified).

3.6 Anchored Training Data

This layer addresses the research gap in the existing literature that refers to reducing the probability of imbalanced data (including overfitting/underfitting).

The Anchored Training Data is applicable with the D_{train}.

The method evaluates the right number of samples to reduce the overfitting and underfitting probability. Reducing this probability also reduces the ratio of imbalanced data. This is because the best sample amount contains the best quality combination of samples. Better quality samples result in a more balanced ratio of attributes in each variable. The ratio is likely to be (more) balanced compared to the original dataset specification.

The selected technique is to find the best number of samples for each experiment. This purpose is to reduce the overfitting/underfitting issue as it can be a common problem. The method involves shuffling and splitting the data into segments s where $s = 5$. Each segment s_i contains an even ratio of samples from the original dataset D. This is processed in a range of iterations i where $i = s$. In each iteration i_n, it collects the current segment s_i and all previous segments $s_{i-1}, s_{i-2}, ..., s_{i-n}$ (if any). These segments are merged together to form one (temporary) dataset D_s. The next step is to split D_s into two (further) sets: training D_{train} (67%) & testing D_{test} (33%) sets. This is followed by inserting the D_{train} & D_{test} to a Logistic Regression (Sigmoid function) and producing the predictions. The benchmark metric is then executed (in this case, the Accuracy benchmark is used). This same process is repeated for all i. The sample amount with the best benchmark performance is the selected number of samples to apply to the experiments.

Figure 3.2 provides a visual representation of the process.

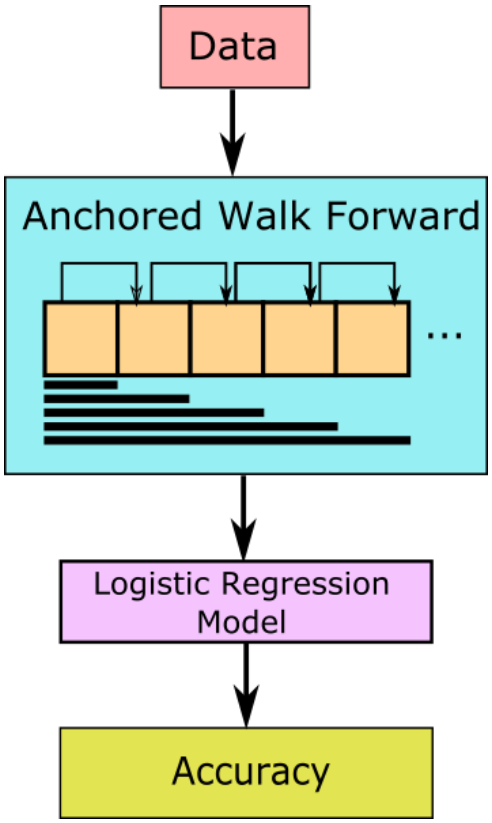

FIGURE 3.2: The process of how the Anchored Walk Forward process leads to identifying the best number of samples.

After this layer, the Model layer is executed. From this point, D_{train} is filtered with the best-suited number of samples. There are no changes with D_{test}.

3.7 Model

Integrating Supervised Learning and Unsupervised Learning is applicable to the model architectures within this layer. The judgment of models is subjective, though it is important to note that the models should suit the problem. However, it is preferable that several models are included which can be multiple architectures of a given model and/or unique decision-making algorithms. The Model layer should also include start-of-the-art methods to maximise potential performance. In general, an exploration of 5 models in total should generally be sufficient. These may include, for instance:

- Random Forest
- AdaBoost
- K-Nearest Neighbours
- Gradient Boost
- Extreme Gradient Boost
- Gaussian Process
- Passive Aggressive

- Extra Trees
- (Feed Forward) Neural Networks
- K-Modes
- K-Prototype

In addition, it depends on the given dataset as some models may not be eligible. For example, integrating Unsupervised Learning models only accept specific data type (e.g. K-Modes only accepts qualitative attributes).

This layer addresses several research gaps in the existing literature: (1) conducting an exploration of Unsupervised Learning in a model pipeline process (as a pre-processing step to group relevant student data before producing predictions) to assist student performance predictions; (2) applying multiple (Feed Forward) Neural Network architectures with different hidden layers (in one experiment); (3) predicting student performance ranges which describe the highest & lowest academic abilities.

In this part, the model collects D_{train} version outputted from the Anchored Training Data. That D_{train} is trained with Machine Learning algorithms to predict the performance ranges for each student in D_{test}. The outcomes are then fed into the benchmark process.

There are two types of architectures:

- Parallel Architecture - Executed for decision-making models that require a target vector and a feature matrix/vector (Figure 3.3).
- Popularity Architecture - Executed for decision-making models that require a feature matrix/vector (Figure 3.4).

The Parallel Architecture integrates Supervised Learning in its pipeline process. In this architecture, for each student in D_{test}, the minimum and maximum academic performance are predicted, which becomes the performance range. This does mean that two executions are computed per student. The machine learning algorithms in the Parallel Architecture are the main step for predictions. These algorithms are built-in with the feature to predict future events. For this reason, this architecture includes fewer processes compared to the remaining ones.

The Popularity Architecture integrates Unsupervised Learning in its pipeline process. Each student's variables in D_{test} are grouped with students in D_{train}. In other words, similar student data are grouped together (which is the purpose of Unsupervised Learning). Those grouped samples are collected and the popularity of academic performance is filtered. The popularity judgment is based on the number of occasions each academic performance occurred above the median. The minimum and maximum academic performance from the popular set becomes the performance range (becomes the prediction). Furthermore, before the usage of Unsupervised Learning, one must determine the right cluster c_i. Fortunately, there are cluster performance methods to determine the outcome. Each cluster's performance differs and their specification must be met. The machine learning algorithms in the Popularity Architecture are a pre-processing step for predictions. The duty is to collect student samples with similar characteristics. These algorithms are not built-in with the feature to predict future events. This architecture includes more processes compared to the remaining.

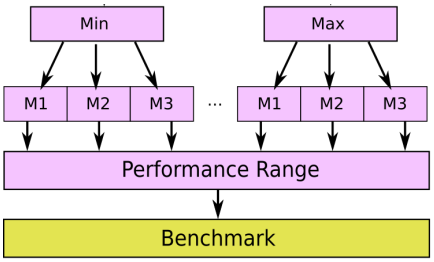

FIGURE 3.3: Parallel Architecture integrates Supervised Learning.

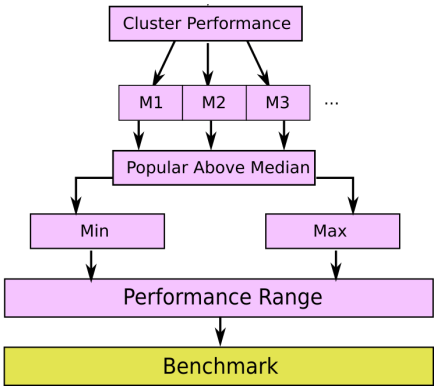

FIGURE 3.4: Popularity Architecture integrates Unsupervised Learning.

After this layer, the Benchmark layer is executed. From this point, the models have predicted the performance range.

3.8 Benchmark

This layer also addresses the aforementioned research gap related to predicting performance ranges. The association is on the basis of evaluating if the prediction is correct or not.

The traditional benchmarks are implemented to compare outcomes between x & y vectors $\mathbb{R}^n$ that contain only singular values. x is the prediction vector and y is the true vector. The given framework outputs a range and is therefore not compatible with the well-known benchmarks.

As a result, the framework includes a compatible benchmark, which this thesis refers it as the *Valid Range*. The benchmark compares a range with a singular value and evaluates predictability if the singular value is within the range. If the singular value is within the range is correct, otherwise, it is not. The output is binary-based and executed on each model. This benchmark is **not** an improvement to the traditional version, it is not compatible with the output results from this work.

Table 3.1 & Figure 3.5 provides an example. The *Min* & *Max* columns are the performance ranges (prediction output from this framework). The *Correct* column is the benchmark output ($\mathbb{R}^n$) from this framework, that verifies whether the *True* column (by default its the average performance) is within *Min* & *Max*. If so, then the range is correct, otherwise, the range is not correct. The output is binary-based but can be converted to percentages, as represented in Equations 3.2 & 3.3.

$$Correct\ Ratio\ (\%) = \frac{Number\ of\ Correct\ Samples}{Total\ Number\ of\ Samples} \tag{3.2}$$

$$Incorrect\ Ratio\ (\%) = \frac{Number\ of\ Incorrect\ Samples}{Total\ Number\ of\ Samples} \tag{3.3}$$

TABLE 3.1: Display an example of the output from this project.

Student	Min	Max	True	Correct
ID 1	0.3	0.7	0.5	1
ID 2	0.4	0.8	0.9	0
ID 3	0.4	0.5	0.4	1
...	...	...	...	...

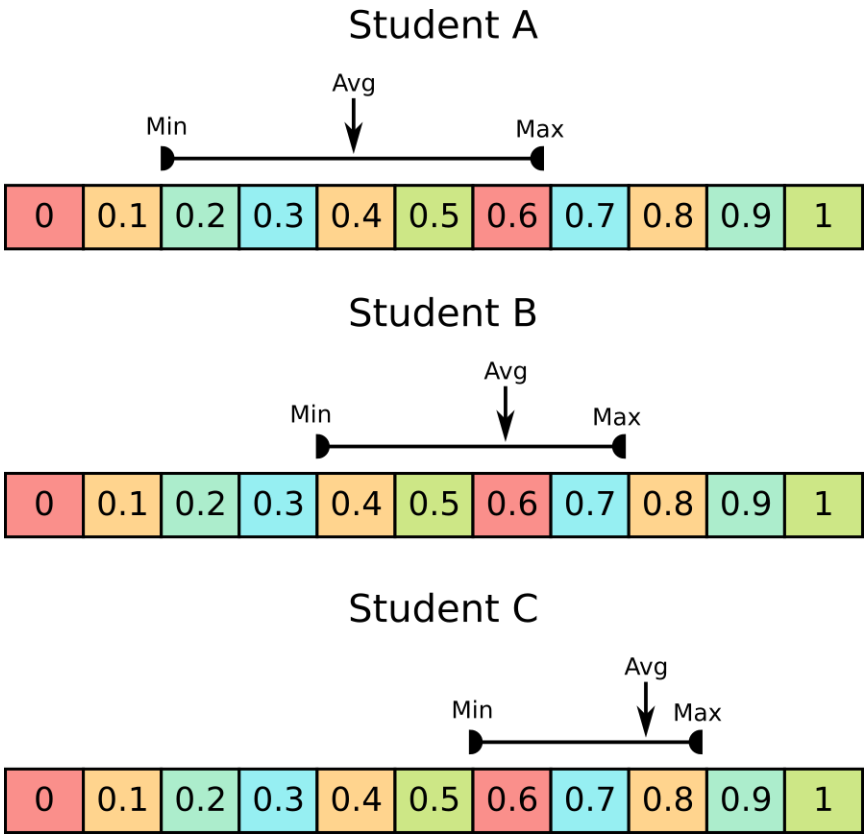

FIGURE 3.5: Example of the benchmark using a scale (for demonstration purposes).

This is the final layer and completes the proposed computational framework. The benchmark shows the prediction accuracy of each model. Then one can perform an evaluation.

Chapter 4

Experiment

This chapter validates the applicability of the computational framework presented in Chapter 3 by applying it to a given dataset. Each section represents a layering process and shows the configuration, comparison of existing literature, discussion about choices, and/or results (when necessary) to predict student performance. Again, the specification of each layer is mentioned in Chapter 3.

Figure 4.1 provides the process executed in this experiment. Two experiments are applied to the computational framework. Experiment One applies the Parallel Architecture and Experiment Two applies the Popularity Architecture.

4.1 Data & Global Pre-Processing

In this section, a background understanding of the data is presented. This includes providing the origin of the data, a description of the data, and the Global Pre-Processing tasks (Section 3.1).

4.1.1 Data Supplier

The dataset is given by The University of Huddersfield. The University of Huddersfield is an institution based in West Yorkshire, Huddersfield, United Kingdom. The University features in all Higher Education information and ranking portals, such as *The Complete University Guide* [77]. They offer undergraduate, postgraduate, and doctoral degree courses. In the period of time the dataset was collected there were 231 full-time courses, 168 sandwich/placement courses, and & 8 part-time courses. The University of Huddersfield prides itself in that 100% of undergraduates undertake professional work-related experience after their studies [44].

4.1.2 Data Description

The given dataset is in the form of student records that consist of completed modules between $2014 - 2018$. In total, the original dataset contains $200,000$ rows (size of data) that contribute to over $27,000$ students (when grouped). Out of these, about $14,000$ students are graduates. There are 52 columns but several variables have the same definition; the difference is that one presents the code, and the other presents the title (e.g. department code and name). Therefore, 27 variables contain unique definitions. All students are undergraduates, between years $1 - 3$. The variables contain several demographic data such as ethnicity, gender, and parental education. The remaining variables are assessment and course details such as attendance, grades, and course title.

Table 4.1 presents further details (e.g. data attributes, types) of each variable within the (original) dataset. There are 15 variables that are nominal, 8 are ordinal,

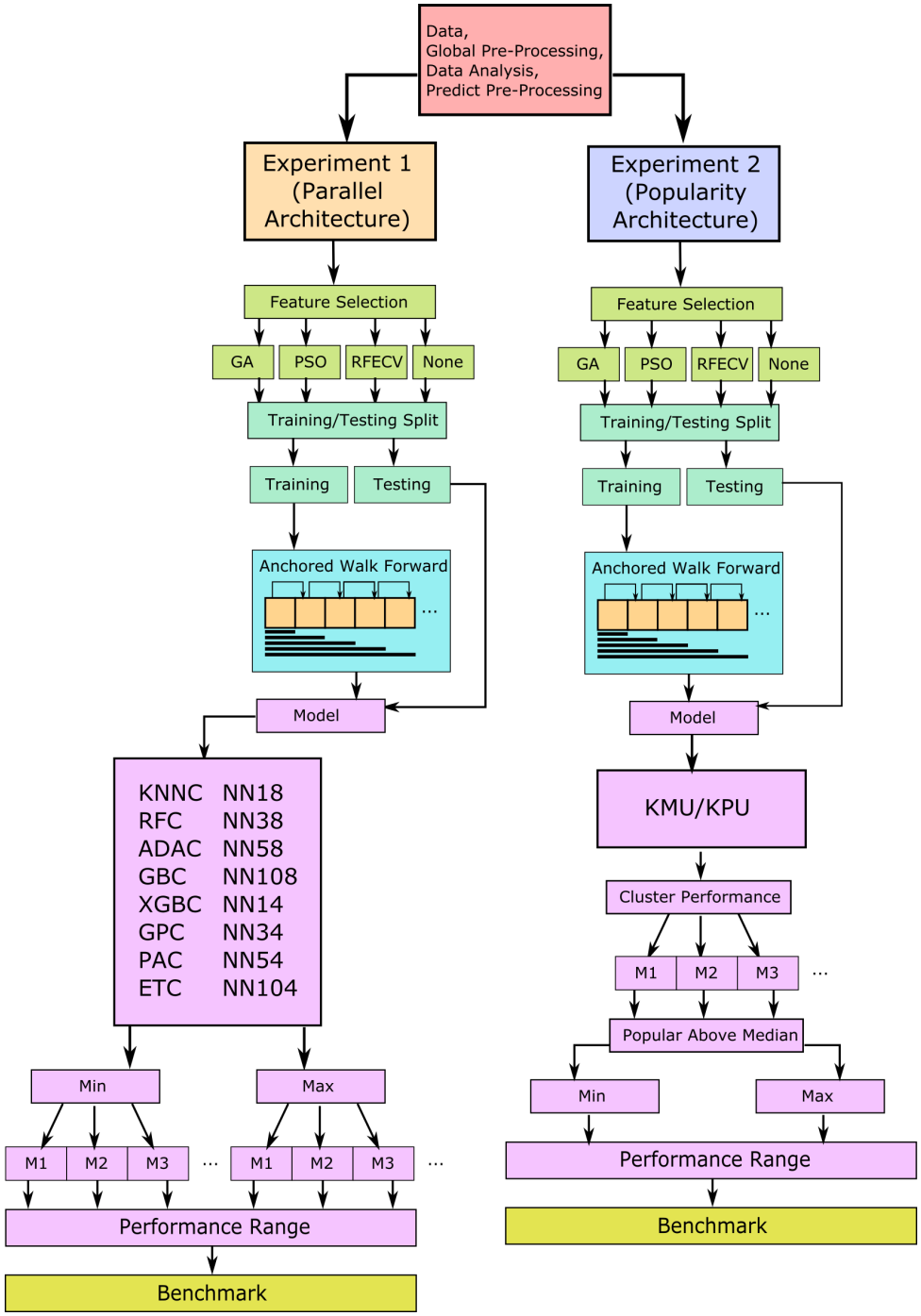

FIGURE 4.1: A diagram of the pipeline process applied to this experiment. Both model architectures are applied, and the experiment follows the specification explained in Chapter 3.

and 3 are continuous. Most variables are qualitative (24 variables) and the remaining are quantitative (3 variables). Also, the data format is mainly non-numerical (21 variables) rather than numerical (6 variables). Most variables in each row tend to duplicate especially demographic data. The non-demographic data deliver the uniqueness between rows. The common non-demographics include attendance, grades, and swipe time.

Table 4.2 provides an example of the representation of the dataset (not real data). Each row represents a completed module. This means one row or more can be associated with a student. For example, row 1 can be *student A* and completed *Module A* with 50%. Row 2 can be *student A* and completed *Module B* with 79%. Row 3 can be *student B* and completed *Module B* with 62%.

There are several strengths and weaknesses with the given dataset. The strengths include sufficient sample size, sufficient variables, an adequate ratio of demographic and course variables, sufficient academic year cycles, sufficient student groups, and the fact that the dataset contains both current students and graduates. The weaknesses include incomplete data (missing or incorrect details such as an address, & travel type), and imbalanced data (e.g. ethnicity is biased to the White background only).

Differences from datasets in literature

There are several differences between the dataset used in this research and those in the existing literature and these are discussed next. The dataset in this experiment, the sample size, and the variable amount is higher than in past papers, and the range tends to differ depending on the past paper (e.g. [18], [2], [88], [60]). Normally, it is around 10000 for samples (or more) and around 9 for variables (or more). For example, Bilal et al. [18] dataset contains 76 samples and 17 variables, Zhang and Rangwala [88] dataset contains 13000 samples and 11 variables.

Most variables contain a wide range of attributes and there is no particular pattern among them. The attributes in the Student ID variable should equal the total number of students within the dataset. But the remaining attributes can correspond to a collection of students in each variable. This results in several student groups which can contain high/low amounts of samples and/or variables (depending on the data filtering).

This dataset contains several variables but also misses other variables compared to the literature (and vice versa). Examples of missing variables include pre-entry academic performance such as A Level grades (or equivalent), health status (e.g. disability, anxiety) & household income. It can provide more insight into identifying patterns between their circumstances and their academic performance. The variables that are missing in this dataset do exist in past works (Aggarwal, Mittal, and Bali [5], Xu, Moon, and Schaar [85]). On the other hand, all variables in the given dataset have existed in some way in other datasets (e.g. age, domicile, & grades) but the particular combination of variables is unique.

Also, this dataset contains uncommon & useful variables such as IMD, POLAR 3/4, and socioeconomic. These variables are available in UK institutions only (to the author's knowledge). In the UK, institution data are collected from UCAS, so student records from UK universities contain similar variables. Of course, the data format and structure between UK universities can be different. For example, one institution may record based on each module but some may not apply that practice (e.g. stored in a separate database).

Most existing research that uses datasets from UK institutions does not fully express the (data) specification. The best case scenario is one paper by Thiele et al. [75] that mentions collecting data from a UK institution. It does provide sufficient details that variables IMD, POLAR 3/4, and socioeconomic (or other) are included. However, the number of samples is below 6000 which is (far) less in comparison to the data provided in this thesis (which is the difference).

In summary, this dataset contains a unique combination of academic year cycles, modules, courses, student home/term residency, sample amount (rows), and more. Which also delivers novelty compared to past dataset usage.

Similarities to datasets in literature

In addition, in the existing literature, the datasets used in past papers and this dataset share similarities. The datasets in past works are collected from institutions, and/or third-party organisations (e.g. Coursera). The challenges (ethics and law) of collecting data are relatable (e.g. data protection) (Chaudhry and Kazim [24], Dignum [26]). Within those datasets, most variables are qualitative, minor variables are quantitative. Most variables that are qualitative are demographic data and most variables that are quantitative are academic/course data. The students are commonly undergraduates (or equivalent level). Furthermore, incomplete data and imbalanced data exist in this dataset and in past papers datasets.

Moreover, the case studies (Alwarthan, Aslam, and Khan [8], Hellas et al. [40]) explore more past research and provide the common and uncommon trends between this dataset and those datasets. The common patterns include most variables being course details, the number of variables being around 25, most variables being qualitative, and the assessment variables being grade values. The uncommon pattern(s) include the (popular) sample size being between $30 - 300$. These trends show the dataset in this experiment shares more similarities than differences. The differences are not beneficial to this dataset or the experiment.

TABLE 4.1: Details of the original dataset such as definitions & attributes. Note there are many variables that have the same definitions but the representation differs, the common one is represented as name and code.

Column	Description	Format	Attributes	Type
Student ID	The student identity.	String	27,018	Nominal
Academic Year	The occurred academic year.	String	4	Ordinal
Status	The student's current circumstance.	String	15	Nominal
Gender	The student's gender.	String	2	Nominal
Department	The department the student studied.	String	26	Nominal
School	The school the student studied.	String	7	Nominal
Module	The module the student studied.	String	1,235	Nominal
Module Grade	The student's % grade achieved from a module.	Float	$0 \leqslant x \leqslant 100$	Continuous
Graduate Grade	The student's final classification grade.	String	4	Ordinal
Attendance	The student's attendance per module.	Float	$0 \leqslant x \leqslant 1$	Continuous
Home Postcode	The student's home residence.	String	18,006	Nominal
Term Postcode	The student's study term residence.	String	12,411	Nominal
Travel Type	The student's method of traveling to the institution.	String	3	Nominal
Accommodation	The name of the student's study term residence.	String	8	Nominal
Domicile	The student's country of origin.	String	113	Nominal
Ethnicity	The student's ethnicity.	String	19	Nominal
Entry Qualification	The student's latest & previous qualification.	String	48	Ordinal
Parental Education	The student's degree educated or not.	String	6	Ordinal
Last School	The student's last institution.	String	1,750	Nominal
Route	The student's course choice on their application.	String	461	Nominal
Course	The student's current course during studying.	String	168	Nominal
IMD	The student's home residence deprived rank.	Integer	5	Ordinal
POLAR 3/4	The student's rank of participation in higher education.	Integer	5	Ordinal
Socioeconomic	The student's allocated occupation.	String	8	Ordinal
Swipe Time	The student's time attendance on each scheduled lesson.	Float	$-30 \leqslant x \leqslant 377$	Continuous
D.O.B	The student's date of birth.	String	1,760	Ordinal

TABLE 4.2: Dummy rows that represent the actual original dataset. Each row corresponds to a completed module.

Column	Row 1	Row 2	Row 3
Student ID	ID 1	ID 1	ID 2
Academic Year	14/15	14/15	14/15
Status	Current	Current	Current
Gender	Female	Female	Male
Department	Computer Science	Computer Science	Engineering
School	Computing and Engineering	Computing and Engineering	Computing and Engineering
Module	Programming	Mathematics	Electronics
Grade	69.8	56.9	61.9
Graduate Grade	Upper Class Division 1	Upper Class Division 1	Upper Class Division 2
Attendance	0.8	0.9	1
Home Postcode	JJJ 4DF	JJJ 4DF	HJJ 888
Term Postcode	HD1	HD1	HD1
Travel Type	Not-Commuter	Not-Commuter	Commuter
Accommodation	Parents	Parents	Huddersfield Halls
Domicile	England	England	China
Ethnicity	White	White	Asian
Entry Qualification	A Levels	A Levels	Level 3
Parental Education	No	No	Yes
Last School	Huddersfield College	Huddersfield College	Hong Kong College
Route	Computer Science	Computer Science	Engineering
Course	Computer Science	Computer Science	Engineering
IMD	1	1	Not Classified
POLAR 3	2	2	Not Classified
POLAR 4	4	4	Not Classified
Socioeconomic	Higher Managerial	Higher Managerial	Not Classified
Swipe Time	-18.5	1.6	33.5
D.O.B	12/05/1998	12/05/1998	01/02/1997

4.1.3 Global Pre-Processing

As mentioned in the framework specification, the Global Pre-Processing layer is the first applied to a given dataset. The dataset is grouped by student ID. The quantitative variables (e.g. module grades) return the average value, and the qualitative variables return the first attribute (the attributes are duplicates). Variables like Academic Year and Module are removed as they deliver no purpose after being grouped by student ID. This dataset contains variables with the same definitions, for this experiment, only variables with different definitions are collected. Samples are removed that contain any incomplete data, this means each row must not be empty.

Table 4.3 displays additional variables computed using algorithms. The (new) academic years are calculated by counting the number of students that are allocated to each academic year cycle. The age is calculated by computing the time year difference between a student's date of birth to the time of execution. The age outcomes depend on the time of execution. The distribution is the important factor and it is useful to identify patterns. The home and term distance is calculated by converting the location postcodes to latitude and longitude and then computing their distance from the university location. This can be implemented through available open-source APIs, e.g. ones written in Python. The outcomes are in kilometers (KM). The standard grade compares the average module grade of each student against a 60% threshold, the outcome is either 1 or 0. The threshold is marked as 60% due to the popularity of entry qualifications in postgraduate courses and graduate schemes.

The Consistency-Scale is a simplified version of the module grade (in the given dataset). It converts the module grade variable into a rank scale. The technique involves rounding the module grades to the nearest 10^{th} and then dividing them by 100. Each student receives a rank score between $0-1$. 0 is the lowest rank score and 1 is the highest rank score, each interval is by 0.1. It is possible to divide it by 10 rather than 100. This results in a range between $0-10$ (rather than $0-1$) which means intervals are by 1. In any case, the definition remains the same. The Consistency-Scale can also be referred to as a Grade-Scale.

With the given dataset, unique grade outcomes are represented in 10^{th} intervals (e.g. $50, 60, 70$). The remaining values do not provide any unique academic outcome. So, it is sensible to round the variable to the nearest 10^{th}. Also, narrowing down the outcomes improves the prediction performance from the decision-making models (this may not be the case for all). Therefore, the Consistency-Scale variable is produced to improve the time execution, prediction performance & simplicity (in this experiment).

For experimentation purposes, the $0-1$ range is used (to fulfill the specification of all the decision-making algorithms). Figure 4.2 presents the Consistency-Scale diagram.

Given these requirements, Table 4.4 provides the finalised dataset version applied to this experiment.

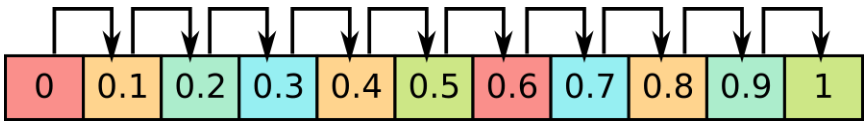

FIGURE 4.2: Illustration of the Consistency-Scale rank. The worst to best rank score starts from left to right.

TABLE 4.3: New variables added to the original dataset such as definitions & attributes.

Column	Description	Format	Attributes	Type
Academic Years	The student's total number of years.	String	4	Ordinal
Standard Grade	The student's average module grade is 60% or above.	Integer	2	Ordinal
Age	The student's current age	Integer	$22 \leqslant x \leqslant 85$	Discrete
Modules (Amount)	The total number of modules.	Integer	$1 \leqslant x \leqslant 23$	Discrete
Term Distance	The student's accommodation KM distance to the university.	Float	$0 \leqslant x \leqslant 458$	Continuous
Home Distance	The student's home KM distance to the university.	Float	$0 \leqslant x \leqslant 5949$	Continuous
Consistency-Scale	The student's performance scale (min, average, & max versions).	Float	10	Ordinal

TABLE 4.4: This is the conversion version of the dataset. It is applied to the computational framework.

Column	Description	Format	Attributes	Type
Student ID	The student identity.	String	27,018	Nominal
Academic Years	The student's total number of years.	String	4	Ordinal
Status	The student's current circumstance.	String	15	Nominal
Gender	The student's gender.	String	2	Nominal
Department	The department the student studied.	String	26	Nominal
School	The school the student studied.	String	7	Nominal
Modules (Amount)	The student's total number of modules.	Integer	$1 \leqslant x \leqslant 23$	Discrete
Module Grade	The student's average % grade achieved.	Float	$0 \leqslant x \leqslant 100$	Continuous
Graduate Grade	The student's final classification grade.	String	4	Ordinal
Attendance	The student's average attendance.	Float	$0 \leqslant x \leqslant 1$	Continuous
Home Postcode	The student's home residence.	String	18,006	Nominal
Term Postcode	The student's study term residence.	String	12,411	Nominal
Travel Type	The student's method of traveling to the institution.	String	3	Nominal
Accommodation	The name of the student's study term residence.	String	8	Nominal
Domicile	The student's country of origin.	String	113	Nominal
Ethnicity	The student's ethnicity.	String	19	Nominal
Entry Qualification	The student's latest & previous qualification.	String	48	Ordinal
Parental Education	The student's degree educated or not.	String	6	Ordinal
Last School	The student's last institution.	String	1,750	Nominal
Route	The student's course choice on their application.	String	461	Nominal
Course	The student's current course during studying.	String	168	Nominal
IMD	The student's home residence deprived rank.	Integer	5	Ordinal
POLAR 3/4	The student's rank of participation in higher education.	Integer	5	Ordinal
Socioeconomic	The student's allocated occupation.	String	8	Ordinal
Swipe Time	The student's average time attendance on each scheduled lesson.	Float	$-29 \leqslant x \leqslant 263$	Continuous
D.O.B	The student's date of birth.	String	1,760	Ordinal
Standard Grade	The student's average module grade is 60% or above.	Integer	2	Ordinal
Consistency-Scale	The student's performance scale (min, average, & max versions).	Float	10	Ordinal
Age	The student's current age	Integer	$22 \leqslant x \leqslant 85$	Discrete
Term Distance	The student's accommodation KM distance to the university.	Float	$0 \leqslant x \leqslant 458$	Continuous
Home Distance	The student's home KM distance to the university.	Float	$0 \leqslant x \leqslant 5949$	Continuous

4.2 Data Analysis

In this section, a number of statistical analyses are applied to help understand the data. Figure 4.3 provides a step-by-step pipeline process of the steps that need to be carried out.

The data analysis phase aims to explore the data in-depth and discover any hidden patterns that can deliver benefits to predicting student performance. The following data analysis is applied to this experiment:

- Frequencies: Attributes of qualitative variables and statistics of quantitative variables.
- Quantitative: Pearson correlation, distribution hypothesis tests, and many statistical analyses.
- Dimensionality Reduction: Explore with PCA, MCA & FAMD. To verify if the data can be shrunken accurately.

- Above/Below 60% grade: Explore student performance patterns with many qualitative and quantitative variables (when necessary).
- Grade Classification: Explore student performance patterns with many qualitative and quantitative variables (when necessary).
- D.O.B: Explore patterns between date/time and student performance (when necessary).

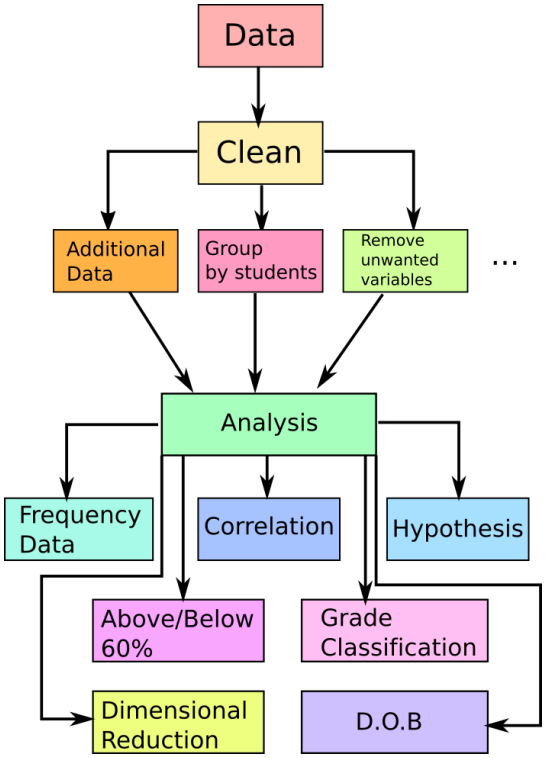

FIGURE 4.3: Display the pipeline process of the data analysis and statistics.

4.2.1 Frequencies

Tables 4.5, 4.6 and 4.7 provide details of each variable. The qualitative variables include the number of attributes associated with each student. Due to the number of attributes in some variables, all attributes are not displayed. The quantitative variables contain statistics as it is more suitable. This includes the minimum, median, and maximum.

TABLE 4.5: Display statistics of quantitative data, and attributes of qualitative data. Note that due to the number of samples, some data columns are not presented here. Data that contain no details are ignored.

Column	Stats / Values	Freq (%)
Academic Years	1 Year Cycle: 12846 2 Year Cycles: 8446 3 Year Cycles: 5910 4 Year Cycles: 500	46.37 30.49 21.33 1.8
Status	Current Student: 25580 Debtor Current Student: 495 ...	92.34 1.79 ...
Gender	Male: 12589 Female: 15112	45.44 54.55
Department	Behavioural and Social Sciences: 1780 Computer Science: 1642 Health Sciences: 2941 ...	6.43 5.93 10.6 ...
School	Applied Sciences: 2655 Art, Design and Architecture: 3625 Computing and Engineering: 3937 Education and Professional Development: 1620 Huddersfield Business School: 6852 Human and Health Sciences: 5804 Music, Humanities and Media: 3209	9.58 13.09 14.21 5.85 24.7 20.95 11.58
Modules (Amount)	Mean (std) : 7.83 (4.15) Min $<$ Med $<$ Max : $1 < 6 < 23$ IQR (CV) : 6 (53.06)	23 distinct values
Module Grades	Mean (std) : 56.03 (17.49) Min $<$ Med $<$ Max : $0.0 < 59.75 < 95.46$ IQR (CV) : 16.5 (31.22)	100 distinct values
Graduate Grade	1st Class: 4240 2:1 Class: 5931 2:2 Class: 3469 3rd Class: 584 Other: 237	29.32 41 23.99 4 1.64
Attendance	Mean (std) : 0.7 (0.19) Min $<$ Med $<$ Max : $0.0 < 0.72 < 1.0$ IQR (CV) : - (27.08)	100 distinct values

TABLE 4.6: Statistics of quantitative data, and attributes of qualitative data. Note that due to the number of samples, some data columns are not presented here. Data that contain no details are ignored.

Column	Stats / Values	Freq (%)
Ethnicity	Whites: 15550	68.74
	Asian Pakistani: 3275	14.48
	...	...
Entry Qualification	A/AS level: 7594	27.5
	Other qualification at level 3: 1293	4.68
	...	...
D.O.B	1999-01-10: 50	0.18
	1996-08-18: 102	0.37
	...	...
Parental Education	Do not know: 2231	8.4
	Information refused: 2099	7.9
	No: 12965	48.6
	No response given: 5	0.019
	Yes: 9399	35.2
Socioeconomic	Higher managerial and professional occupations: 2381	
	Intermediate occupation: 2435	10.87
	Lower managerial and professional occupations: 4044	11.12 18.46
	Lower supervisory and technical occupations: 984	4.5 26.08
	Not classified: 5711	8.14
	Routine occupations: 1782	13.21
	Semi-routine occupations: 2893	7.64
	Small employers and own account workers: 1675	
IMD	Rank 1: 6972	31.3
	Rank 2: 4786	21.49
	Rank 3: 3582	16.08
	Rank 4: 3917	17.59
	Rank 5: 3017	13.54
POLAR 3	Rank 1: 3843	17.06
	Rank 2: 6414	28.46
	Rank 3: 5726	25.42
	Rank 4: 4091	18.16
	Rank 5: 2456	10.9
Domicile	England: 22339	80.64
	China: 2170	7.83
	...	...

TABLE 4.7: Statistics of quantitative data, and attributes of qualitative data. Note that due to the number of samples, some data columns are not presented here. Data that contain no details are ignored.

Column	Stats / Values	Freq (%)
POLAR 4	Rank 1: 3634 Rank 2: 5713 Rank 3: 6121 Rank 4: 4221 Rank 5: 2841	16.13 25.35 27.17 18.74 12.61
Course	BA (Hons) Business Management SW/FT: 1049 MPharm: 535 ...	3.79 1.93 ...
Age	Mean (std) : 28.15 (6.05) Min $<$ Med $<$ Max : 22 $<$ 27 $<$ 85 IQR (CV) : 4 (21.49)	85 distinct values
Term Distance	Mean (std) : 11.42 (22.85) Min $<$ Med $<$ Max : 0.0 $<$ 4.91 $<$ 457.92 IQR (CV) : - (200.0)	inf
Home Distance	Mean (std) : 57.88 (271.7) Min $<$ Med $<$ Max : 0.31 $<$ 23.15 $<$ 5948.53 IQR (CV) : - (469.46)	inf
Travel Type	Commuter: 12094 Not Commuter: 10446	53.66 46.34
Swipe Time	Mean (std) : -0.43 (15.31) Min $<$ Med $<$ Max : -28.87 $<$ -3.14 $<$ 263.06 IQR (CV) : - (-3521.81)	inf

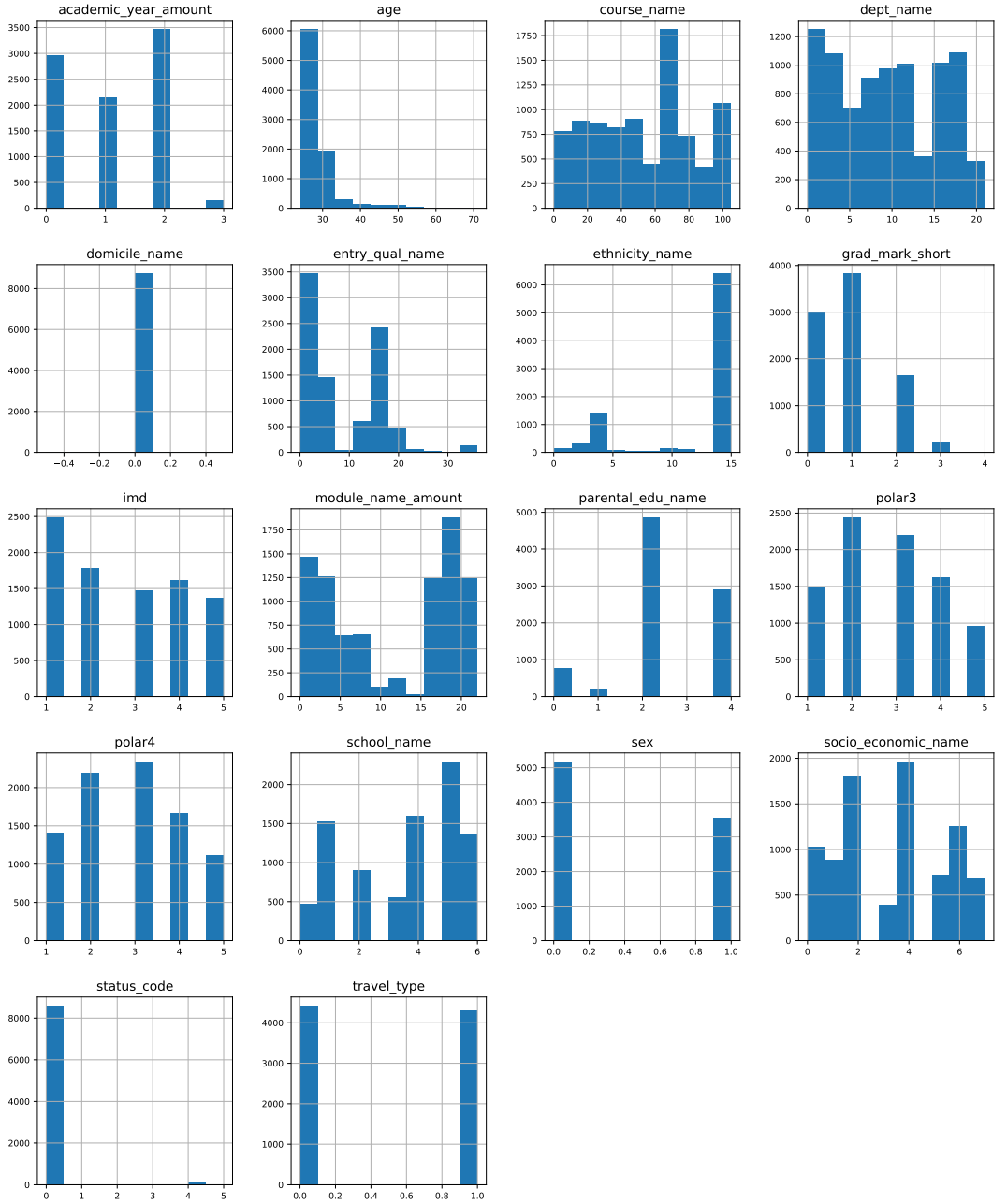

FIGURE 4.4: Distribution of all qualitative variables within the dataset. Variables that are non-numerical (e.g. entry qualification) are label-encoded first and then the distribution is computed. All other variables remain unchanged.

Statistical analysis shows several imbalanced data ratios within the dataset such as ethnicity and schools. The ethnicity shows mainly students of white background which should be expected given that the majority of the population in the United Kingdom are people of British white or other white backgrounds. If one explores each nationality based on different characteristics such as term distances, and home residence, (in most cases) the outcomes remain the same. Figures 4.5 and 4.6 display the students based on ethnicity against other relevant variables. In all cases grouped by ethnicity, Whites dominate on quantity and express the ratio to be biased. Therefore, it shows the imbalance between white backgrounds against other backgrounds.

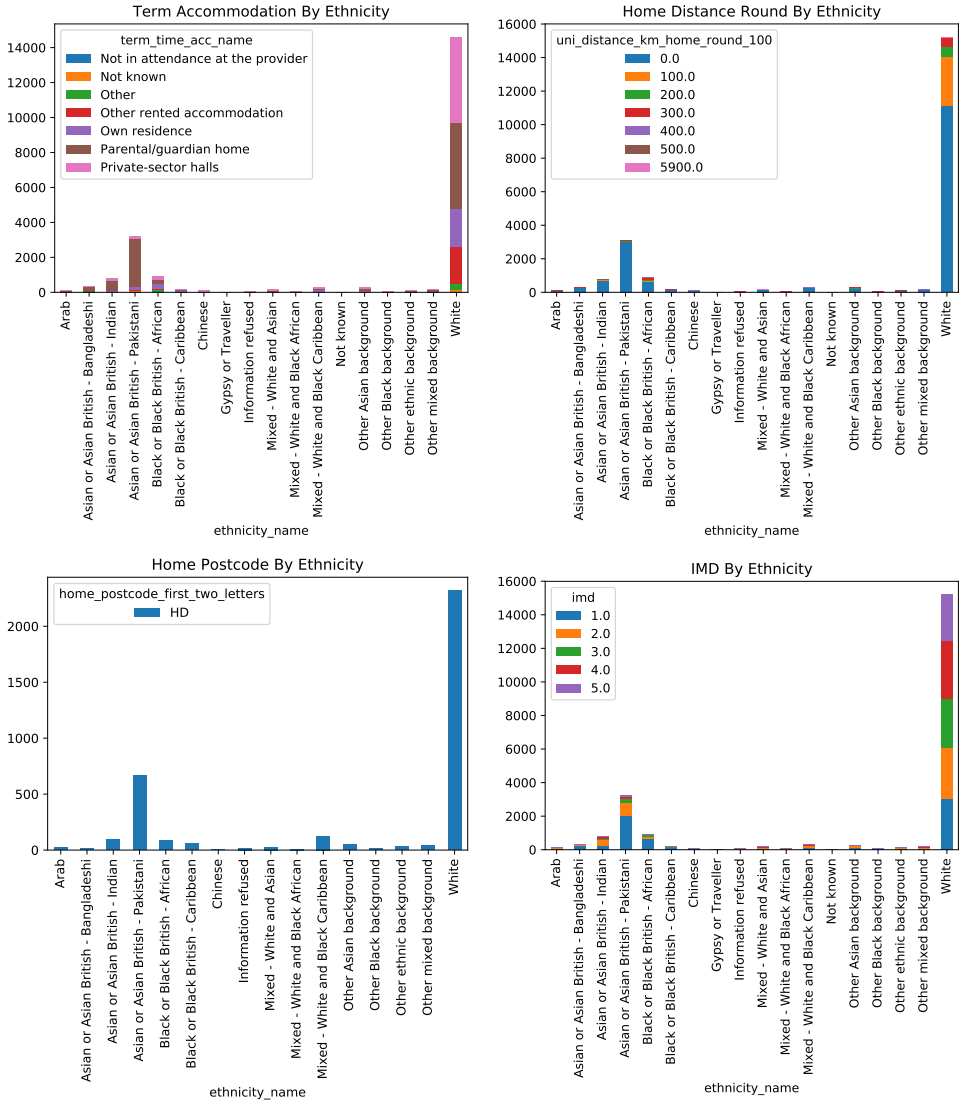

FIGURE 4.5: Display the number of students based on ethnicity against other variables. It is obvious from all conditions that Whites dominate all conditions.

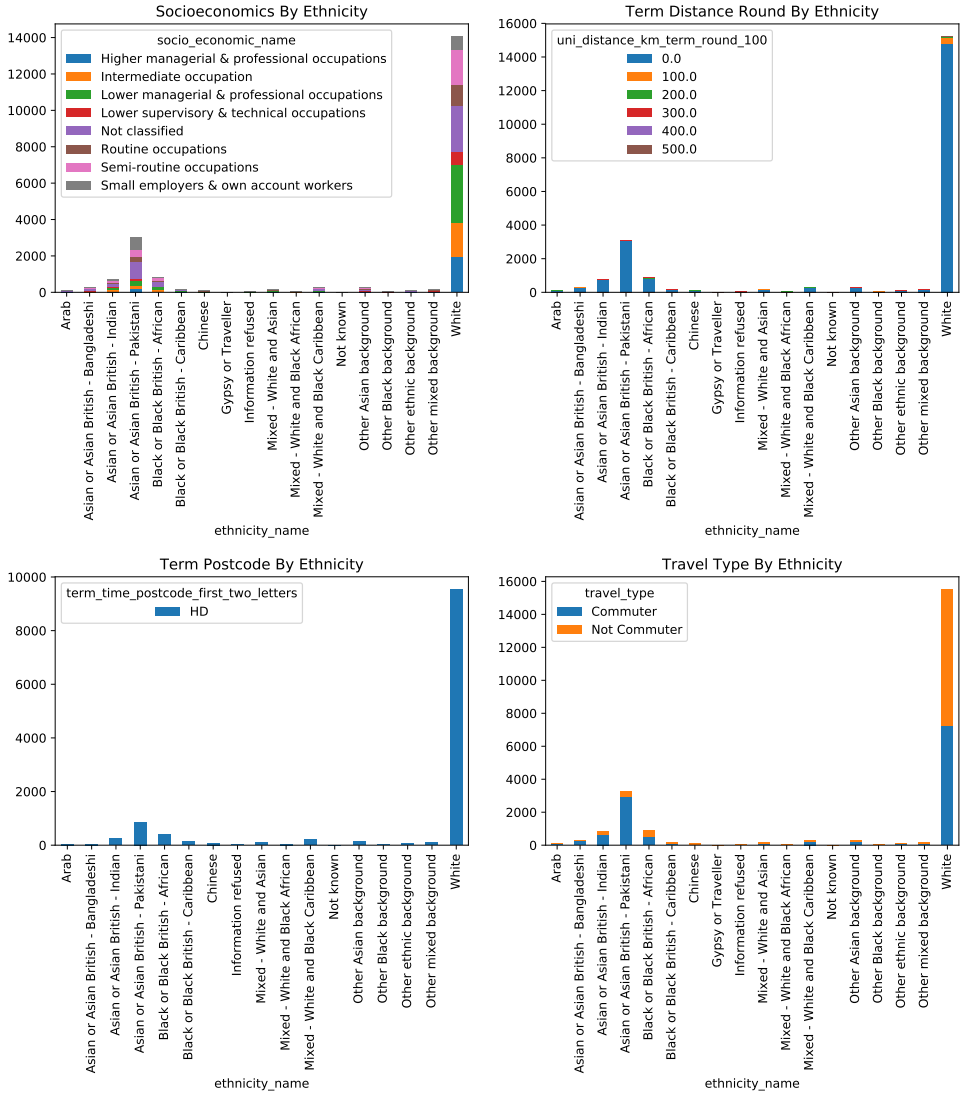

FIGURE 4.6: The number of students based on ethnicity against other variables. The dominance of students from a white background is evident.

With regard to academic schools in the University, analysis shows more students in non-science schools. This is again expected given that most courses are non-science related. Figure 4.7 illustrates the number of courses associated with each school.

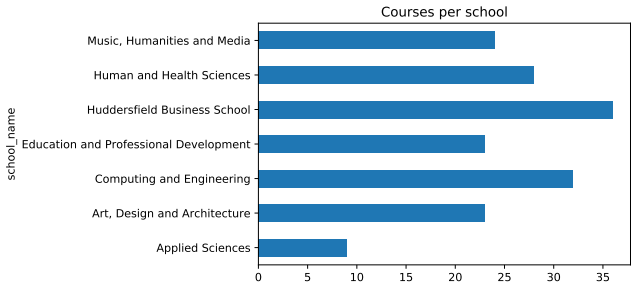

FIGURE 4.7: The number of courses based on schools.

Unfortunately, imbalanced data is the norm in this application area. Although it is possible to reduce it, there are variables (e.g. ethnicity) that are hard to execute given the current circumstances. The overall goal for both students and the institution is for students to achieve an Upper-Class Division I (60%) or above. This is because most postgraduate courses and jobs (graduate schemes or not) require this as a minimum grade. The statistics show most students graduate with Upper-Class Division I or above.

4.2.2 Quantitative

Looking further into the quantitative variables, some are understandable, and some are not. The distance variables show students are nearby for both home (5 KM) and term (23 KM). One can assume that living close to the university may result in a better quality of education. According to Figure 4.10, the correlation relationship between module grades and term distance is weak. But if it is illustrated as Figure 4.8 between distances and grade, it provides a much clearer picture.

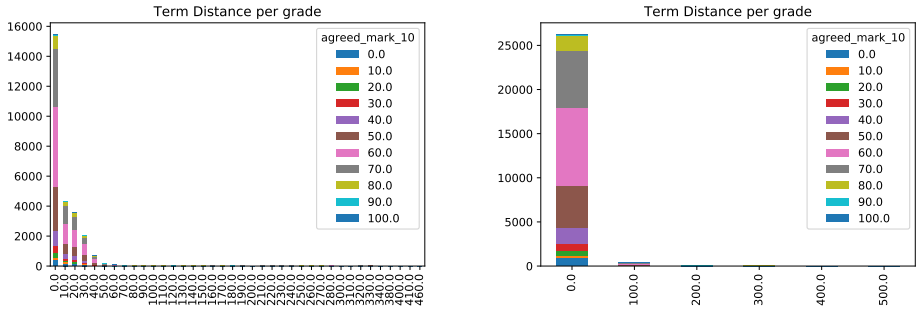

FIGURE 4.8: Module average grade from students based on term distances rounded to the nearest 10^{th} and 100^{th}. The majority of students are achieving 60% or above and therefore can prove the relationship between living close and higher performance.

The home distance is quite close, which shows strong signs of students from the North West and Yorkshire region. There are reasons why students are from these residences including household income, traveling, and family circumstances. The number of commuters and non-commuters is almost 50 : 50. The university is close to many types of public transport, the most common ones are buses and trains. If the topic is explored with several other variables as presented in Figure 4.9, the pattern repeats in most variables except the term distance. But this is expected as students living closer are more likely to be non-commuters.

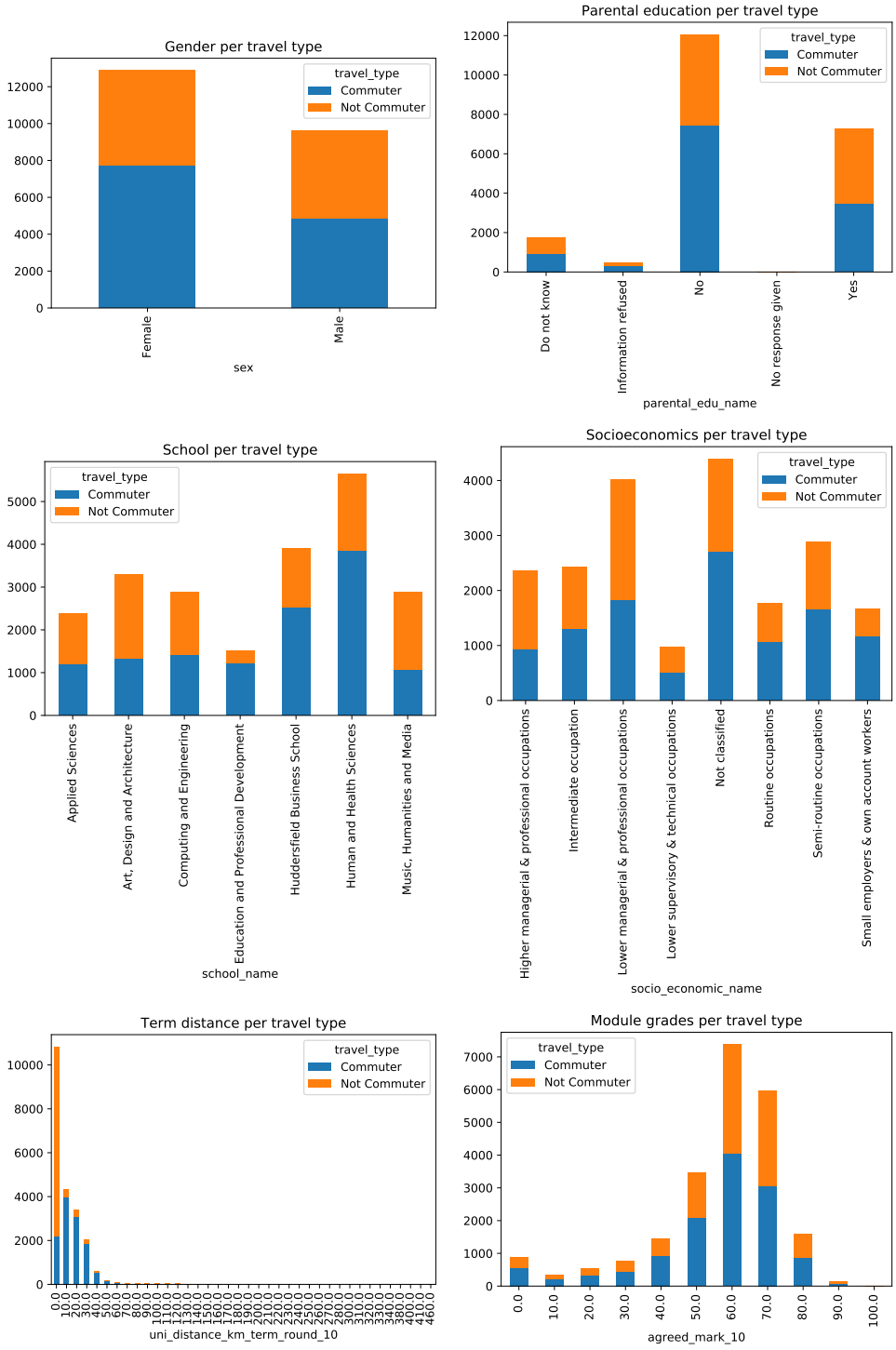

FIGURE 4.9: Ratio of travel types between other relevant variables. The results show the ratio of 50 : 50 repeats occasionally when grouped with other variables. There are exceptions such as term distance with more non-commuters, this is expected as they live nearby.

However, looking into the module grade statistics, the average and medium show most students do not achieve 60% or above. This may mean those students may have fewer opportunities after graduation. It seems the non-graduates are the group that is responsible for this outcome. Avoiding these students returns at least 60%, one sign of this is evaluating the graduate grade classification.

Attendance data shows mediocre performance that is not bad or good. It can be assumed that mediocre attendance could be the culprit for the grade outcomes. This is because attending lectures is a key requirement for all students as the lectures produce the assessments. The correlation relationship between attendance and grades presented in Table 4.10 shows to be the strongest compared to the other variables but it is not above 0.5. Therefore it shows a weak relationship and eliminates the assumption. If both attendance and grades are grouped (and rounded when necessary), one can see the relationship between good/bad attendance and grades. This is demonstrated in Figure 4.10.

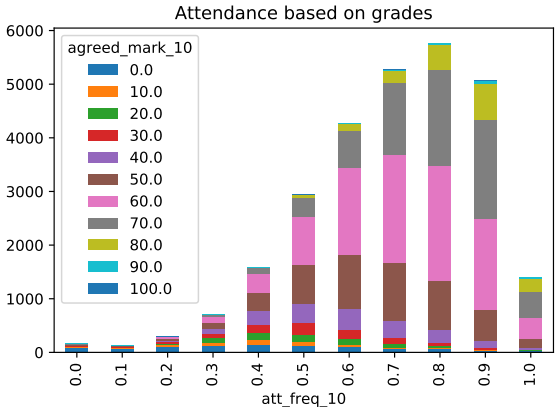

FIGURE 4.10: Number of achieved grades based on attendance. Attendance is between $0-1$ and rounded by 1 decimal place. The module grade is between $0-100$ and rounded by the whole number.

Swipe time shows students attend lectures before the scheduled time, so the time of arrival is what an institution expects and does not provide any harm to the performance of the grade. Of course, swipe time also has an influence on grades, as swiping early means students are aware of all the events that occurred in a lecture. Also, distance has a relationship with grades but in this case, it is probably not because most students live nearby. The swipe time correlation in Table 4.10 shows a weak relationship towards module grades. If it is grouped as presented in Figure 4.11 then one can see the relationship between early/late swipe time and good/bad grades.

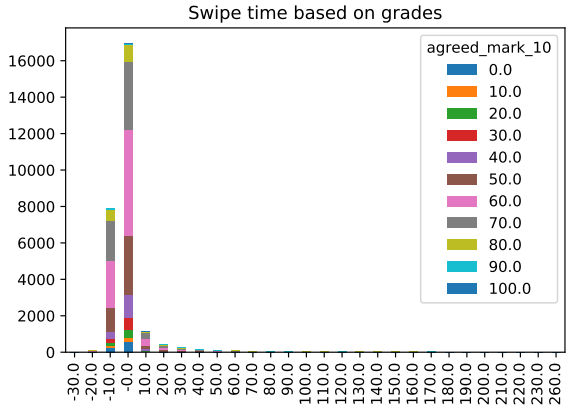

FIGURE 4.11: Number of achieved grades based on each swipe time. Attendance is between $-20 - \text{inf}$ and rounded by 1 decimal place. The module grade is between $0 - 100$ and rounded by the whole number.

Table 4.8 show in-depth statistics with all quantitative variables. The top 5 rows (e.g. maximum) are enough to have a good idea of the variables but further details (e.g. Skew) can be useful to have a deeper understanding. Figure 4.12 displays the distributions of all quantitative variables. The distributions can deliver interesting patterns such as the most/least popular. Some statistics are not presented due to the imbalance of the distribution.

TABLE 4.8: Statistics for quantitative variables.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Modules (Amount)	Age
Mean	56.03	-0.43	0.7	11.42	57.88	7.83	28.15
SD	17.49	15.31	0.19	22.85	271.7	4.15	6.05
Max	95.36	263.06	1	457.92	5948.53	23	85
Min	0	-28.87	0	0	0.31	1	22
Median	59.75	-3.14	0.72	4.91	23.15	6	27
Q1	50.5	-5.42	0.58	0.81	11.31	5	25
Q2	59.75	-3.14	0.72	4.91	23.15	6	27
Q3	67	-0.41	0.84	16.15	43.49	11	29
Q4	83.6	69.9	1	95.42	306.36	18	54
Var	306.02	234.28	0.04	522.07	73820.46	17.25	36.59
IQR	16.5	-	-	-	-	6	4
CV	31.22	-3521.81	27.08	200	469.46	53.06	21.49
Skew	-1.47	-	-	-	-	0.7	2.81
Kurtosis	2.35	-	-	-	-	-0.19	9.75
SEM	0.11	-	-	-	-	0.02	0.04
Moment	0	0	0	0	0	0	0
Mode	60	-4	0.8	0	11	6	26

Quantile analysis shows interesting behaviour between each variable. $Q1$ is 25%, $Q2$ is 50%, $Q3$ is 75%, $Q4$ is 99%. For the module grades, most of the distribution is above 50%, which means most students are on target for Upper-Class Division II. The Q3 (75%) shows students are above 60% or above. The swipe time shows a minor difference between each quantile and they are preferred, it shows that most students attend lectures before it starts. The attendance fluctuates between the quantiles and shows a group of students sensitive to attendance and another group that is mediocre. The distance variables are similar as expected and most students do not live far during their studies. Also, the age statistic shows most students commence their studies during their youth (when applying measurement to the academic year cycles). This is expected generally continue their studies from college. The modules definitely fluctuate as the dataset contains students from different academic cycles

(most students are graduates). The lower number of modules may possibly refer to students who discontinued their studies at an early stage or started their studies. The modules with greater value are from graduates or students in their final academic year.

The variance (Var) and standard deviation (SD) show all variables except attendance fluctuate with no common pattern. The students definitely do have different behaviour and clarify that students come from different backgrounds and circumstances. Attendance is not acting in this manner because students might live nearby. It is already discovered that students live nearby and access to public transport is easy. So the pattern is similar because the method of travel is similar. All variables are not evenly distributed and this is normal. Unfortunately, it is rare to achieve an even distribution. The best case excluding attendance is modules and ages.

The skew for most variables is unable to produce an outcome due to the imbalanced distribution. The module grade is shown to be more on the negative side and shows more bias distribution on the left side rather than the right side. But the bias is quite minor and it is more favorable toward lower grades. The skew for modules and age is reasonable with more biased on the positive side. The mode (popular value) on the other hand shows that 60% is the most popular module grade among students. The attendance mode is 80% which is better than the median and it shows their interest in their studies. The swipe time success is repeated with the mode which is -4, this means most students arrive before the scheduled lesson. Both distances show most students are nearby the university. The minor home distance could be a sign of a large portion of students living with their parents. Please note, the mode's computation involves rounding the variables to the whole number except for attendance. The attendance is rounded by 1 decimal point.

The given analysis delivers interesting details of the dataset. Despite its richness, there are some concerns that could affect its performance. One is the imbalances distribution; unfortunately, this is an issue in general within this application area. Second is the range, the SD and Var show a large range between the variables. The third is data errors, there is obvious unusual activity in the dataset. A perfect example is the Swipe Time, which shows up to a 263 minutes delay to a scheduled lesson. Although practically it is possible it does not make sense to store this value, it could also be a system error. It is evident that the Module Grade is probably the most important variable for this research and it is better compared to most variables.

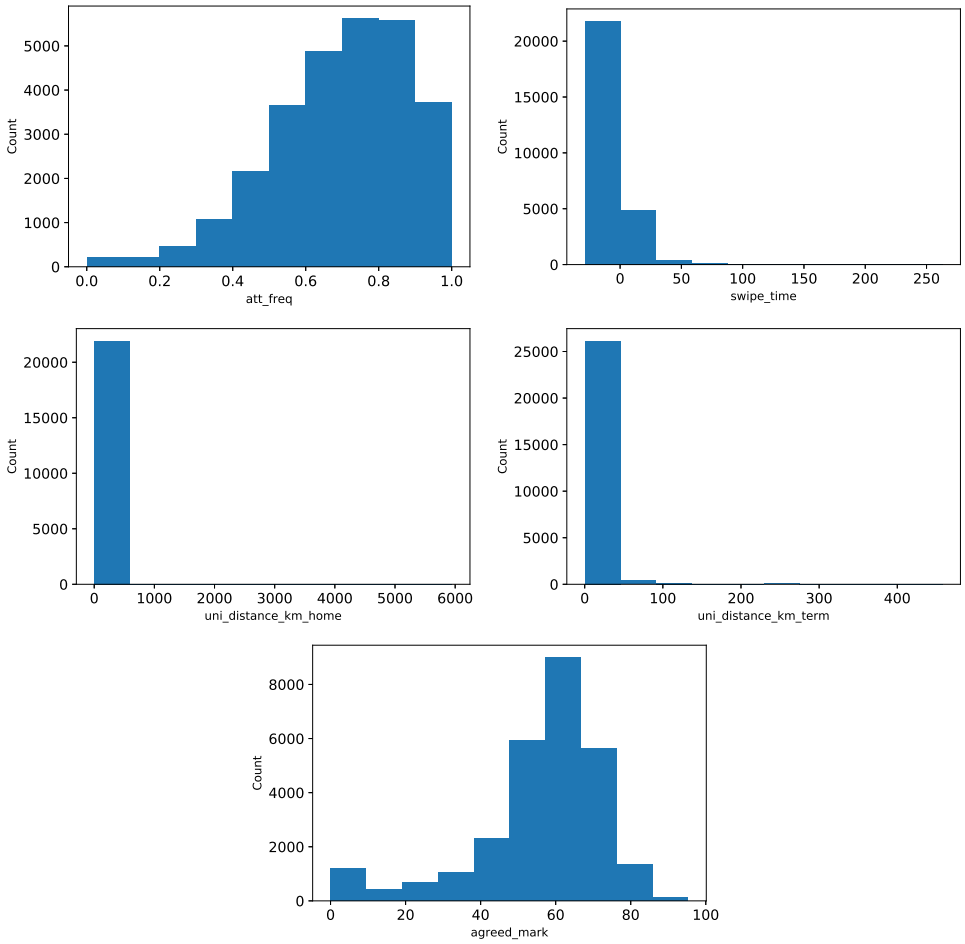

FIGURE 4.12: Quantitative histograms from the dataset (no filters).

Hypothesis

It is interesting to see the normality between quantitative variables. This can be accomplished using null-hypothesis tests such as Shapiro-Wilk. These forms of tests confirm if the assumption of distribution passes a condition. Table 4.9 presents the null-hypothesis test with quantitative variables. Good distribution is great as it avoids unfair outcomes from decision-making models. Unfortunately, with the given variables, most variables reject the null hypothesis tests. This means the distribution does encounter biased behaviour. It is a concern within this area as decision-making models produce their outcomes based on what is given. When biased behaviour occurs, it can affect the decision and direct processes to the wrong routes. The threshold here is 0.05, and the outcomes below the threshold mean the null hypothesis is rejected. The quantitative variables are the only suitable tests worth exploring within this dataset. The remaining form of tests is not useful for this research.

TABLE 4.9: Quantitative null hypothesis tests to verify the normality of distribution. Only specific tests are used due to their suitability, these are mainly one-group tests.

Tests	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Modules (Amount)	Age
Cramér-von Mises	0.0667	9.454E-08	7.504E-10	7.368E-08	6.140E-08	1.9499E-08	6.956E-08
Jarque–Bera	0	0	0	0	0	0	0
Kolmogorov–Smirnov	0.027	0	3.208E-16	0	0	0	0
Kurtosis	2.797E-10	0	2.888E-09	0	3.112E-235	0	0
Shapiro-Wilk	5.957E-12	0	1.489E-34	0	0	0	0
Skew	6.545E-11	0	6.392E-62	0	0	0.0004	0

Correlation

The module grade (variable) is probably the most important variable within education and is widely used in published papers. It is important to verify the relationship between the grades towards other sensible variables. Therefore, Table 4.10 presents the correlation matrix relationship between each variable. The most important variable is module grade, as expected, attendance has the strongest correlation relationship with module grades. Home Distance shows a better correlation relationship towards module grade than term distance. The age, swipe time, and module (amount) variables are shown to have the weakest relationship toward module grade. It also shows most variables have a positive relationship to module grade, but most of the variables are not strong either. Most correlations are neither strong nor weak, this is an indication that most variable combinations might not deliver any good uses for decision-making models. Some variables make sense to have a negative relationship such as attendance and home distance.

It is worth noting that good/bad correlation relationship does not necessarily mean poor outcomes from the decision-making models. Occasionally, the results can show positive signs but bad performance when applied. It also does not mean it evaluates the full aspect relationship between variables.

TABLE 4.10: Display the Pearson correlation matrix to verify the correlation relationship between each quantitative variable. The metric determines any hidden patterns with the variables that can deliver importance to the research.

Columns	Module Grade	Modules (Amount)	Swipe Time	Attendance	Age	Term Distance	Home Distance
Module Grade	1.0	-0.0	0.038	0.388	0.042	-0.007	0.053
Modules (Amount)	-0.0	1.0	0.147	-0.055	0.163	0.047	0.043
Swipe Time	0.038	0.147	1.0	-0.053	-0.004	0.02	0.074
Attendance	0.388	-0.055	-0.053	1.0	0.082	-0.023	-0.021
Age	0.042	0.163	-0.004	0.082	1.0	0.116	-0.082
Term Distance	-0.007	0.047	0.02	-0.023	0.116	1.0	0.033
Home Distance	0.053	0.043	0.074	-0.021	-0.082	0.033	1.0

There are previous studies done that describe the correlation relationship with educational datasets [81], [37], [70]. It shows positive relationships between characteristics and performance but it is also worth noting the outcome depends on the quality of data and resources. The popularity of correlation within this research domain delivers motivation to explore correlation analysis with the given dataset. Of course, the correlation outcomes need to make sense otherwise the results are worthless. In this case, computing correlation with quantitative data types is the appropriate analysis. Although it does not provide the full details between variables, it does provide enough details for sensible future work.

4.2.3 Dimensionality Reduction

The study below demonstrates the data preciseness in lower dimensions. The purpose of these results is to evaluate if the data can be represented, if so, then the decision-making models can perform better. All dimensionality tests involve reducing to 5 dimensions and depending on the type, it returns the loading factor. It displays the performance with and without conditions. The analysis can be represented with all variables or the first 2 dimensions. This is because it is hard to represent all data professionally. The studies below show that it is not represented in lower dimensions and therefore **the data should not be transformed**. It does provide drawbacks and uncertainty regarding the results.

Principal Component Analysis (PCA)

PCA provides the relationship between each variable and principal components. It is possible with this test as the number of quantitative variables is minor. The PCA shows that most data have a weak relationship towards the first 2 dimensions in all aspects but the grade filtering provides more insight into the difference. Figure 4.14 shows many variables tend to have a stronger relationship towards grade classifications. For example, First Class shows the distance (above 60%) variables contain the strongest relationship. But for Third Class, it is modules (above 80%) that contain the strongest relationship. It is difficult to identify the grade classification relationship in general as all contain similar patterns (just allocated to different variables). In addition, Figure 4.15 shows students above 60% have better associations compared to students below 60%. This means students above 60% are more represented in lower dimensions. The distance variables show the strongest relationship with students above 60%.

The following variables are used with PCA:

- Module Grade
- Attendance
- Swipe Time
- Term Distance
- Home Distance
- Age
- Modules (Amount)

TABLE 4.11: Loading factor of each dimension using only quantitative data.

Dimension	Cumulative Variance	Variance %
1	0.21	0.21
2	0.40	0.19
3	0.56	0.16
4	0.7	0.14
5	0.82	0.12

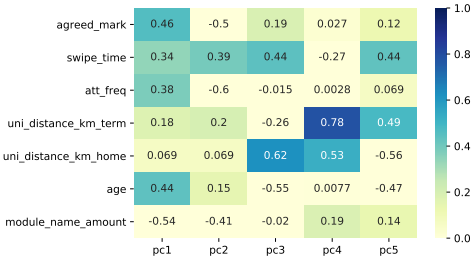

FIGURE 4.13: Relationship between each quantitative variable and principal component (PC_n). The PC_1, PC_2, PC_n are the principal components. The *Column* displays which columns from the dataset are applied.

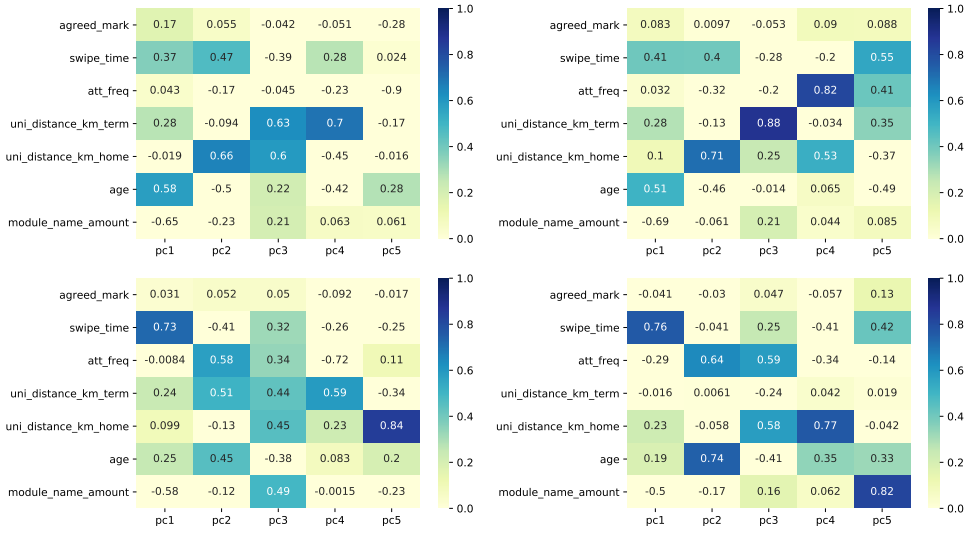

FIGURE 4.14: Relationship between each variable and principal component (PC_n). The PC_1, PC_2, PC_n are the principal components. The *Column* displays which columns from the dataset are applied. The top left is students with First Class, the top right is Upper-Class Division I, the bottom left is students with Upper-Class Division II and the bottom right is students with Third Class.

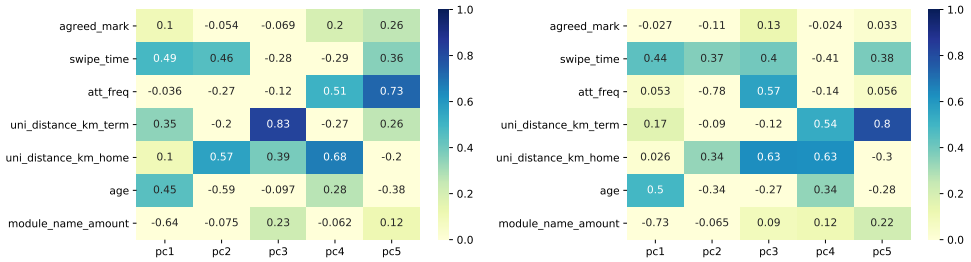

FIGURE 4.15: Relationship between each column and dimension using PCA. The PC_1, PC_2, PC_n are the principal components. The *Column* displays which columns from the dataset are applied. The left is students with 60% or above, the right is students with below 60%.

Figure 4.13 presents a correlation relationship between each quantitative variable against each PC (5 dimensions). The ranges are from $-1 - 1$, where most scores are

on the positive side. The swipe time presents the strongest correlation, in all PC, the variable is on the positive side. The modules present the weakest correlation, in most PC, the variable is roughly in the middle with a slight bias towards the negative side. The best score is from the term distance variables in PC_3. The worst score is from the module variable. PC_1 followed by PC_4 shows the strongest combination of scores, and PC_3 shows the weakest combination of scores. The PC_1 and PC_2 show a strong correlation with all variables except the distance variables. Overall, the relationship is not very strong or weak, with more positive relationships. Therefore, transformation probably does not provide any greater benefit.

Figure 4.14 displays the relationship scores filtered by grade classification. In a nutshell, they all contain similar patterns with different ranges of scores. Most relationships are on the positive side but not so strong. Identifying the best/worst relationship is difficult as they are very similar with minor differences. The common unique pattern is the popular highest values in Upper-Class Division II and Third Class. Having said that, all grade classifications show that transformation delivers more drawbacks than benefits. The module grades are strongest in the First Class, which technically makes sense as First Class grades are awarded to high-achieving students.

Figure 4.15 displays the relationship scores filtered by students above/below 60%. Again, there is not so much unique behaviour between both conditions with an exception of slightly better performance with students above 60%. But the difference is hard to identify. Having said that, above/below 60% shows that transformation delivers more drawbacks than benefits. The module grades are strongest in the above 60%, which technically makes sense as they are high-achieving students.

In addition, Table 4.11 presents the eigenvalues for the first 5 dimensions. With only quantitative variables, the data is not very represented in lower dimensions and suggests that transformation is not a good option for no bias outcomes. But it is worth claiming that the PCA eigenvalue performance is better than the MCA and FAMD. It is possibly the number of variables, the relationship between them, and/or the distributions.

Multiple Correspondence Analysis (MCA)

MCA provides the associations with just the first two dimensions with qualitative variables only. Figures 4.17 and 4.18 show the relationship between each qualitative variable's attribute. It displays just the first 2 dimensions. Each attribute shows no unique pattern between its counterparts. The ranges are between -0.3 to 0.3, which means the attributes are neither biased to one side. Unfortunately, the results do not show interesting aspects between the variables.

The following variables are used with MCA:

- Academic Years
- Gender
- Department
- School
- Graduate Grade
- Travel Type
- Ethnicity

- Entry Qualification
- Parental Education
- Course
- IMD
- POLAR 3 & 4
- Socioeconomic

TABLE 4.12: Loading factor of each dimension using only qualitative data.

Dimension	Cumulative Variance	Variance %
1	0.036	0.036
2	0.068	0.032
3	0.096	0.029
4	0.125	0.029
5	0.153	0.028

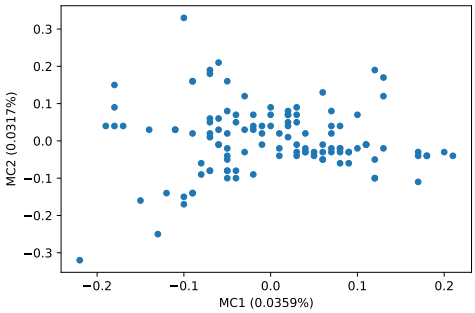

FIGURE 4.16: Relationship of each qualitative variable with MCA. It only presents the first 2 dimensions. Due to the sample amount, plots are represented without labels.

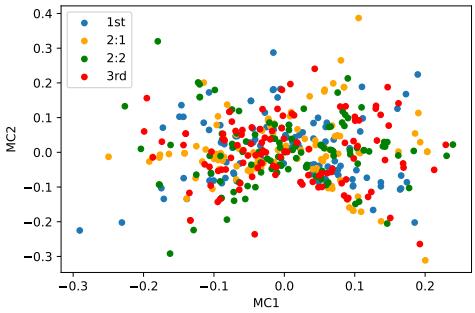

FIGURE 4.17: Relationship of each qualitative variable with MCA between grade classification. It only presents the first 2 dimensions. Due to the sample amount, plots are represented without labels.

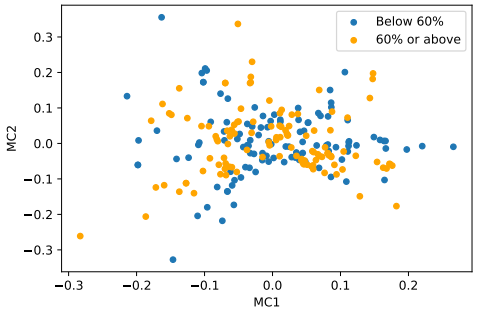

FIGURE 4.18: Relationship of each qualitative variable with MCA above/below 60%. It only presents the first 2 dimensions. Due to the sample amount, plots are represented without labels.

According to Figure 4.16, the vast majority of variables are close to each other with minor relationship differences. The variable shows evenly balanced on both positive and negative sides. What is certain is that all correlations in the first 2 dimensions are not very biased to one side. With no biased relationships, it is uncertain how the variables behave in experiments. It could be a sign of variables not delivering any value to the work. The pattern is repeated with the grade classification and above/below 60% conditions. Each filter does not show any unique differences from its counterparts. These outcomes are presented in Figures 4.17 and 4.18. The relationship between all variables no matter the conditions returns the same outcome.

Table 4.12 presents the eigenvalues for the first 5 dimensions. The variance shows it is not accurately represented in lower dimensions. Therefore, transformation is not a good approach. The high number of variables and the amount of imbalance is probably why it performed poorly.

Factor Analysis of Mixed Data (FAMD)

FAMD provides the loading factor and the associations with just the first two dimensions. Figures 4.20 and 4.21 show the relationship between each qualitative variable's values and each quantitative variable. It displays just the first 2 dimensions. There are some attributes/variables that are away from the majority but it is not enough to show any specific pattern. The ranges are between -0.4 to 0.8, which means the attributes/variables are quite broad. Unfortunately, the results do not show interesting aspects between the variables.

The following variables are used with FAMD:

- Academic Years
- Gender
- Department
- School
- Modules (Amount)
- Graduate Grade
- Travel Type
- Ethnicity

- Entry Qualification
- Parental Education
- Course
- IMD
- POLAR 3 & 4
- Socioeconomic
- Age
- Module Grade
- Attendance
- Swipe Time
- Term Distance
- Home Distance

Dimension	Cumulative Variance	Variance %
1	0.32	0.32
2	0.39	0.07
3	0.44	0.05
4	0.49	0.04
5	0.53	0.04

TABLE 4.13: Loading factor of each dimension using all data types (no conditions).

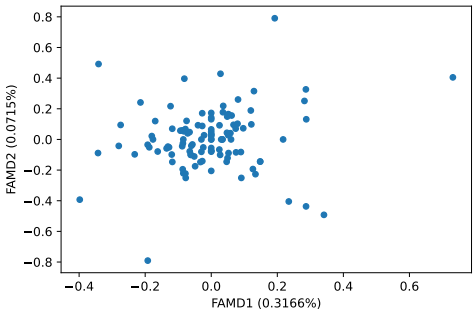

FIGURE 4.19: Relationship of each qualitative and quantitative variable with FAMD. It only presents the first 2 dimensions. Due to the sample amount, it is suitable to represent the plots without labels.

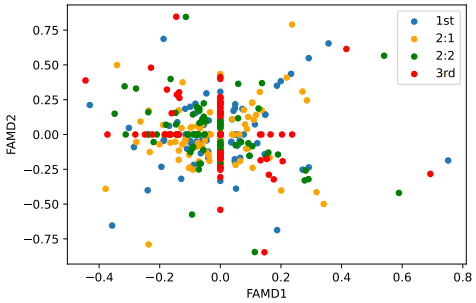

FIGURE 4.20: Relationship of each qualitative and quantitative variable with FAMD between grade classification. It only presents the first 2 dimensions. Due to the sample amount, plots are represented without labels.

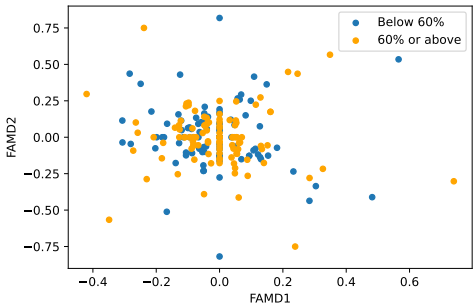

FIGURE 4.21: Relationship of each qualitative and quantitative variable with FAMD above/below 60%. It only presents the first 2 dimensions. Due to the sample amount, plots are represented without labels.

Figure 4.19 shows that nearly all variables are close to each other regardless of their type. The strongest and unique plot is from the ethnicity (Whites) variable. Unlike MCA, the outcome group is shifted towards the positive side. This represents an improvement but even then the relationship is not strong nor weak. Table 4.13 displays the eigenvalues in the first 5 dimensions. Unfortunately, the data is not accurately represented in lower dimensions and suggests a transformation is not a good option to pursue. It is great to see the first 2 dimensions with the highest variance but not enough for accurate outcomes.

Figures 4.20 and 4.21 display a correlation relationship based on grade classifications and above/below 60%. In all cases, the conditions do not provide unique patterns between the students. In the grade classifications filter, the strongest correlation is associated with First Class students. The relationship is major on the positive side than the negative side. On both sides, the relationship is spread out (not clustered together). In the above/below 60% filter, the strongest relationship is associated with students above 60%. The relationship is more spread out on the positive side than the negative side. As a result, it demonstrates that even filtering on performances, transformation is not a useful approach.

Table 4.13 presents the eigenvalues for the first 5 dimensions. The variance shows it is not accurately represented in lower dimensions and therefore, transformation is not a good approach. One can see from all PCA, MCA, and FAMD that as more variables are applied, the preciseness reduces. The number of variables could be the main culprit of the poor outcomes. Perhaps another reason could be the imbalance

ratio from the qualitative variables, and/or non-symmetric distributions from the quantitative variables. A common pattern one can see is the 2 dimensions contain the most variance with the remaining to be either decreasing or remaining the same.

4.2.4 Above/Below 60%

Achieving 60% or more is the intended target for many students as it opens the highest number of opportunities for graduates. Almost all postgraduate courses and careers generally request 60% or above. As a result, performing the in-depth analysis can show the academic performance of all students within the dataset. Again, this considers all students whether or not they are graduated. The 60% is based on the Module Grade variable.

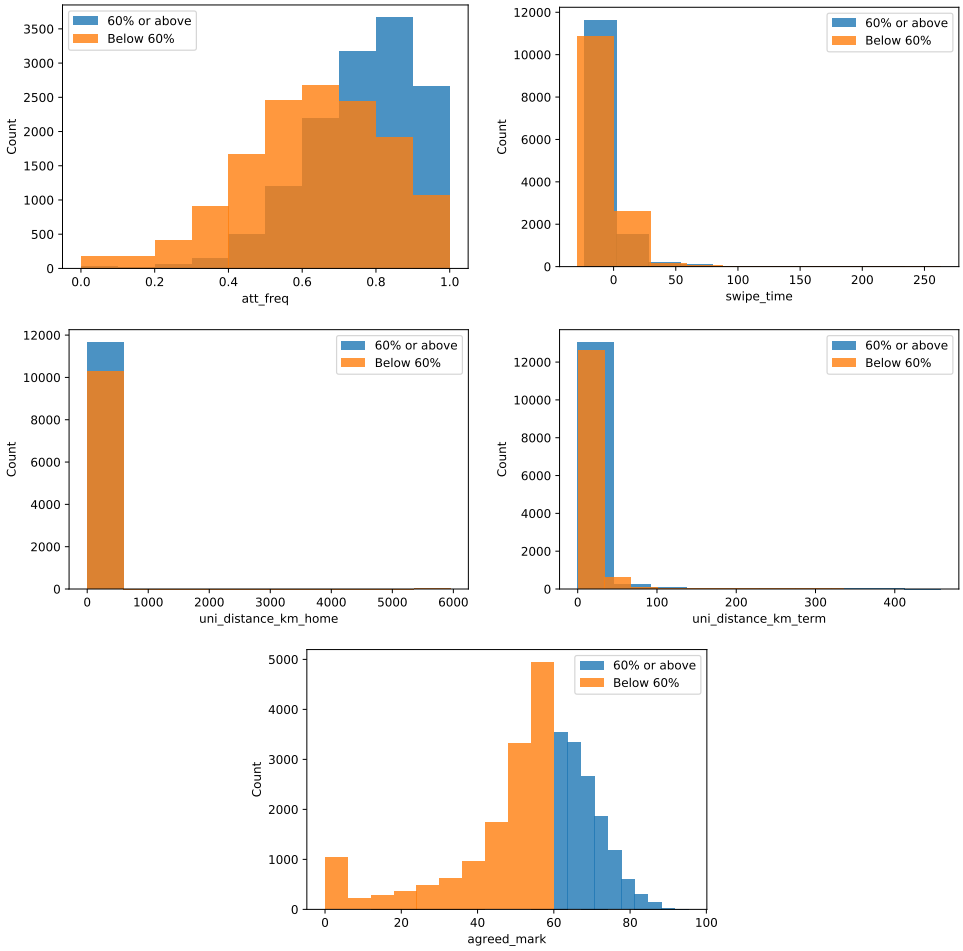

FIGURE 4.22: Quantitative histograms from the dataset that is above/below 60%.

Below 60% Frequencies and Statistics

FIGURE 4.23: Qualitative analysis when the dataset is filtered by module grades below 60%.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	55.0	-2.81	0.64	6.49	22.89	28	9
Max	59.94	201.26	1.0	282.88	337.28	71	23
Min	19.1	-23.27	0.03	0.23	0.57	24	1
SD	5.18	21.67	0.17	18.91	60.89	4.67	4.49
Mean	53.75	2.37	0.64	12.4	43.85	28.64	9.26
Q1	51.25	-5.34	0.52	1.32	11.76	26	5
Q2	55.0	-2.81	0.64	6.49	22.89	28	9
Q3	57.67	0.5	0.76	18.63	39.7	29	13
Q4	59.94	201.26	1.0	282.88	337.28	71	23
Var	26.84	469.46	0.03	357.59	3707.14	21.81	20.19

TABLE 4.14: Quantitative analysis when the dataset is filtered by module grades below 60%.

Above 60% Frequencies and Statistics

FIGURE 4.24: Qualitative analysis when the dataset is filtered by module grades 60% or above.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	67.0	-3.13	0.76	4.93	25.31	28	9
Max	95.25	224.66	1.0	399.59	404.22	65	23
Min	60.0	-19.07	0.1	0.0	0.44	24	1
SD	5.68	22.34	0.15	22.28	64.22	4.94	4.34
Mean	68.03	2.66	0.74	11.69	50.13	28.76	8.94
Q1	63.54	-5.56	0.64	1.09	12.32	26	5
Q2	67.0	-3.13	0.76	4.93	25.31	28	9
Q3	71.4	0.41	0.86	16.68	57.04	29	12
Q4	95.25	224.66	1.0	399.59	404.22	65	23
Var	32.3	499.06	0.02	496.59	4124.74	24.37	18.88

TABLE 4.15: Quantitative analysis when the dataset is filtered by module grades 60% or above.

Figure 4.22 displays the difference of students above/below 60%. In all cases, the students with 60% or above have surpassed all quantitative variables. High-performing students have higher attendance, the common attendance is around 80%. The attendance is lower with the low-performing students, the common attendance is around 50%. The difference shows how attending lectures/tutorial affect their end goal. The high-performing students are attending on time compared to low-performing students. Attending the lectures before the scheduled time shows the relationship to their performance. This is probably because they know all the events that occurred during that occasion. The distances between both home and term residences show that students who live nearby to the institution perform better. The reasons can be several, for example, nearby students have closer contact/support to/from their family. On the other hand, those nearby students might be living at home during their studies rather than in accommodation.

More students are achieving below 60%. Most students above 60% are achieving around the $60\% - 69\%$ range. The best-performing students above 60% are achieving around the $80\% - 100\%$ range. This could mean the students find their studies difficult or they just want the minimum grade possible to pursue their next career/course. Most students below 60% are achieving around the $50\% - 59\%$ range. Least students below 60% are achieving around the $10\% - 19\%$ range. There are unusual patterns occurring near the $0\% - 10\%$. A potential reason can be abandoning their studies or it could be system/human errors. The reasons could be they find the course difficult or they lose interest in the course.

Figures 4.23 and 4.24 show the distributions of several variables. If one investigates the differences, the entry qualification, IMD and ethnicity show the major differences. Students that are above 60% show more students in fewer deprived areas (using IMD) compared to those that are below 60%. This means the deprivation of their home residences shows a lack of support for the individual and results in a lower outcome. Students who are above 60% show fewer students from all ethnicity excluding Whites. The distribution shows people who are identified as White are more likely to achieve the national grade (or above). One can make an assumption that most students that are achieving good marks are likely to be Whites. The entry requirement shows most students achieving above 60% are from an A-Level background compared to students below 60%. That is not to say that another entry qualification is not useful but due to the preparation for A-Level, it develops a better foundation. There is also a slight increase of females achieving above 60% than below. Tables 4.14 and 4.15 express more statistics to a range of variables. Looking closer at the statistics, the difference is not huge but it shows better performance with students above 60%. For example, students above 60% attend lectures early according to the median but the difference is minor.

4.2.5 Grade Classification

It is worth noting that 60% in grade classification is Upper-Class Division I. Table 4.16 displays the grade classification equivalent in percentage. Not all students have a grade classification as they are not completed their registered course.

TABLE 4.16: Percentage equivalent to each grade classification (rounded by the whole number).

Classification	Grade %	Other Names
First Class	70% – 100%	1st
Upper Class Division I	60% – 69%	2:1
Upper Class Division II	50% – 59%	2:2
Third Class	40% – 49%	3rd
Fail	0% – 39%	-

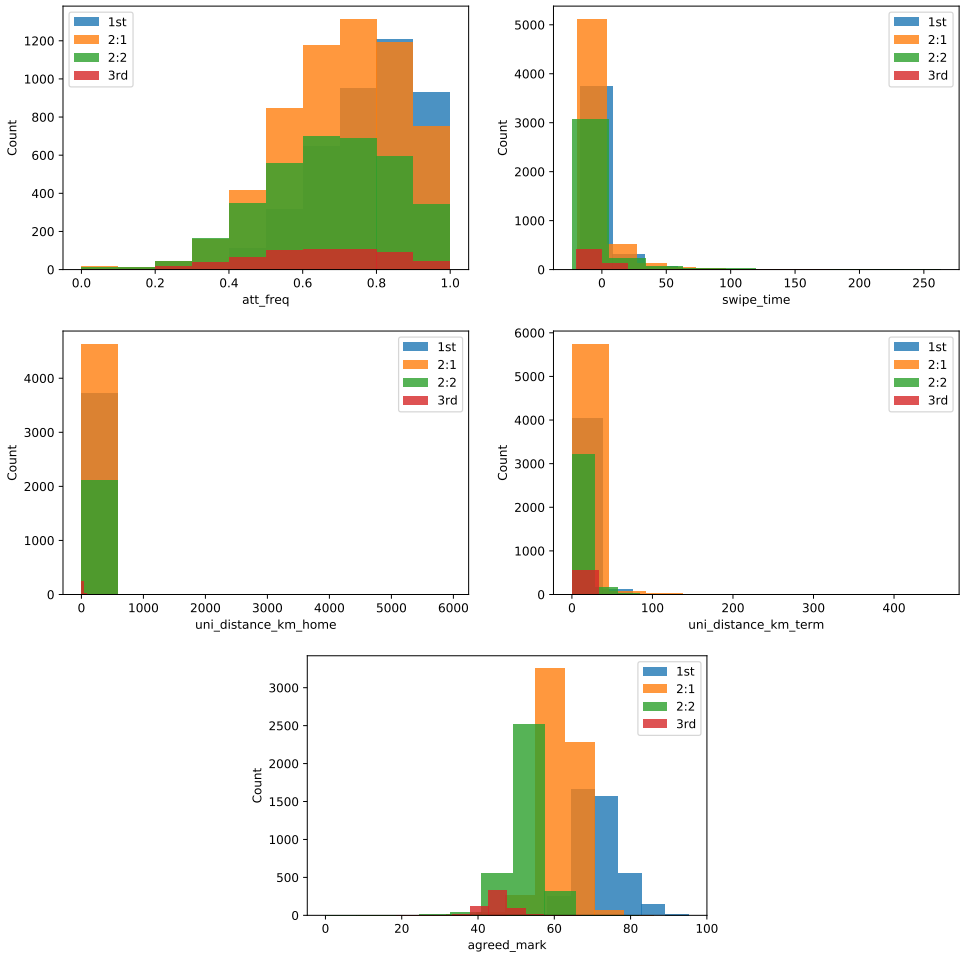

FIGURE 4.25: Quantitative histograms from the dataset based on each grade classification.

First Class Frequencies and Statistics

FIGURE 4.26: Qualitative analysis when the dataset is filtered by First Class.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	70.89	-3.22	0.8	4.94	25.22	28	10
Max	95.25	224.66	1.0	380.39	383.3	62	23
Min	33.5	-16.65	0.1	0.0	0.44	24	1
SD	5.65	21.11	0.14	20.71	64.89	5.11	4.38
Mean	71.4	2.2	0.78	11.3	50.87	29.01	9.18
Q1	67.83	-5.65	0.69	1.08	12.09	26	5
Q2	70.89	-3.22	0.8	4.94	25.22	28	10
Q3	74.67	0.2	0.89	16.3	59.83	29	13
Q4	95.25	224.66	1.0	380.39	383.3	62	23
Var	31.91	445.77	0.02	428.96	4211.36	26.07	19.17

TABLE 4.17: Quantitative analysis when the dataset is filtered by First Class.

Upper Class Division I Frequencies and Statistics

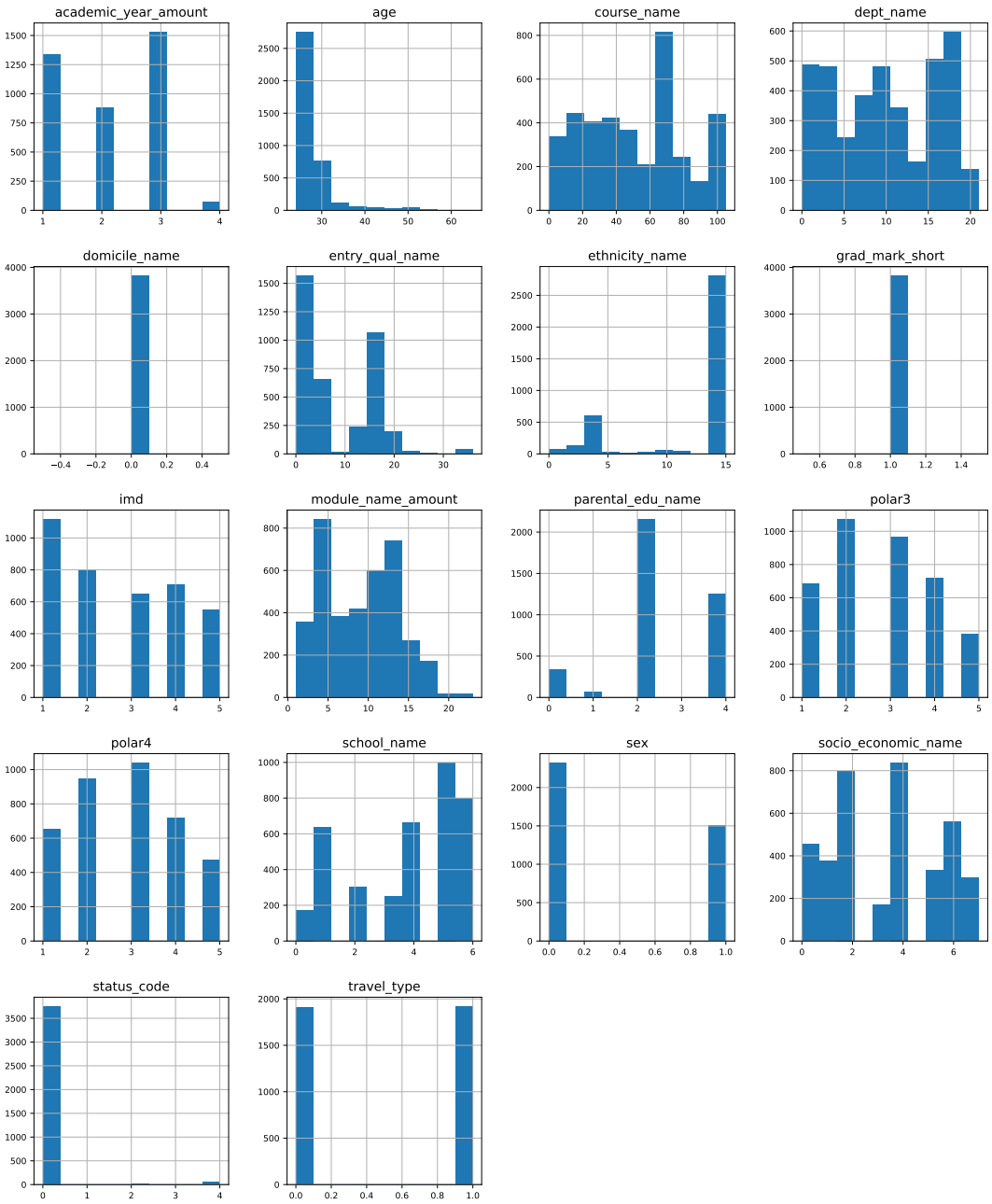

FIGURE 4.27: Qualitative analysis when the dataset is filtered by Upper-Class Division I Class.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	61.92	-3.05	0.7	4.94	24.95	27	9
Max	77.0	211.36	1.0	399.59	404.22	65	23
Min	24.6	-19.07	0.11	0.1	0.46	24	1
SD	4.55	21.28	0.16	21.59	63.6	4.75	4.37
Mean	61.67	2.09	0.69	11.96	48.43	28.5	9.04
Q1	59.0	-5.46	0.58	1.16	12.34	26	5
Q2	61.92	-3.05	0.7	4.94	24.95	27	9
Q3	64.79	0.17	0.81	17.93	48.27	29	12
Q4	77.0	211.36	1.0	399.59	404.22	65	23
Var	20.66	452.79	0.02	466.17	4045.39	22.59	19.13

TABLE 4.18: Quantitative analysis when the dataset is filtered by Upper-Class Division I Class.

Upper Class Division II Frequencies and Statistics

FIGURE 4.28: Qualitative analysis when the dataset is filtered by Upper-Class Division II Class.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	53.31	-2.59	0.63	6.95	22.42	28	9
Max	70.0	193.38	1.0	282.88	332.2	71	23
Min	22.4	-23.27	0.03	0.23	0.85	24	1
SD	4.45	25.0	0.17	21.5	58.5	4.43	4.51
Mean	52.95	3.89	0.62	12.94	41.99	28.61	8.89
Q1	50.6	-5.29	0.51	1.31	11.78	26	5
Q2	53.31	-2.59	0.63	6.95	22.42	28	9
Q3	55.8	0.87	0.74	18.68	37.87	29	12
Q4	70.0	193.38	1.0	282.88	332.2	71	23
Var	19.83	624.75	0.03	462.42	3422.35	19.63	20.34

TABLE 4.19: Quantitative analysis when the dataset is filtered by Upper-Class Division II Class.

Third Class Frequencies and Statistics

FIGURE 4.29: Qualitative analysis when the dataset is filtered by Third Class.

Statistics	Module Grade	Swipe Time	Attendance	Term Distance	Home Distance	Age	Modules (Amount)
Medium	45.8	-1.75	0.58	9.44	22.22	28	8
Max	67.0	176.49	1.0	127.99	305.77	65	22
Min	22.25	-19.61	0.17	0.23	0.72	24	1
SD	4.48	25.48	0.18	15.72	61.34	5.31	4.34
Mean	45.4	5.42	0.59	12.92	43.04	29.15	8.71
Q1	43.08	-4.88	0.45	1.71	10.97	26	5
Q2	45.8	-1.75	0.58	9.44	22.22	28	8
Q3	47.71	3.55	0.72	19.54	39.71	29.5	11
Q4	67.0	176.49	1.0	127.99	305.77	65	22
Var	20.06	649.41	0.03	247.05	3763.18	28.25	18.87

TABLE 4.20: Quantitative analysis when the dataset is filtered by Third Class.

Figure 4.25 displays the difference between all grade classifications. In all cases, students with First Class grade classification have the best outcomes. As the grade classification increases, attendance increases. Students from a First Class have more students with attendance near 100%. More students with low-grade classifications have low attendance. As the lecturers are the assessment makers and markers, avoiding lectures deliver more misunderstandings about the subject. This results in falling behind compared to their peers. The swipe time shows the majority of students from First Class arrive at lectures before the scheduled time. The Upper-Class Division I have more students but that may be because of the imbalance among those students. As the grades decrease, there is a wider range of time delays from the scheduled time to the student's arrival. Most students achieving Upper-Class Division I or above live nearby. There are more students from lower grade classifications from far distances. A possible reason can be the effect of traveling. Traveling far to a destination can affect concentration during studies. Living close is also convenient as a student can have more assistance from staff. The outcome is similar to the analysis done with just students above/below 60%.

Figures 4.26, 4.27, 4.28 and 4.29 show the distributions of several variables. The department variable in all grade classifications shows different outcomes, since the majority of students are from the Business department, it is likely to assume that most students in all grade classifications are from the Business area. In each department, the course difficulty differs and therefore could result in a bias of one subject area maintaining higher grades. Perhaps most students with an Upper-Class Division I or above are more likely from a none science background. One can see this repeat with the school variable, one is more biased compared to the remaining schools. The entry qualification variable show that most students accomplishing higher grade classification are from an A-Level background. As the grade classification increases, the number of students from other entry qualifications starts to decrease. Perhaps the experience of A-Level assessment preparation is best suited to high-achieving grades. Many variables show that increasing the grade classification results in one group having more bias than the remaining. Gender is one variable, as the grade classification increases, one gender (females) is more biased. Starting from Third Class, there is a more equal proportion between the genders and then ends up imbalanced to one group. Another variable is ethnicity, starting from Third Class, there is more fluctuation among all ethnicity, and as the grade increases, it becomes biased toward one ethnicity (Whites). Tables 4.17, 4.18, 4.19 and 4.20 express more statistics to a range of variables. Looking closer at the statistics, the difference is not huge but it shows better performance with First Class. For example, students with First Class show more completed modules. This means the students are more engaged in their studies.

4.2.6 Date of Birth (D.O.B.)

In this section, an analysis based on D.O.B. is presented. Figure 4.30 presents the number of students based on different time intervals. The distributions show no unique patterns in the dataset. There is some imbalance such as the months and weeks. Exploring D.O.B. may not bring value to the research but it definitely contains a few unique patterns. One pattern is the months, most students are from January and December which is out of the ordinary. There may be a valid reason or it could be just a coincidence. Most students are from the 1900*s* but some students go back to the 1950*s* (it could be a system error or not). Theoretically, there should not be any patterns (from the D.O.B.) & relationship to student performance. Also, no past studies show any evidence (at this current time).

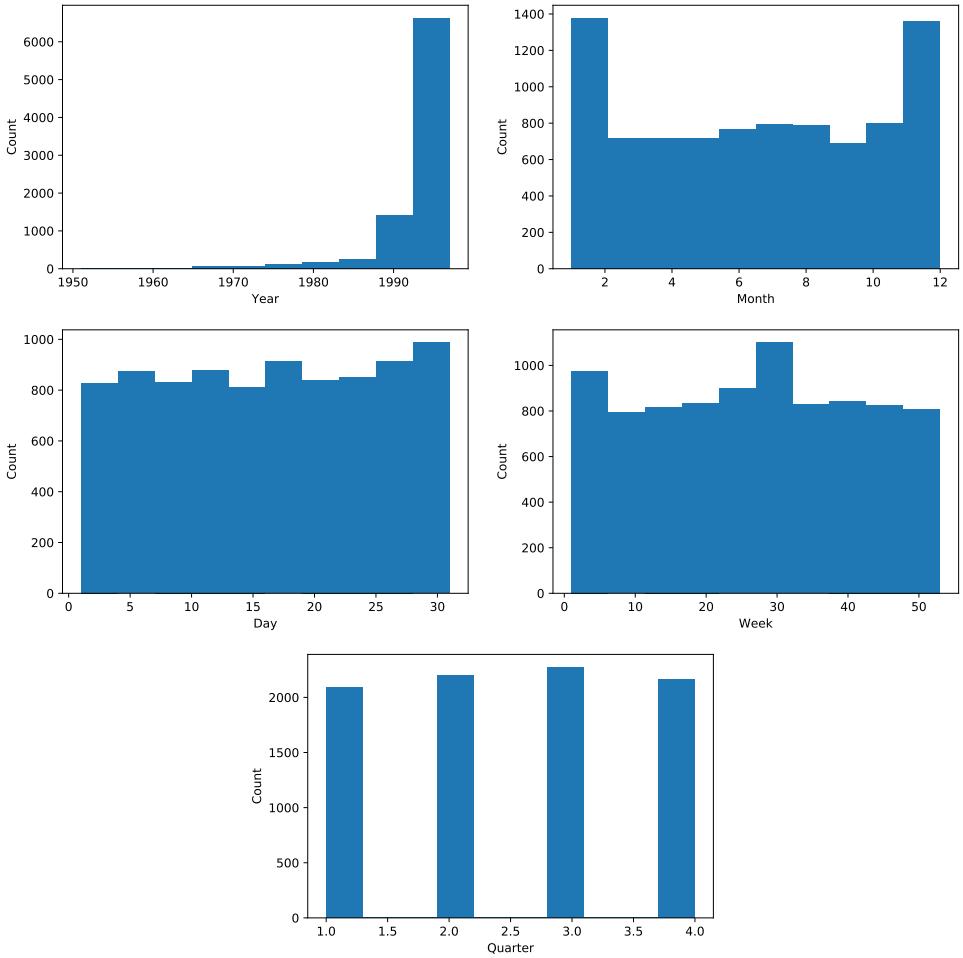

FIGURE 4.30: Atudent histograms based on D.O.B.). The graphs are filtered with different time intervals such as month and year.

4.2.7 Discussion

Data analysis is an important part to understand the condition of a dataset. Past works have used data analysis to evaluate the condition of their datasets. These are used to understand the relationship between variables. Typical examples of statistics explored are correlation and dimensional reduction. The common statistics include FAMD (Pereira [64]), & Pearson (Zacharis [87]). In rare cases, other forms of data analysis are applied such as Single T-Test ANOVA (Lau, Sun, and Yang [53]).

Pearson correlation is used in the vast majority of analyses. Oftentimes, datasets in this field are mainly qualitative and the main focus on performance is generally quantitative. There are other forms of correlation like Kendall & Spearman but mathematically because the data types are not eligible, it's not applied. Using dimensional reduction to identify patterns in student performance really depends on the data type except for FAMD (it's portable). It can also be used to verify if the dataset can be accurately represented in lower dimensions. Doing so can help improve the prediction accuracy of the decision-making models.

Even though there are past works that embrace statistical analysis but the research in relation to predicting student performances doesn't explore many kinds of data analysis. The area is quite limited in the exploration of (deep) data analysis. Normally, the data specification is written such as their samples, attributes, and variables. If one look at papers such as Aguiar et al. [6], Zhang and Rangwala [88], Kotter et al. [51], many generic statistical analyses are used. The generic analysis includes data origin, exploring high/low performing students, correlation, & a number of attributes/samples/variables with bar charts, heatmaps & histograms. Other than that, it is rare to find other types.

The best-case scenario with sufficient data analysis is Pereira [64] (to the author's knowledge) that applies many forms of analysis such as hypothesis tests, & FAMD. But still, many missing analyses are not explored such as exploring other types of dimensional reductions (e.g. PCA & MCA), deeper statistics with quantitative variables (e.g. quantiles, skew), exploring (hidden) patterns between students (e.g. grades), and using several hypothesis tests (if compatible).

Unlike statistical analysis in past works, the novelty here is the wide variety of analyses. The analysis in this research (thesis) includes exploring attributes from qualitative variables, statistics (e.g. median, skew), correlation with quantitative variables, distribution hypothesis tests with quantitative variables, exploring patterns between high/low performing students, exploring potential patterns with D.O.B, and dimensional reduction with several types.

The number of attributes alongside their percentage of popularity is presented to determine the balance between them. The statistics do not just explore generic ones (e.g. mean & max) but also quantiles (Q1, Q2, Q3, Q4), kurtosis, mode, variance, and much more.

Although past work provides analyses between high-performing and low-performing students, it is expressed briefly. In this thesis, several student groups are explored in greater detail. The student groups include D.O.B (e.g. weekdays, weeks, months) and grades (classification grades & module grades). This includes exploring patterns with several variables within these student groups such as their attendance, time of arrival in scheduled lectures, (home/term) distance to the university, & type of travel.

The dimensional reduction novelty is exploring several types (PCA, MCA & FAMD) to evaluate patterns between students with and without certain data types. Then compare the patterns in all tests (e.g. difference between them, the best and worst performing attributes/variables, the reliability in lower dimensions, and the weak and strong data types). Also, verify if the data can be accurately represented in lower dimensions (individually). If so, then it can be useful when predicting student performance.

Overall, the analysis expressed in this thesis is in greater detail and contains new types of analyses that have not been explored in past works.

4.3 Predict Pre-Processing

In addition to the Global Pre-Processing, here are the student group and variables used for this experiment:

- Only graduate students: This is done by collecting the rows that have a degree classification.
- Only enrollment characteristics: The audience is students that assume to be first-time enrolling. This means students that are not undergraduates such as college students that are finishing off their A Levels (or equivalent).

This particular student group is chosen due to the large majority of them existing in the dataset. The same applies to the selection of the variables, most variables (enrollment characteristics) suit very well with this student group. Again, this is a preference, not a requirement for the framework.

The prediction produced from this experiment represents their overall grade achievement (boundaries) at the end of their course (final year).

The total list of features (variables) is presented below:

- Student (ID)
- Socioeconomic
- Entry Qualification
- Gender
- Travel Type
- Last School
- Parental Education
- Term Distance (KM)
- Home Distance (KM)
- Age
- Ethnicity
- IMD
- POLAR 3
- POLAR 4
- Consistency-Scale (Minimum, Average, Maximum)

Once the above conditions are met, then the remaining Predict Pre-Processing requirements are executed (Section 3.3).

4.4 Feature Selection

From an educational perspective, the effectiveness of applying feature selection is dependent on many factors such as data (e.g. samples, variables, size) and methodology (pipeline process). In these types of research, the purpose is to help improve the prediction accuracy of machine learning algorithms. This is done by collecting a combination of features that have a positive correlation to student (academic) performance. However, feature selection can harm the benchmark performance as it can return overfitting/underfitting issues (or similar). In this case, not applying feature selection is a better strategy and can deliver other benefits such as shorter execution time.

In the existing literature, feature selection algorithms are uncommon. It is considered an optional layer to their methodologies. Examples include using feature variance (Li, Lynch, and Barnes [55]) and coefficient correlation (Badr et al. [10], Hussain et al. [45]). In rare cases, a paper applies a unique algorithm for feature selection (Hussain and Khan [46]).

Past papers apply one feature selection algorithm to their methodology. The novelty here is using a combination of feature selection algorithms in one setting. Doing so enables a deeper analysis of its usefulness. This is important because feature selection requires more execution time & can affect the prediction performance (e.g. overfitting/underfitting) from machine learning algorithms. Also, the Particle Swarm Optimisation for feature selection has not been explored before which also delivers some novelty.

Note that there is also the option of not performing any feature selection, as discussed in the previous chapter.

Tables 4.21, 4.22, 4.23 present the configurations for the Feature Selections.

TABLE 4.21: Configuration of the Genetic Algorithm.

Configuration	Value
Population	10
Generation	2
Model	Logistic Regression

TABLE 4.22: Configuration of the Particle Swarm Optimisation.

Configuration	Value
Iterations	10
Particles	50
Model	Logistic Regression

TABLE 4.23: Configuration of the Recursive Feature Elimination.

Configuration	Value
Model	Logistic Regression
Step	1
CV (Folds)	5

4.4.1 Feature Selection For Parallel Architecture

Decision-making model(s) that involve features and a target apply the same Feature Selection (due to their specification).

Table 4.24 displays the total number of variables. The experiment excluding the Feature Selection includes all features available.

TABLE 4.24: The list of features and target. The Consistency-Scale (T) is the (target) predictor. Each row represents the data of each student. $F1$, $F2$, ..., F_n. I is the index column (student identity).

Columns	Row 1	Row 2	Row 3	Row 4
Student (I)	ID 1	ID 2	ID 3	ID 4
Socioeconomic ($F1$)	1	2	2	2
Entry Qualification ($F2$)	3	3	3	2
Gender ($F3$)	1	0	1	1
Travel Type ($F4$)	1	0	1	1
Last School ($F5$)	1	2	3	4
Parental Education ($F6$)	1	2	1	1
Term Distance (KM) ($F7$)	140	78	15	21
Home Distance (KM) ($F8$)	200	160	190	80
Age ($F9$)	1	2	1	1
Ethnicity ($F10$)	1	2	1	1
IMD ($F11$)	1	2	1	1
POLAR 3 ($F12$)	1	2	1	3
POLAR 4 ($F13$)	1	2	1	4
Consistency-Scale (T)	7	5	3	9

The following variables are chosen using the Genetic Algorithm Feature Selection:

- Term Distance (KM)
- Home Distance (KM)
- Socioeconomic
- Age
- Travel Type
- Last School
- POLAR 3
- POLAR 4
- Ethnicity
- IMD

The following variables are chosen using the Particle Swarm Optimisation Feature Selection:

- POLAR 3
- POLAR 4

- Ethnicity
- IMD
- Term Distance (KM)
- Entry Qualification
- Gender
- Travel Type

The following variables are chosen using the Recursive Feature Elimination Feature Selection:

- Socioeconomic
- Entry Qualification
- Gender
- Travel Type
- Last School
- POLAR 4
- Age
- Ethnicity
- IMD
- Parental Education

TABLE 4.25: The total number of times each feature is selected by the Feature Selection techniques.

Columns	Count
Age	2
Entry Qualification	2
Ethnicity	3
IMD	3
Last School	2
Parental Education	1
POLAR 3	2
POLAR 4	3
Gender	2
Socioeconomic	2
Travel Type	3
Term Distance (KM)	2
Home Distance (KM)	1

4.4.2 Feature Selection For Popularity Architecture

Decision-making model(s) that involve (only) features apply the same Feature Selection (due to their specification).

Table 4.26 displays the total number of variables. The experiment excluding the Feature Selection includes all features available.

TABLE 4.26: The list of features. Each row represents the data of each student. $F1$, $F2$, ..., F_n. I is the index column (student identity).

Columns	Row 1	Row 2	Row 3	Row 4
Student (I)	ID 1	ID 2	ID 3	ID 4
Socioeconomic ($F1$)	1	2	2	2
Entry Qualification ($F2$)	3	3	3	2
Gender ($F3$)	1	0	1	1
Travel Type ($F4$)	1	0	1	1
Last School ($F5$)	1	2	3	4
Parental Education ($F6$)	1	2	1	1
Term Distance (KM) ($F7$)	140	78	15	21
Home Distance (KM) ($F8$)	200	160	190	80
Age ($F9$)	1	2	1	1
Ethnicity ($F10$)	1	2	1	1
IMD ($F11$)	1	2	1	1
POLAR 3 ($F12$)	1	2	1	3
POLAR 4 ($F13$)	1	2	1	4

The following variables are chosen using the Genetic Algorithm Feature Selection:

- POLAR 4
- IMD
- Parental Education

The following variables are chosen using the Particle Swarm Optimisation Feature Selection:

- Last School
- Term Distance (KM)
- Home Distance (KM)
- Gender
- Travel Type
- POLAR 3
- POLAR 4
- Ethnicity
- IMD

The following variables are chosen using the Recursive Feature Elimination Feature Selection:

- Term Distance (KM)
- Home Distance (KM)
- Gender
- Travel Type
- POLAR 3
- POLAR 4
- Age
- Ethnicity
- IMD
- Parental Education

TABLE 4.27: Total number of times each feature is selected by the Feature Selection techniques.

Columns	Count
Age	1
Ethnicity	2
IMD	3
Last School	1
Parental Education	2
POLAR 3	2
POLAR 4	3
Gender	2
Travel Type	2
Term Distance (KM)	2
Home Distance (KM)	2

4.5 Training/Testing Split and Anchored Training Data

In total, the dataset D contains 10785 students for this experiment (after the data-cleaning steps). The training set D_{train} contains 7225 students and the testing set D_{test} contains 3560 students. The chosen ratio for this experiment is 67% : 33% ($D_{train} : D_{test}$).

From an educational perspective, the effectiveness of applying methods to reduce imbalanced data is dependent on the dataset. The dataset must contain sufficient samples, otherwise, after balancing, the number of samples is not sufficient for the decision-making algorithms. This can increase the chances of overfitting/underfitting issues. If this is not the case, then balancing the dataset is beneficial (and successful). Imbalanced data is a common issue with educational datasets. It is impossible to fully remove the issue, the best-case scenario is reducing the probability. Doing so results in a lower chance of overfitting/underfitting issues. Therefore, this increases the prediction accuracy of decision-making models.

In the existing literature, resolving imbalanced data is lacking despite its popularity. The past papers that attempt to tackle the issue include using dimensional

reduction (Yang et al. [86], Pereira [64]) or feature selection algorithms (Hussain and Khan [46], Ding et al., [27], Li, Lynch, and Barnes [55]).

Past papers do not include sufficient solutions for tackling imbalanced data. The novelty here is to present a unique algorithm that determines the right number of samples. Collecting the right number of samples decreases the probability of imbalanced data, and overfitting/underfitting issues. Therefore, increases the probability of accurate predictions. Again, reducing the probability by 1% is better than nothing.

Tables 4.28 & 4.29 showcase the number of samples applied to predict student performance.

4.5.1 Sample Amount For Parallel Architecture

TABLE 4.28: Display the total number of samples used for Parallel Architecture. The models are trained with D_{train} to predict the performance of each student in D_{test}. Also, the operation is executed in parallel to define the minimum/maximum performance.

Feature Selection	D_{train}	D_{test}	Range
Genetic Algorithm	1445	3560	Min
Particle Swarm Optimisation	2890	3560	Min
Recursive Feature Elimination	2890	3560	Min
None	2890	3560	Min
Genetic Algorithm	7225	3560	Max
Particle Swarm Optimisation	4335	3560	Max
Recursive Feature Elimination	7225	3560	Max
None	7225	3560	Max

4.5.2 Sample Amount For Popularity Architecture

TABLE 4.29: Display the total number of samples used for Popularity Architecture. The models are trained with D_{train} to predict the performance of each student in D_{test}. Also, the operation is executed standalone which outputs both the minimum/maximum using the same grouping identity.

Feature Selection	D_{train}	D_{test}
Genetic Algorithm	7225	3560
Particle Swarm Optimisation	2890	3560
Recursive Feature Elimination	4335	3560
None	4335	3560

4.6 Model

From an educational perspective, the effectiveness of AI technologies is popular & successful in past works due to their intelligent algorithms. It can foresee the future using current and/or past events better. Compared to other approaches (e.g. manually produced by humans), they are accurate, efficient, fast, solve more complex problems, are able to explore more hidden patterns, are easily integrable with databases (time-efficient, more secure), and are less likely to have errors. Therefore, the results produced from AI techniques are more likely reliable. In addition, it is more practical and cost-efficient commercially. The benefits of AI in education

outweigh the losses (e.g. data cleaning, imbalanced data, coding). So far, past works have shown AI approaches do provide accurate predictions in most cases ([39], [46]). Since the benchmark performance is more positive (than negative), they are repeatedly going to be applied (including this experiment).

The success ratio of AI technologies returning accurate predictions depends on the data quality (e.g. sample size, variables) applied. That is why data analysis is conducted beforehand (with the given dataset) in this experiment. This allows one to determine the best student group that decreases the chance of overfitting/underfitting. Also, the AI algorithm specification must be compatible with the data (e.g. type, variables) and problem, otherwise, the outcomes are unreliable.

In the existing literature, AI techniques are mostly used to evaluate student academic performances in future events ([68], [69]). The vast majority of the time, Supervised Learning models are used ([2], [87]). Other types of AI technologies (e.g. NLP) are uncommon for student performance predictions due to their incompatibility with the problem. But it is used in other problems related to AI in education ([7], [54]). It is uncommon to find past papers that don't apply AI technologies (for student performance predictions). These uncommon past papers generally apply statistics such as correlation for their evaluations ([64], [53]).

For this research, machine learning approaches are used to predict student performance. The motivation is derived from the popularity & successes of its uses in past works. This identifies high & low performing students in future occasions (e.g. the next semester). Knowing their performances in advance allows institutions to understand the level of support required to ensure high-grade achievements from all students. Identifying low-performing students is more important than high-performing students. This is because it affects both students and institutions in negative ways. For students, it affects career opportunities after course completion (e.g. fewer career options). For institutions, it affects their reputation which may result in future students not enrolling. Although high-performing students may not deliver any risks, it does provide insights into which of them do not require additional support. Therefore, students that are not on track have more assistance & resources from staff.

Supervised Learning (excluding Feed Forward Neural Networks) is the most popular family for predicting student performance (Yang et al. [86], Aggarwal, Mittal, and Bali [5], Mueen, Zafar, and Manzoor [61]). The common practice includes using either old decision-making algorithms or using a mixture of old and recent ones that are heavily biased toward one side (normally to older models). The novelty here is applying a more balanced ratio of old & recent models in one experiment (setting). Another novelty can be the number of models applied, which is 8. The maximum amount found in past works is 7 (Rodríguez-Hernández et al. [69]). This can also be said with the hyperparameters applied to models (e.g. L1/L2 regularisation, iterations, trees, and neighbours) which contain a unique combination of configurations. The combination of Supervised Learning algorithms with other families (e.g. Unsupervised Learning) in one experiment delivers some novelty as well.

A better-balanced ratio of old & recent models can uncover more hidden patterns. They could perform better than the well-known models. Having more models tested with the same data and applying the same process deliver fairer observation when analysing the benchmarks. This is better than comparing the prediction accuracy of models from several papers. Examples of popular models include AdaBoost, & Random Forest, and examples of unpopular models include Gradient Boost & Extra Trees (Alwarthan, Aslam, and Khan [8]).

Feed Forward Neural Networks (excluding other Supervised Learning) have been used to predict student performance (Bilal et al. [18], Mengash [60], Zhao et al. [89]). The common approach (in past papers) is applying (only) one architecture that generally consists of an input layer, $1-2$ hidden layers (neurons/nodes based on preference), and an output layer (1 node). The concern here is the lack of architectures with different hidden layers in one setting (experiment). The novelty of (Feed Forward) Neural Networks in this experiment is applying several architectures with different hidden layers. It verifies if hidden layers provide any benefit to the prediction accuracy. For fairer observation, the models should be similar as possible (e.g. neurons/nodes). Having more than 1 Feed Forward Neural Network set of the following requirement with another set of nodes is an added bonus of novelty.

Unsupervised Learning has been used before as a pre-processing step for predicting performance in rare cases (Alwarthan, Aslam, and Khan [8]). The novelty of Unsupervised Learning in this experiment is introducing a model pipeline process to assist student performance predictions. Due to the clustering nature, it is treated as pre-processing step. The usage is collecting relevant students based on their information (before producing predictions). The prediction accuracy of this family may outperform other families (e.g. Supervised Learning). Also, it should be addressed that clustering algorithms are dependent on the data types given in a dataset. This means not all clustering algorithms are going to be eligible. A further novelty is the selected clustering algorithm(s) in this experiment. For this experiment, the chosen algorithm(s) is not been applied in past educational papers, not even as a layering process. Also, the use of applying benchmarks to determine the best cluster group (in this case Bouldin Davies) is not been used in past research which also delivers some novelty.

4.6.1 Parallel Architecture

Here are the models used in this model architecture:

- (Feed Forward) Neural Network 1 Hidden Layer & 8 nodes (NN18)
- (Feed Forward) Neural Network 3 Hidden Layers & 8 nodes (NN38)
- (Feed Forward) Neural Network 5 Hidden Layers & 8 nodes (NN58)
- (Feed Forward) Neural Network 10 Hidden Layers & 8 nodes (NN108)
- (Feed Forward) Neural Network 1 Hidden Layer & 4 nodes (NN14)
- (Feed Forward) Neural Network 3 Hidden Layers & 4 nodes (NN34)
- (Feed Forward) Neural Network 5 Hidden Layers & 4 nodes (NN54)
- (Feed Forward) Neural Network 10 Hidden Layers & 4 nodes (NN104)
- Random Forest (RFC)
- AdaBoost (ADAC)
- K-Nearest Neighbours (KNNC)
- Gradient Boost (GBC)
- Extreme Gradient Boost (XGBC)

- Gaussian Process (GPC)
- Passive Aggressive (PAC)
- Extra Trees (ETC)

In this case, Supervised Learning is integrated into this model architecture (Figure 3.3).

A selection of Feed Forward Neural Network architectures is applied and justifies if hidden layers have any impact on their prediction performance. What makes this family unique is the backpropagation applied to reduce error loss which results in a higher probability of accurate predictions.

Moreover, there are other types of architectures that are not applied to this experiment which are CNN, RNN & LSTM (for good reasoning). In brief, CNN is designed for image processing that uses components such as pixels, and RGB. It is used widely in Computer Vision. CNN works well with spatial relationship problems such as autonomous vehicles & identifying objects in relationship to something else with images. RNN is designed for speech/text mining processing that uses components such as sentences and words. It is used widely in Natural Language Processing (NLP). It is most suited to sequence prediction problems such as generating handwriting & language translation. An LSTM is part of RNN that requires less training and has a slightly different architecture. The LSTM is suitable for time-series problems such as forecasting stock markets.

The main barrier is the specification, they are designed for specific problems which must be met in order for them to deliver their purpose. Another barrier is the dataset, it must be compatible (e.g. image data for CNN). The dataset applied in this experiment is not suitable for these architectures. Even in existing literature, it's rarely used for predicting student performances. For example, Ding et al. [27] applied an LSTM (as a process to predict student performances) & the research is heavily focused on a time-series problem.

Also, they are designed for very rich datasets, which is difficult in this research field due to the ethics and legal barriers. In this case, assuming the problem matches their (architecture) specification, despite the data being sufficient, it may not be rich enough for these architectures to deliver reliable outcomes. This may result in poor predictions or some cases no results being produced (because of errors and/or conditions not being met). Which provides more reasoning why these architectures are ignored.

In terms of other Supervised Learning models (excluding Feed Forward Neural Networks), the Multi-Classification version of Supervised Learning is chosen and applies the parallel architecture as described in the framework (Model layer, Figure 3.3).

Not all models are adequate for this experiment. Examples of inadequate models include Support Vector Machine and Naive Bayes. Support Vector Machine is likely to run into overfitting/underfitting issues as the data applied is noisy, this results in poor predictions (hyperplane is not fitted well). Naive Bayes is a probability model, it does not make sense to compute probability with qualitative variables.

Random Forest is an ensemble algorithm that produces outcomes based on a selection of t Decision Trees votes. It also avoids overfitting/underfitting issues, and due to this reason, it is suitable to predict Multi-Classification problems. Decision Trees are an important factor. Since Decision Trees are rule-based models, they can identify patterns between variables and assign them as rules. Extra Trees

is somewhat similar to Random Forest but instead of computing the locally optimal feature/split combination, a random value is used instead.

AdaBoost uses the Boosting algorithm, since the method involves improving weak classifiers by providing more weights in a number of iterations, it can reduce the overfitting/underfitting problems. Since the dataset contains imbalanced variables, the Boosting algorithm can help reduce the issue when evaluating overall decisions, thanks to its Decision Stump iterator decider.

K-Nearest Neighbours (KNN) produces the decision based on nearby data points, these nearby data points are classed as similarities to a specific coordinate location (in a sub-space). As a result, whenever it produces outcomes, it collects students that have similarities and collects the mode value from a selection of k nearest neighbours. Nearest data points are more related to the problem than data points that are far away.

Passive Aggressive is focused on large-scale learning with regularisation to help improve prediction performance. With this being said it can identify hidden patterns and use them to its advantage. Also, since it doesn't change when the prediction is correct, therefore, it can increase the prediction accuracy.

Gradient Boost relies upon the intuition from the next model that is best, combining it with previous models, which helps reduce the error loss and improve the prediction accuracy. The Extreme version is different from the generic Gradient Boost with more regularisation options (L1 & L2). The Gaussian Process is designed to evaluate reliable estimates of its own uncertainty using probabilistic measures. Therefore, it can help reduce overfitting/underfitting and return accurate outcomes.

Tables 4.30 & 4.31 displays the configuration applied in this experiment.

TABLE 4.30: Configuration of each (Feed Forward) Neural Network (excluding other Supervised Learning) applied to this experiment.

Model	Configuration	Value	Feature Selection
NN18	Hidden Layers	1	All
NN18	Nodes	8	All
NN18	Output	1	All
NN18	Batch Size	1000	All
NN18	Epochs	300	All
NN38	Hidden Layers	3	All
NN38	Nodes	8	All
NN38	Output	1	All
NN38	Batch Size	1000	All
NN38	Epochs	300	All
NN58	Hidden Layers	5	All
NN58	Nodes	8	All
NN58	Output	1	All
NN58	Batch Size	1000	All
NN58	Epochs	300	All
NN108	Hidden Layers	10	All
NN108	Nodes	8	All
NN108	Output	1	All
NN108	Batch Size	1000	All
NN108	Epochs	300	All
NN14	Hidden Layers	1	All
NN14	Nodes	4	All
NN14	Output	1	All
NN14	Batch Size	1000	All
NN14	Epochs	300	All
NN34	Hidden Layers	3	All
NN34	Nodes	4	All
NN34	Output	1	All
NN34	Batch Size	1000	All
NN34	Epochs	300	All
NN54	Hidden Layers	5	All
NN54	Nodes	4	All
NN54	Output	1	All
NN54	Batch Size	1000	All
NN54	Epochs	300	All
NN104	Hidden Layers	10	All
NN104	Nodes	4	All
NN104	Output	1	All
NN104	Batch Size	1000	All
NN104	Epochs	300	All

TABLE 4.31: Configuration of each Supervised Learning (excluding Feed Forward Neural Networks) applied to this experiment.

Model	Configuration	Value	Feature Selection
Random Forest	Criterion	Gini	All
Random Forest	Trees	30	All
Random Forest	Max-depth	None	All
Random Forest	Max Features	Auto	All
Random Forest	Min Samples Leaf	1	All
Random Forest	Min Samples Splits	2	All
K-Nearest Neighbours	Neighbours	5	All
K-Nearest Neighbours	Distance	Euclidean	All
AdaBoost	Iteration	20	All
AdaBoost	Learning Rate	1	All
Gradient Boost	Learning Rate	0.1	All
Gradient Boost	Iteration (Boosting Stages)	100	All
Gradient Boost	Min Samples Leaf	1	All
Gradient Boost	Min Samples Splits	2	All
Gaussian Process	Iteration (Newton)	100	All
Passive Aggressive	Regularisation	1	All
Passive Aggressive	Iteration (Passes)	1000	All
Extra Trees	Trees	10	All
Extra Trees	Criterion	Gini	All
Extra Trees	Max-depth	None	All
Extra Trees	Min Samples Leaf	1	All
Extra Trees	Min Samples Splits	2	All
Extreme Gradient Boost	Learning Rate	1	All
Extreme Gradient Boost	Objective	Softmax	All
Extreme Gradient Boost	Iterations	1000	All
Extreme Gradient Boost	Max Depth	5	All
Extreme Gradient Boost	L1	0.3	All
Extreme Gradient Boost	L2	0.5	All

4.6.2 Popularity Architecture

Here are the models used in this model architecture:

- K-Modes/K-Prototype (KMU/KPU)

In this case, Unsupervised Learning is integrated into this model architecture (Figure 3.4).

In this scenario, one model is used but two possible types can be chosen. The decision is based on the variable data types. The majority of the dataset is qualitative with a minority of quantitative variables. The K-Modes cluster data with only qualitative variables, and the K-Prototype cluster data with a mixture of qualitative and quantitative variables. The work uses Feature Selection (and not) and may choose a list of features with both data types or one.

Other models such as K-Means and Hierarchical are not suitable as they generally perform better with datasets with quantitative variables (the majority). The K-Means group's data is based on a mathematical function in each iteration, the process does not make sense with qualitative variables. Hierarchical can be more

suited but due to its slow execution time with large samples and usefulness with quantitative variables, it would not perform well.

There is a chance that the Feature Selection can only choose quantitative variables but this is extremely rare due to the ratio of qualitative and quantitative variables.

For Unsupervised Learning, it is important to derive the best number of clusters to increase the chances of a better outcome. Fortunately, there are metrics available to determine the best cluster. In this work, Davies Bouldin's clustering performance is applied. Tables 4.32, 4.33, 4.34 and 4.35 display the score performance of each cluster.

The Davies Bouldin cluster performance is chosen due to its suitability with qualitative variables. The computation involves comparing the distance between clusters with the quantities of the clusters. Other types (e.g. Calinski and Harabasz & Silhouette Coefficient) normally involve computing the averages and/or variances, which do not make sense with qualitative variables (as the majority). Even though there are some quantitative variables, the benefits outweigh the drawbacks.

Table 4.36 displays the configuration applied in this experiment.

TABLE 4.32: The Davies Bouldin scores of each cluster using the Genetic Algorithm chosen features. The lowest score is classed as the best cluster number.

Cluster	Score
5	52.064443
6	47.233691
7	42.540461
8	39.685189
9	34.807184
10	30.789081
11	**24.770243**

TABLE 4.33: The Davies Bouldin scores of each cluster using the Particle Swarm Optimisation chosen features. The lowest score is classed as the best cluster number.

Cluster	Score
5	**14.266843**
6	22.258627
7	28.603664
8	164.136815
9	89.868902
10	45.931413
11	51.889532

TABLE 4.34: The Davies Bouldin scores of each cluster using the Recursive Feature Elimination. The lowest score is classed as the best cluster number.

Cluster	Score
5	**14.941357**
6	20.603112
7	17.587010
8	19.013560
9	28.770707
10	32.305778
11	26.799300

TABLE 4.35: The Davies Bouldin scores of each cluster using no Feature Selection. The lowest score is classed as the best cluster number.

Cluster	Score
5	**12.481492**
6	18.481842
7	12.834236
8	19.962476
9	17.084352
10	35.522522
11	27.694579

TABLE 4.36: Configuration of each Unsupervised Learning model applied to this experiment.

Model	Configuration	Value	Feature Selection
K-Modes/K-Prototype	Clusters	5/11/5/5	PSO/GA/RFE/None
K-Modes/K-Prototype	Init	Cao	All
K-Modes/K-Prototype	Iterations	100	All

4.7 Benchmark

The results (benchmark scores) from this experiment are presented in this section. Tables 4.37, 4.39, and 4.38 displays the benchmark scores. The *Min* & *Max* is the minimum and maximum Consistency-Scale (C_{min}, C_{max}), the *True* is the average Consistency-Scale (C_{avg}).

In the existing literature, several performance metrics have been used in relation to Classification or Regression problems ([36], [76], [74], [64]). Classification metrics compare actual and predicted outcomes and determine whether it is identical or not. The judgment is normally based on 1 or more of the following: Accuracy, AUC, F1 Score, & PRAUC. Regression metrics compute actual and predicted outcomes and determine the error loss by calculating the differences. The judgment is normally based on 1 or more of the following: MAE, & MSE. In this experiment, as explained in Chapter 3 a compatible benchmark known as *Valid Range* is applied. Traditional metrics in the existing literature are not compatible.

The benchmarks in the existing literature are appropriate if the outcome values are singular. These metrics are not appropriate for this framework due to different

formats. In this case, the outcome is a range that consists of boundary limits (minimum & maximum).

The Valid Range compares a range with a singular value. A correct outcome is when the singular value is within the range. Otherwise, it is incorrect. It is suitable for (Multi) Classification problems. The benefit of this approach (for predicting student performance) is that one can justify if the range is overlapping their normal performance. If it does not overlap with their normal performance then the outcome is more likely to be incorrect. Note that a range vs range (true vs predicted) approach would not be beneficial to the intended outcome.

A range outcome for student performance can be another solution as it considers the fluctuation due to struggles of student life (e.g. assessment, finance, & time management).

Please note about Tables 4.37, 4.39, and 4.38. For easier clarification, the number of correct is converted into percentages (represented in the *Accuracy* column). The percentages are the benchmark scores and represent the number of correct occasions (Equation 3.2).

TABLE 4.37: Display the Parallel Architecture benchmark scores (integrated with Supervised Learning). The results are subdivided into Feature Selection. The score highlighted in bold represents the best score.

Model	Feature Selection	Accuracy
KNNC	GA	85.31
ADAC	GA	77.53
RFC	GA	**93.01**
GBC	GA	88.68
GPC	GA	88.17
PAC	GA	81.74
ETC	GA	91.52
XGBC	GA	46.46
KNNC	PSO	84.55
ADAC	PSO	61.40
RFC	PSO	89.02
GBC	PSO	88.74
GPC	PSO	**89.5**
PAC	PSO	72.75
ETC	PSO	87.80
XGBC	PSO	30.84
KNNC	RFE	85.76
ADAC	RFE	74.61
RFC	RFE	**93.06**
GBC	RFE	87.84
GPC	RFE	89.5
PAC	RFE	79.78
ETC	RFE	92.75
XGBC	RFE	46.66
KNNC	None	85.8
ADAC	None	80.7
RFC	None	**93.51**
GBC	None	89.29
GPC	None	90.31
PAC	None	81.74
ETC	None	91.52
XGBC	None	45.22

TABLE 4.38: Display the Parallel Architecture benchmark scores (integrated with Feed Forward Neural Networks). The results are subdivided into Feature Selection. The score highlighted in bold represents the best score.

Model	Feature Selection	Accuracy
NN18	GA	86.63
NN38	GA	86.32
NN58	GA	84.3
NN108	GA	86.94
NN14	GA	**87.22**
NN34	GA	83.37
NN54	GA	84.94
NN104	GA	86.35
NN18	PSO	**93.68**
NN38	PSO	88.65
NN58	PSO	88.23
NN108	PSO	89.38
NN14	PSO	92.61
NN34	PSO	88.26
NN54	PSO	86.97
NN104	PSO	88.46
NN18	RFE	86.83
NN38	RFE	87.67
NN58	RFE	**89.83**
NN108	RFE	88.85
NN14	RFE	82.16
NN34	RFE	84.57
NN54	RFE	87.02
NN104	RFE	86.83
NN18	None	88.34
NN38	None	87.78
NN58	None	**89.13**
NN108	None	87.02
NN14	None	72.08
NN34	None	87.78
NN54	None	87.61
NN104	None	81.21

TABLE 4.39: Display the Popularity Architecture benchmark scores (integrated with Unsupervised Learning). The results are subdivided into Feature Selection. The score highlighted in bold represents the best score.

Model	Feature Selection	Accuracy
KMU/KPU	GA	95.87
KMU/KPU	PSO	94.10
KMU/KPU	RFE	92.64
KMU/KPU	None	**96.91**

Chapter 5

Results and Discussion

5.1 Introduction

This chapter provides a detailed evaluation of the experiment results produced in Section 4.7, discussing these results in detail. A baseline comparison is also provided, between this research and related work. Tables 4.37, 4.38, and 4.39 display the benchmarks for all models using the computational framework. Further discussion in this chapter relates to key points drawn from the research, feature selection, and potential explanations for particular results.

It is important to note that one cannot perform a direct baseline comparison as the outcome of the proposed computational framework is a performance range (not a performance value), and the dataset used in the experiment in the previous chapter is unique to this thesis and has not been used previously. This precludes direct critical comparison as commonly performed in literature [17] Hence, the comparison focuses on metric values to understand relative performance. Research in this area has commonly used metrics such as Accuracy, F1 Score, Precision-Recall curves, Root Mean Square Error & Mean Square Error ([60], [2], [74], [86]). Due to compatibility reasons, existing benchmark metrics are not applied. The framework explained in this thesis introduces a compatible metric.

Two types of baseline comparison can be considered in this context: 1) comparison to the average prediction accuracy of the most relevant research works; 2) comparison to the average prediction accuracy of any research that involves predicting student performance. Choosing the latter can be more problematic as it may not involve similar applications (e.g. data specification, machine learning algorithms). In this evaluation, the former is followed, with the baseline comparison focusing on the average performance of the 11 most relevant past papers identified in Chapter 2 ([18], [78], [60], [39], [2], [89], [87], [34], [55], [69], [31]). A paper is considered relevant if it contains similar input data specifications & methodologies (similar processes to produce predictions). This includes comparing the data variables, the data types, the student group niche applied, the type of decision-making models, and pre-processing (e.g. label encoding). It should be expected that the experiments contained in these papers are not fully identical to the ones conducted in this thesis, but these papers represent the most similar ones in the literature.

The paper amount is a good ratio of relevancy and several decision-making techniques. Based on the identified papers, the baseline target of prediction accuracy is 81%. While the full prediction accuracy range is $40\% - 100\%$, most commonly prediction accuracy values fall within the $70\% - 90\%$ range, which is the one used for boundary comparisons. In other words, papers that have a conclusion of benchmark performance or papers that show the vast majority of benchmark performances are considered in this range decision. The exceptional low and high scores occurred in

a few papers ([78], [31]). The baseline comparison can be done based on each family of approaches as well as across all.

The main purpose of the comparative analysis that follows is to evaluate the extent to which the proposed computational framework delivers equivalent or better prediction accuracy to related research.

5.2 Comparative Analysis

In this section, a benchmark comparison analysis between this experiment and the baseline is explored. Tables 5.1, 5.2 & 5.3 provides a summary between them. The *Baseline Accuracy* & *Baseline Min/Max Accuracy* are the baseline benchmark scores. The *Accuracy* is the experiment benchmark score for each integrated machine learning algorithm within each model architecture. The *Accuracy Difference* & *Min/Max Accuracy Difference* is the difference between the baseline benchmark scores and the experiment benchmark scores (in percentage). The difference can be positive or negative.

TABLE 5.1: Display the benchmark comparison analysis between the baseline and experiment average scores. The numerical units are percentages. This is part 1 of the scores.

Baseline Accuracy	Accuracy	Accuracy Difference	Integration	Feature Selection	Architecture
81	85.31	4.31	KNNC	GA	Parallel
81	77.53	-3.47	ADAC	GA	Parallel
81	93.01	12.01	RFC	GA	Parallel
81	88.68	7.68	GBC	GA	Parallel
81	88.17	7.17	GPC	GA	Parallel
81	81.74	0.74	PAC	GA	Parallel
81	91.52	10.52	ETC	GA	Parallel
81	46.46	-34.54	XGBC	GA	Parallel
81	84.55	3.55	KNNC	PSO	Parallel
81	61.4	-19.6	ADAC	PSO	Parallel
81	89.02	8.02	RFC	PSO	Parallel
81	88.74	7.74	GBC	PSO	Parallel
81	89.5	8.5	GPC	PSO	Parallel
81	72.75	-8.25	PAC	PSO	Parallel
81	87.8	6.8	ETC	PSO	Parallel
81	30.84	-50.16	XGBC	PSO	Parallel
81	85.76	4.76	KNNC	RFE	Parallel
81	74.61	-6.39	ADAC	RFE	Parallel
81	93.06	12.06	RFC	RFE	Parallel
81	87.84	6.84	GBC	RFE	Parallel
81	89.5	8.5	GPC	RFE	Parallel
81	79.78	-1.22	PAC	RFE	Parallel
81	92.75	11.75	ETC	RFE	Parallel
81	46.66	-34.34	XGBC	RFE	Parallel
81	85.8	4.8	KNNC	None	Parallel
81	80.7	-0.3	ADAC	None	Parallel
81	93.51	12.51	RFC	None	Parallel
81	89.29	8.29	GBC	None	Parallel
81	90.31	9.31	GPC	None	Parallel
81	81.74	0.74	PAC	None	Parallel
81	91.52	10.52	ETC	None	Parallel
81	45.22	-35.78	XGBC	None	Parallel

TABLE 5.2: Display the benchmark comparison analysis between the baseline and experiment average scores. The numerical units are percentages. This is part 2 of the scores.

Baseline Accuracy	Accuracy	Accuracy Difference	Integration	Feature Selection	Architecture
81	86.63	5.63	NN18	GA	Parallel
81	86.32	5.32	NN38	GA	Parallel
81	84.3	3.3	NN58	GA	Parallel
81	86.94	5.94	NN108	GA	Parallel
81	87.22	6.22	NN14	GA	Parallel
81	83.37	2.37	NN34	GA	Parallel
81	84.94	3.94	NN54	GA	Parallel
81	86.35	5.35	NN104	GA	Parallel
81	93.68	12.68	NN18	PSO	Parallel
81	88.65	7.65	NN38	PSO	Parallel
81	88.23	7.23	NN58	PSO	Parallel
81	89.38	8.38	NN108	PSO	Parallel
81	92.61	11.61	NN14	PSO	Parallel
81	88.26	7.26	NN34	PSO	Parallel
81	86.97	5.97	NN54	PSO	Parallel
81	88.46	7.46	NN104	PSO	Parallel
81	86.83	5.83	NN18	RFE	Parallel
81	87.67	6.67	NN38	RFE	Parallel
81	89.83	8.83	NN58	RFE	Parallel
81	88.85	7.85	NN108	RFE	Parallel
81	82.16	1.16	NN14	RFE	Parallel
81	84.57	3.57	NN34	RFE	Parallel
81	87.02	6.02	NN54	RFE	Parallel
81	86.83	5.83	NN104	RFE	Parallel
81	88.34	7.34	NN18	None	Parallel
81	87.78	6.78	NN38	None	Parallel
81	89.13	8.13	NN58	None	Parallel
81	87.02	6.02	NN108	None	Parallel
81	72.08	-8.92	NN14	None	Parallel
81	87.78	6.78	NN34	None	Parallel
81	87.61	6.61	NN54	None	Parallel
81	81.21	0.21	NN104	None	Parallel
81	95.87	14.87	KMU/KPU	GA	Popularity
81	94.1	13.1	KMU/KPU	PSO	Popularity
81	92.64	11.64	KMU/KPU	RFE	Popularity
81	96.91	15.91	KMU/KPU	None	Popularity

TABLE 5.3: Display the benchmark comparison analysis between the baseline and experiment range scores. The numerical units are percentages.

Baseline Min/Max Accuracy	Accuracy	Min Accuracy Difference	Max Accuracy Difference	Integration	Feature Selection	Architecture
70 - 90	85.31	15.31	-4.69	KNNC	GA	Parallel
70 - 90	77.53	7.53	-12.47	ADAC	GA	Parallel
70 - 90	93.01	23.01	3.01	RFC	GA	Parallel
70 - 90	88.68	18.68	-1.32	GBC	GA	Parallel
70 - 90	88.17	18.17	-1.83	GPC	GA	Parallel
70 - 90	81.74	11.74	-8.26	PAC	GA	Parallel
70 - 90	91.52	21.52	1.52	ETC	GA	Parallel
70 - 90	46.46	-23.54	-43.54	XGBC	GA	Parallel
70 - 90	84.55	14.55	-5.45	KNNC	PSO	Parallel
70 - 90	61.4	-8.6	-28.6	ADAC	PSO	Parallel
70 - 90	89.02	19.02	-0.98	RFC	PSO	Parallel
70 - 90	88.74	18.74	-1.26	GBC	PSO	Parallel
70 - 90	89.5	19.5	-0.5	GPC	PSO	Parallel
70 - 90	72.75	2.75	-17.25	PAC	PSO	Parallel
70 - 90	87.8	17.8	-2.2	ETC	PSO	Parallel
70 - 90	30.84	-39.16	-59.16	XGBC	PSO	Parallel
70 - 90	85.76	15.76	-4.24	KNNC	RFE	Parallel
70 - 90	74.61	4.61	-15.39	ADAC	RFE	Parallel
70 - 90	93.06	23.06	3.06	RFC	RFE	Parallel
70 - 90	87.84	17.84	-2.16	GBC	RFE	Parallel
70 - 90	89.5	19.5	-0.5	GPC	RFE	Parallel
70 - 90	79.78	9.78	-10.22	PAC	RFE	Parallel
70 - 90	92.75	22.75	2.75	ETC	RFE	Parallel
70 - 90	46.66	-23.34	-43.34	XGBC	RFE	Parallel
70 - 90	85.8	15.8	-4.2	KNNC	None	Parallel
70 - 90	80.7	10.7	-9.3	ADAC	None	Parallel
70 - 90	93.51	23.51	3.51	RFC	None	Parallel
70 - 90	89.29	19.29	-0.71	GBC	None	Parallel
70 - 90	90.31	20.31	0.31	GPC	None	Parallel
70 - 90	81.74	11.74	-8.26	PAC	None	Parallel
70 - 90	91.52	21.52	1.52	ETC	None	Parallel
70 - 90	45.22	-24.78	-44.78	XGBC	None	Parallel
70 - 90	86.63	16.63	-3.37	NN18	GA	Parallel
70 - 90	86.32	16.32	-3.68	NN38	GA	Parallel
70 - 90	84.3	14.3	-5.7	NN58	GA	Parallel
70 - 90	86.94	16.94	-3.06	NN108	GA	Parallel
70 - 90	87.22	17.22	-2.78	NN14	GA	Parallel
70 - 90	83.37	13.37	-6.63	NN34	GA	Parallel
70 - 90	84.94	14.94	-5.06	NN54	GA	Parallel
70 - 90	86.35	16.35	-3.65	NN104	GA	Parallel
70 - 90	93.68	23.68	3.68	NN18	PSO	Parallel
70 - 90	88.65	18.65	-1.35	NN38	PSO	Parallel
70 - 90	88.23	18.23	-1.77	NN58	PSO	Parallel
70 - 90	89.38	19.38	-0.62	NN108	PSO	Parallel
70 - 90	92.61	22.61	2.61	NN14	PSO	Parallel
70 - 90	88.26	18.26	-1.74	NN34	PSO	Parallel
70 - 90	86.97	16.97	-3.03	NN54	PSO	Parallel
70 - 90	88.46	18.46	-1.54	NN104	PSO	Parallel
70 - 90	86.83	16.83	-3.17	NN18	RFE	Parallel
70 - 90	87.67	17.67	-2.33	NN38	RFE	Parallel
70 - 90	89.83	19.83	-0.17	NN58	RFE	Parallel
70 - 90	88.85	18.85	-1.15	NN108	RFE	Parallel
70 - 90	82.16	12.16	-7.84	NN14	RFE	Parallel
70 - 90	84.57	14.57	-5.43	NN34	RFE	Parallel
70 - 90	87.02	17.02	-2.98	NN54	RFE	Parallel
70 - 90	86.83	16.83	-3.17	NN104	RFE	Parallel
70 - 90	88.34	18.34	-1.66	NN18	None	Parallel
70 - 90	87.78	17.78	-2.22	NN38	None	Parallel
70 - 90	89.13	19.13	-0.87	NN58	None	Parallel
70 - 90	87.02	17.02	-2.98	NN108	None	Parallel
70 - 90	72.08	2.08	-17.92	NN14	None	Parallel
70 - 90	87.78	17.78	-2.22	NN34	None	Parallel
70 - 90	87.61	17.61	-2.39	NN54	None	Parallel
70 - 90	81.21	11.21	-8.79	NN104	None	Parallel
70 - 90	95.87	25.87	5.87	KMU/KPU	GA	Popularity
70 - 90	94.1	24.1	4.1	KMU/KPU	PSO	Popularity
70 - 90	92.64	22.64	2.64	KMU/KPU	RFE	Popularity
70 - 90	96.91	26.91	6.91	KMU/KPU	None	Popularity

5.2.1 Parallel Architecture

Supervised Learning algorithms are integrated into the Parallel Architecture (Figure 3.3).

For (Feed Forward) Neural Networks (integration) excluding other Supervised Learning models, the model with 1 hidden layer (and 8 nodes) is the best predictor and the worst is the model with 1 hidden layer (and 4 nodes). The difference (all

models) in accuracy is around 22% (all models). The feature selection using Particle Swarm Optimisation produces the best benchmark score.

The fluctuation is quite high, if the worst benchmark score is removed as an outlier then the difference in accuracy is around 12%. Including & excluding the worst benchmark score returns no (direct hidden) patterns despite the number of hidden layers in the architectures. The case remains the same regardless of the number of nodes. Having different nodes does not provide any difference in accuracy performance. There is a good ratio of models and the accuracy performance is somewhat similar.

The accuracy is 87% on average ($72\% - 94\%$), which means a 6% increase compared to past works (baseline comparison). The difference shows an increase in both the upper and lower limits: lower: 2%; upper: 4%. In past works that apply Feed Forward Neural Networks, the popular amount of the hidden layer is 2 or below ([18], [87]). The hidden layer amount is motivated in this work due to the promising outcomes in past works. In rare cases, a higher number of hidden layers are used ([80]). The number of nodes applied in past works tends to be random.

However, it is important to explore several (Feed Forward) Neural Networks with different hidden layers to evaluate if they improve accuracy. The work here shows it does not provide any difference (in accuracy). This could be a reason why (Feed Forward) Neural Networks with several hidden layers are not frequently used in this context. One other possible reason is the execution time. Hidden layers are sensitive when predicting outcomes (like this) and the data fed into them is important. The level of noise within a dataset can also affect the outcome. As it stands, choosing a precise number of hidden layers and/or nodes is not required.

Past papers have shown 1 Feed Forward Neural Network being used in experiments. Applying more than 1 model with different architectures (hidden layers) in one setting (experiment) is not explored. Normally, it is applied by itself or combined with several Supervised Learning models (Mengash [60]). This experiment explores models with several hidden layers (and nodes) on one occasion to detect any unique behaviour.

For other Supervised Learning (integration) excluding (Feed Forward) Neural Networks, the best predictor is the Random Forest and the worst is Extreme Gradient Boost followed by AdaBoost. The difference in accuracy is around 33% (all models) excluding the worst model. With that exclusion, it is significantly higher. It should be noted that performing no feature selection produces the best benchmark score. This may be due to the better quality of the dataset, as feature selection usually leads to better performance when a dataset includes more features with less predictive capability [1]. It tends to be more fluctuated compared to other decision-making families. The reason can be the number of models, as this category contains the highest number of models in one experiment which increases the probability of a fluctuation of benchmark scores. Most models generally perform the same which reduces the accuracy difference significantly.

The accuracy, excluding the Extreme Gradient Boost model, is 85% on average ($61\% - 94\%$), which means a 4% increase compared to past works (baseline comparison). The difference shows an increase in the upper limit and a decrease in the lower limit: lower: 9%; upper: 4%. In past works that explore Supervised Learning, the most commonly explored models include Random Forest, Logistic Regression, SVM, Naive Bayes, etc (Adekitan and Noma-Osaghae [2], Francis and Babu [31]). They do provide promising benchmark scores but most past works do not consider more recent models such as Gradient Boost, & Extra Trees.

If the Extreme Gradient Boost model is included in the benchmark performance then the overall accuracy is 80% on average ($31\% - 94\%$). This shows a 1% decrease compared to past works (baseline comparison). The difference shows an increase in the upper limit and a decrease in the lower limit: lower: 39%; upper: 4%. There is no effect on the upper limit which is far more important. But it is worth noting this the Extreme Gradient Boost model is uncommon in past works, only a few occasions are identified (Niyogisubizo et al. [62], Kiss et al. [50]).

Including & excluding the Extreme Gradient Boost is due to the out-of-the-ordinary benchmark performance. So, the separation is there to showcase the potential of the framework in a general case scenario. Furthermore, the possible reason for Extreme Gradient Boost's poor performance can be the suitability of data types applied to the model. Recall that the data applied to this research is largely biased with qualitative variables. Operating with qualitative variables is not always suitable across all algorithms [16]. Extreme Gradient Boost may be more appropriate for datasets with primarily quantitative variables due to the regularisation options it offers. Another possible reason is noisy data which is likely to lead to overfitting issues. These concerns may be the reason why the uses of Extreme Gradient Boost in past works are not popular.

If the AdaBoost lowest performance (61%) is ignored from the benchmark scores then the lower limit range in this work becomes greater than past works by 2% (the new lower limit becomes 72%). Both AdaBoost & Extreme Gradient Boost apply the Boosting algorithm. The benchmark shows that algorithms applying the Boosting algorithm are more likely to deliver worse performance than others.

Looking further, when observing many models in one experiment, the prediction performance tends to be fairer. Even though the effect of applying a high number of models is useful (in this experiment), the benchmark performance does not show much difference between them in most models. But it is fairer & reliable compared to analysing different models in several papers that use different data and methodologies (not all models follow one specification). The maximum number of machine learning algorithms found in past papers is 7 (Rodríguez-Hernández et al. [69]).

Unlike past papers, this thesis provides a better balance of popular, unpopular, recent, and more established machine learning algorithms) are applied. The accuracy scores might provide some initial thoughts on the reasoning behind why some models are less popular. Perhaps the enormous amount of work of usage with these popular models can be the reason why unpopular models are not equivalent in usability. Additionally, the performance and trustworthiness of more established models can be more attractive to researchers than considering algorithms that have not been tried and proven across many application areas.

5.2.2 Popularity Architecture

Unsupervised Learning algorithms are integrated into the Popularity Architecture (Figure 3.4).

For Unsupervised Learning (integration), there is only one model (integrated into the model architecture), so no model comparison is required. The experiment with no feature selection produces the best benchmark score. The difference in accuracy is around 2% (all models). Unlike other decision-making families, Feature Selection plays the least role in the case of unsupervised learning. This can be an advantage as one layer is removed, therefore execution speed is reduced and its application is more practical.

The accuracy is around 95% (95% − 97%), which means a 14% increase compared to past works (baseline comparison). The difference shows an increase in both the upper and lower limits: lower: 25%; upper: 7%. There is not much to say about this family due to its limited usage in past works. Clustering in theory is to group data (in this case students), the outcome here shows it can group similar students using their variables.

Unsupervised Learning performed better than other families. This can be attributed to the process (the Popularity Architecture) used in this (computational) framework (only applied to this model due to specification). In the context of the framework, the given dataset, and the performed experiments, the method of clustering data seems to be beneficial for student performance predictions. Improved results may also be related to the suitability of the data types to the chosen clustering algorithm(s). It should be noted that only one unsupervised model is applied, so results cannot be generalised to apply to unsupervised learning algorithms as a whole.

It is also worth noting that a different combination of features is applied for clustering, which can be another reason behind the difference in performance. Looking through the benchmark performances and the choices of feature combinations, it does not show a pattern. There are cases of a lower number of feature combinations applied and still, the accuracy performance is high (and vice versa).

Performance differences can also be attributed to the variables. The data applied mainly contain characteristics that may be better with clustering techniques. After all, past research has used characteristics to predict student performance (Mengash [60]). On the other hand, a case study (Alwarthan, Aslam, and Khan [8]) explored the trends of features in similar work and discovered the vast majority of papers show no direct evidence of a strong relationship between characteristics and student performance. This may provide a clearer sign that the framework's process is the main reason.

Past papers show Unsupervised Learning has rarely been applied due to its clustering nature (not compatible). It's not a prime predictor algorithm in this thesis, instead, it's treated as a pre-processing step to assist student performance predictions. It assists in grouping student data with similarities (before producing predictions). Despite the complication, the results here show potential. Also, the chosen clustering algorithms in this experiment is not been applied in this manner. Hopefully, it can encourage more usage in the future but researchers should be aware that using this model for predicting student performance is more complex compared to other algorithms (e.g. Supervised Learning) that require fewer preparation steps. Additional complexity is due to the number of clustering algorithm options, it is limited compared to the other decision-making families. It becomes more limited when data type sensitivity comes into play.

5.2.3 Overall Comparison

In an overall comparison (considering all models except the worst integrated model, Extreme Gradient Boost), the average accuracy in this thesis is around 86% (61% − 97%), which means a 5% increase compared to past works. The difference shows an increase in the upper limit and a decrease in the lower limit: lower: 9%; upper: 7%. In addition, including the worst model benchmark performance (Extreme Gradient Boost), the lower limit returns a decreased outcome by 39%. But the upper limit remains the same, the overall accuracy is 84% which is still an increase by 3%.

5.3 Further Discussion

The benchmark scores show in the vast majority of cases the computational framework outperforms past relevant papers (based on the baseline comparison). This applies to both the overall (averaged) prediction accuracy and the prediction accuracy boundaries (lower & upper accuracy limits).

The increase of the lower limit (on most occasions) is higher than the upper limit. Even though the importance is more on the upper limit, the framework does show that even in the worst-case scenarios, it is more likely to be accurate than past relevant works. The upper limit shows an increase in all cases which provides more confidence for returning accurate academic performance. With promising results like this, it is more likely usable commercially & practically.

As mentioned before, the Unsupervised Learning integration in the Popularity Architecture performed the best predictions in all cases. The benchmark performance (scores) show that not applying Feature Selection performed the best, with an exclusion of one. This can be an indication that having more variables is better than a good combination of features. Occasionally, the Feature Selection algorithms provide similar behaviour. It requires execution time which may be time-consuming. This may provide more motivation for avoidance. Note that this performance of feature selection may be directly related to the particular dataset used.

There are rare cases of out-of-the-ordinary benchmark performance in past works. One reason could be the data quality that may be biased to one side. For example, if one collects a dataset with just high-performing students then the outcome is going to be equivalent. Therefore, the prediction accuracy can be phenomenal but the data is not to the right standard for fair observation. That is why the popular benchmark scores are taken into account only as it delivers more confidence that more safeguards are applied.

Another reason for the exclusion of outlier performances is concerns about the reliability of a particular experiment. A valid question is raised, as to why only a single study has achieved such performance and not the vast majority (with a significant amount of data, such as Rodríguez-Hernández et al. [69])? It is also worth noting that some research related to student performance follows a standard data analysis process (e.g. Zacharis [87], Zhao et al. [89]). In such cases, the appropriateness of such a generic process is questionable, given that data either contain insufficient samples/variables or are imbalanced (or both). This also applies to outcomes with unusually low benchmark scores. The reasoning can be similar.

If one explores the most similar past papers and compares them with this thesis (Bilal et al. [18], Mengash [60], Helal et al. [39]), several differences can be identified. The data quality is the obvious one, the number of samples and variables is lower compared to the data considered in this work. The samples are normally under 4000 (in most cases) and under 20 variables (when specified). The dataset used in this experiment contains more than 6000 samples (after all pre-processing). The original dataset samples are significantly higher and it contains 7 additional variables. Also, one should dive into the type of variables. Although they do have interesting factors they do lack some important variables. For example, the distances between student accommodation to the institution, the deprivation level of their home residence, and their socioeconomic background. All of these deliver more insight into their background and show how it influences their academic performance.

Past papers that apply a high volume of data are generally collected through their (third-party) partners and/or they have sufficient privilege to access a high-quality dataset (Hussain and Khan [46], Zhang and Rangwala [88]). Otherwise, it is

difficult collecting a rich dataset due to legal and ethical issues. In this work, over 27000 students are involved but a fraction is applied to the experiment due to the pre-processing in the computational framework.

As discussed in this chapter, performance is also improved compared to past work. If one observes the relevant works in the past and the work conducted here, the computational framework consists of a unique combination of layers. The pipeline process is designed with many safeguards to avoid unfair advantages and return reliable benchmarks.

Improved performance can be attributed to specific characteristics of particular layers:

- The Data Analysis layer includes applying an in-depth analysis of a given dataset that showcases its strengths and weaknesses. Past papers have shown a lack of in-depth analysis and only generic ones are executed such as sample size, basic statistics (e.g. average), and/or variable data types. But it misses several other important analyses such as dimensionality reduction (e.g. FAMD) and other statistics (e.g. skew, quantiles). In this experiment, an example of in-depth analysis is given which allows one to understand its strengths and weaknesses. Doing so allows one to focus more on a student group that does not lead to an unbiased evaluation. Having sufficient samples, balanced data (ratio), relatable variables, and more return reliability to the methodology.

- The Feature Selection layer includes portability with data types that can collect features (variables) without restrictions. Past papers have shown Feature Selection applications to be limited and do not apply several algorithms in one setting. In this experiment, several portable feature selection algorithms are integrated. It also considers a none feature selection observation because sometimes it's not the case of a good combination of features but it can be the number of features (or both).

- The Anchored Training Data layer reduces the imbalanced data, which increases the probability of unbiased outcomes. Past papers have shown the lack of effect applied to resolve the imbalance data problem with educational data. In this experiment, an algorithm is used to help tackle this problem. Tables 4.28 & 4.29 show occasions where the total number of samples is not the recommended amount (to train models). This means, with the given dataset (and pre-processing), the algorithm found a lower sample amount to be better for predictions. The variation of samples shows evidence that the algorithm provided support to the experiment. One cannot remove imbalanced data completely but even reducing the portability by 1% is better than nothing.

- The Model layer applies appropriate Machine Learning algorithms to predict student performance. The model specification must suit the given problem. Otherwise, it may not return fruitful outcomes. Past papers have shown that a performance value is common for predictions. In this experiment, the models return a performance range prediction to each student which presents the minimum/maximum academic capabilities. A performance range (tolerance) considers the circumstances of student life (e.g. finance, health, family). The outcomes from this experiment deliver motivation for its uses in this research community. It can be an alternative output option for problems similar to this thesis.

During this research, several key points have been drawn:

- The challenges of predicting student performance: While it may seem that accurate results can be easily produced by just inserting data into a model, this research shows that it requires more in-depth work to evaluate outcomes such as data analysis & pre-processing. It provides more confidence in the outcome as any issues have been measured and solved.
- Quality of data and ethical challenges: The ethics and legal side give some understanding of why occasionally the data quality tends to be lower than expected. Unless there is the right authorisation (e.g. governing bodies), collecting a rich dataset for research purposes may be quite challenging.
- Room for more variables: The analysis in this thesis and past works (their unique variables) give the motivation that more variables are beneficial. This is because more variables result in the discovery of more information for students.

Overall, it can be claimed that the computational framework delivers usefulness for higher education. In most cases, the prediction accuracy outperforms similar past works. This framework is more likely to produce accurate academic performance, on average, compared to previous research. By addressing the research gaps identified at the beginning of the thesis, issues that may affect past research are dealt with. Moreover, results validate the chosen methods for addressing these research gaps, in the context of student performance prediction and possibly beyond.

Chapter 6

Conclusion

6.1 Summary of Contributions

Predicting student performance provides many benefits to institutions but also challenges in the sense of accurate and reliable predictions. This avoids students from going in the wrong direction. The use of investing time in this topic results in support for the institutions and students. It allows institutions to have early warning of at-risk students and explore potential methods of preventing it. In this topic, datasets generally contain demographic data (e.g. ethnicity, gender) that are classed as personal, but they also contain course details (e.g. attendance, grades). The topic has explored Machine Learning algorithms in effective ways.

The presented work provides a great understanding of the initial aim and objectives. The aim is to develop and execute a computational framework for higher education that consists of filling in several missing gaps in the existing literature. The framework contains a unique combination of layers to predict student performance.

The following are this thesis' identified contributions in relation to Applied Machine Learning and Data Analytics in the context of Education applications.

- Performance Range: This work predicts student performance with a range rather than a value. A range can deliver more insight into a student's capabilities. It is more suitable as it considers the circumstances of student life (e.g. assessments, finance, family). Students are likely to fluctuate in their studies.

- Imbalanced Data Algorithm: This work applies a unique algorithm that is able to reduce the probability of imbalanced data much as possible. This is important as imbalanced data can result in biased outcomes and may not be reliable. In addition, there is an insufficient amount of solutions to tackling this problem.

- Unsupervised Learning Integration: This work applies a model pipeline process that includes Unsupervised Learning as one of the pre-processing steps. Its purpose involves grouping similar student data (before producing predictions). Due to its clustering techniques, Unsupervised Learning has rarely been used in this research niche. And this work is another occasion of its uses.

- Several Feed Forward Neural Networks Architectures: This work applies several (Feed Forward) Neural Network architectures with different hidden layers in one setting. Normally, one (Feed Forward) Neural Network model is applied and mostly uses $1-2$ hidden layers. Having several of them in one setting provides more details of their behaviour.

- In-Depth Data Analysis: This work applies in-depth data analysis with sufficient details of a given dataset. This includes exploring the common data analysis (e.g. correlation) and other useful ones (e.g. dimensional reduction, hypothesis tests, quantile, and skew). Having deep analyses can expose more hidden patterns. It provides a greater understanding of the dataset's strengths and weaknesses.

As a result, the prediction accuracy is around 84%/86% (worst/common case scenario) with a 3%/5% (worst/common case scenario) increase compared to past relevant works. This means most outcomes are able to predict each student's performance correctly. Considering this, the computational framework, through achieving the state's aims and objectives, can have a positive impact on student performance prediction research.

6.2 Future Research

There are several interesting research directions that come out of this thesis.

Firstly, it is important to explore the applicability of the framework further, by using datasets from several institutions. Having datasets from several institutions can provide more analysis and confidence in the proposed framework. This includes understanding whether performance improvement identified in experiments conducted in this thesis carries over to other datasets. At the moment, the dataset applied is rich but having more datasets always makes the outcome more reliable. Having said that, collecting such data is difficult. This is due to ethical and legal issues, including data protection. It requires money and time to collect these datasets. The situation can be accelerated and likely to be executed if local and government authorities are involved.

Secondly, it is worth understanding whether datasets that are more feature rich can result in further improvements. Although the dataset used contains many characteristics of students it of course does not fully cover any available information for a student. Having more information can showcase more hidden patterns (from students) and evolve the methodology. Examples include more details about their parents, current health status, and previous institution. Typical examples of parental data are household income (e.g. £25,000), civil partnership (e.g. divorced), their careers (e.g. Electrical Engineer) & types of qualifications (e.g. BEng). Typical examples of health data are disability (e.g. yes/no), allergic reaction(s) (e.g. flowers) & sporting activities (e.g. football). Typical examples of details from the previous institution are attendance, grades (e.g. A, B), reports from past teachers, subject focus (e.g. Physics, Biology), and hobbies (if any). Collecting these factors can be difficult but possible. Perhaps surveys can be easier if the data does not exist or it is difficult to collect.

Further research can be conducted on reducing data imbalance. New variants are useful in this topic to support future analysis. This research explores one solution but more is required to provide flexibility in analysis & data mining (or similar). One focus can be a rule-based architecture [15]. This means an algorithm that collects the best ratio of each variable that fulfills rules set by an individual.

The specification of Unsupervised Learning provides complexity barriers in this research niche. This thesis provides motivation for its potential and future usage. So, an interesting study can be exploring other clustering techniques to assist student performance predictions (either ranges and/or values) during the pre-processing

stages. It may provide a new field of research and possibly be used in institutions. The difficulty here is more about finding optimal ways of achieving prediction through a clustering architecture.

While research in this thesis explored Feature Selection thoroughly, there is room for further exploration. This is especially true for datasets that only or primarily include qualitative factors. Therefore, potential research can be investigating useful Feature Selection approaches in this context, exercising the appropriate caution in terms of dealing with qualitative vs. quantitative variables.

In terms of delivering impact to academic institutions through this research, the developed computational framework can be packaged as a service-based application [14, 13, 12]. This will allow its integration into existing student performance dashboard systems, allowing academics and administrators to harness its predictive capabilities and drive decision-making to improve outcomes across the student body or targeted efforts, such as reducing differential attainment.

Appendix A

Preliminaries

There is pre-knowledge required to understand the scope of this thesis. Here are the necessary topics required to understand the scope. This is an addon and only the important topics are presented.

A.1 Supervised Learning

Supervised Learning [23] focuses on the study of predicting future events using current/past events. The details are factors and they should be relevant to the event and problem. All factors are represented in a dataset D which is labeled. This means the identity name of each factor is visible and known. It is a requirement for Supervised Learning models. The outcome of all Supervised Learning models is to produce a hypothesis function $h(x)$ using current/past information.

D is divided into two sets: training D_{train} and testing D_{test} sets. Both D_{train} and D_{test} can be single/multiple dimensional sub-space. The sub-space can be represented as a matrix/vector $\mathbb{R}^{n \times m}$. One of the requirements is to have features $\mathbb{R}^{n \times m}$ and target $\mathbb{R}^{n}$. The features are variables (known events) that teach (train) the model(s) about the problem one wishes to predict. The target is a variable (unknown event) that notifies the models about the problem one wishes to predict.

D_{train} contains details of past events and should contain the features and target. There must not be any unknowns in D_{train}. D_{test} contains details of current events that only contain the features but not the target. The target is unknown in D_{test}, it is the duty of the model(s) to predict those unknown (event). Both D_{train} and D_{test} should include the same variables and labels. The label is important as all variables are inside a sub-space.

The model is trained using D_{train}, each feature (variable) represents a dimension with a separate sub-space and is positioned on the x-axis. The target is positioned on the y-axis in each sub-space.

In each sub-space, the model produces a hypothesis function $h(x)$. $h(x)$ is basically the line of best fit using D_{train} and it is used to predict the unknown target variable in D_{test} using Equation A.1. Where y is the prediction outcome (from the D_{train}). The $h(x)$ can be in any polynomial degree (e.g. $2, 3$). In Supervised Learning, depending on the problem, one of the following is suitable: Classification or Regression.

$$y = f(x) \tag{A.1}$$

Although Classification & Regression are different types, several decision-making algorithms exist that apply both concepts. Generally, the difference is the final step. For Classification, the voting method is applied. For Regression, the average method

is applied. Examples of algorithms include AdaBoost ([25]), Random Forest ([57], [56]), & KNN ([73]).

A.1.1 Classification

Classification solves problems that are qualitative (binary/categorical). This means the target should be a whole number. Normally, it is binary outcomes (e.g. $0/1$, Yes/No) that only contain two outcomes. Sometimes, depending on the problem, it may contain three or more outcomes (e.g. $0/1/2$, $Yes/No/Maybe$). This is known as Multi-Class Classification. It is normally suitable for binary-based problems.

The data types for the features in D_{train} and D_{test} are irrelevant and they do not affect predictions (as long they are relevant to the problem). But the target must be qualitative (binary/categorical). The $h(x)$ can be produced using Equation A.2 which is known as the Sigmoid function and is for binary problems.

$$y = \frac{1}{(1 + e^{-z})} \tag{A.2}$$

Metrics are used to count the number of correct and incorrect outcomes between a prediction vector p and true vector t. A model produces p, and the t is used against p to evaluate the accuracy in p. The output score is $0 - 1$ (worst to best accuracy). These outcomes are four elements:

- True Positive (TP) - When the actual and expected are correct.
- False Positive (FP) - When the actual and expected are incorrect.
- True Negative (TN) - When the none expected and actual are correct.
- False Negative (FN) - When the actual and none expected are incorrect.

Common metrics include Area Under The Curve (AUC), Accuracy, Precision/Recall, and F1.

AUC computes the area that is under the so-called ROC (Receiver Operating Characteristic) curve. The ROC curve plots the True Positive Rate or Recall vs. the False Positive Rate (Equation A.7) at different classification thresholds. In the two extreme cases, an AUC equal to 1 denotes a model that is always correct, with AUC equal to 0 signifying a model that never classifies correctly.

Accuracy is the most common metric due to its communication and simplicity. The calculation is simple: count the number of correct and incorrect values in p against t and then compute the percentage. Equation A.6 presents the mathematical notation. Precision and Recall are similar to the Accuracy metric but they do not consider the FN and FP quantities, respectively. Equations A.3 and A.4 calculate the metrics. The F1 score is the average score between Precision (P) and Recall (R). The formula is presented in Equation A.5.

$$Precision = \frac{TP}{TP + FP} \tag{A.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{A.4}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{A.5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.6}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{A.7}$$

A.1.2 Regression

Regression solves problems that are quantitative (e.g. continuous). The $h(x)$ can be formed with any polynomial degree. A linear with the perfect fit $h(x)$ is the best $h(x)$, which means it can predict accurately. The coefficient scores are generated when the $h(x)$ is generated.

$$y = b_0 + (b_1 * x_1) + (b_2 * x_2) + ... + (b_i * x_i) + e \tag{A.8}$$

The data types for the features in the D_{train} and D_{test} are irrelevant and they do not affect predictions. But the target must be quantitative (e.g. continuous). The $h(x)$ is produced using Equation A.8. Where b_0 is the bias, b_i is the coefficient score for each feature x_i, i is each row, and e is the random error component. The coefficient scores are generated when the $h(x)$ is generated.

Metrics compute the error differences between a prediction vector p and true vector t. A model produces p, and the t is used against p to evaluate the error loss in p. The output score is not fixed to a range, it starts from 0 (worst to best accuracy).

Common metrics are Mean Absolute Error (MAE) and Mean Squared Error (MSE).

The MAE produces the error that is absolute using Equation A.10. Where N is the number of data points, y_i is each prediction value, $\overline{y}_i$ is each actual value, and $|*|$ is the absolute value. The MSE produces the error that is not absolute (accept positive and negative error losses) using Equation A.9. Where N is the number of data points, y_i is each prediction value, and $\overline{y}_i$ is each actual value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_i)^2 \tag{A.9}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \overline{y}_i| \tag{A.10}$$

A.2 Unsupervised Learning

Unsupervised Learning [19], [32] focuses on clustering (grouping) data that exhibit similar behaviour. Suppose a dataset $D \in \mathbb{R}^{n \times m}$ and split into training D_{train} and testing D_{test} sets. It trains the model by clustering (grouping) data in D_{train} with data points in D_{test}.

In these models, a popular hyperparameter is used known as cluster c_i. c_i is a whole number $(2, 3, ..., n)$ and it trains a model to cluster the data points in D_{train} to the number of clusters. The data that is allocated to the same c_i means those samples are similar. The specification of matching data differs in each model.

Unsupervised Learning does not predict upcoming events but recommends data that are similar. It also does not need a target vector (not predicting an unknown event), just features.

The suitable cluster c_i enables data to be allocated to the most suitable group. There are many Cluster Performances that apply unique functions to return metric scores to find the best c_i such as Davies-Bouldin [71].

There are many types of clustering algorithms such as K-Modes ([79], [42], [22]), K-Prototype ([79], [43]), & K-Means ([4]).

A.3 Heuristic Search

Heuristic Search is a family of mathematical optimisation techniques [30]. In this family, execution is the unique selling point. Examples of Heuristic Searches include the Genetic Algorithm & Particle Swarm Optimisation.

A.3.1 Genetic Algorithm

The Genetic Algorithm (GA) is motivated by the theory of natural evolution [83]. The process is to generate a number of individuals, which becomes a population and is then interpreted as a family tree. The family tree contains root parents which are normally assigned to the first initialised individuals. Each parent passes down their characteristic (in Biology, this is genetics) to the next generation. In each generation, a fitness score is computed, and the complexity of the fitness score is optional. The fitness score judge which individuals are likely to survive the longest. If an individual in the new generation contains a higher fitness, it means that the individual has a lower mortality rate. This is because the individuals in the current generation have learned to survive from the last generation. The generation update is repeated in a number of iterations $i \in i_1, i_2, ..., i_n$. The generation with the best fitness (fittest) individuals is returned. Those individuals are the solution to the problem.

There are five steps, and these steps are repeated depending on the number of i:

- Initial Population
- Fitness Function
- Selection
- Crossover
- Mutation

A.3.2 Particle Swarm Optimisation

The Particle Swarm Optimisation (PSO) [66] is inserting a population of particles in a sub-space. The sub-space should contain a destination location. The destination is described as the solution to the problem. The particles work together to navigate themselves toward the destination location. This is done by assisting the nearby particles. The method can be described as a bunch of birds searching for food in the sky. The food is stored in a location (destination) and the birds (particles) begin their search until they detect a sign. When one bird detects the (first) sign, the nearby birds start to move in that direction. It also navigates the group of birds closer to the food (destination). This one task is repeated until one bird finds the food (destination). The repeats in the example are the iterations and the stop when the last iteration is completed (and the solution is found).

A.4 Feed Forward Neural Networks

Feed Forward Neural Networks [33] form a simplistic abstraction of the human brain, consisting of nodes with links to predict outcomes. A (Feed Forward) Neural Network generally contains several hidden layers (e.g. 2/3). A Deep Learning model contains more hidden layers (e.g. 10) than a (Feed Forward) Neural Network. There are architectures that are designed for specific problems such as Convolutional Neural Networks (CNN) & Recurrent Neural Networks (RNN) with slight differences. D must contain features and target, D_{train} features & target are known but D_{test} features are only known (the target is unknown). The model is trained with D_{train} to predict the target (vector) in D_{test} using its features.

Each architecture contains an input layer, at least one hidden layer, and an output layer. Each layer contains nodes & the nodes in one layer are linked to the next layer (except the output layer). Each link contains random weights w and it computes the Dot Product. The w are updated in each iteration i. The update is based on the Cost Function which is used to find the local minima.

The input layer inserts a dataset D and each node represents a vector $\mathbb{R}^n$ (1 column). The hidden layers help increase the accuracy by reducing the error losses (depending on the circumstances). In each hidden layer, the node computes an Activation Function using the previous layer's node outputs. Examples of Activation Functions are RELU (Equation A.11, where z is the squared coordinates to compute the intercept function) & Sigmoid (Equation A.2). The output layer returns the predictions which contain at least one node. Also, each node (in the output layer) represents an outcome that can be either Classification or Regression.

$$g(z) = max(0, z) \tag{A.11}$$

In addition, these models contain Back-Propagation to improve accuracy. After Feed-Forward (from one layer to the next layer), the Back-Propagation goes in reverse to decrease the error losses even further (if possible). Lower error losses mean better accuracy. Moreover, there are hyperparameters such as batch size (number of training samples), and epochs (iterations) to help increase accuracy.

Bibliography

[1] Ahmad Abdulla, George Baryannis, and Ibrahim Badi. "Weighting the key features affecting supplier selection using machine learning techniques". In: *7th International Conference on Transport and Logistics, Niš, Serbia, 6 December 2019*. 2019, pp. 15–20. DOI: 10.20944/preprints201912.0154.v1.

[2] A.I. Adekitan and Etinosa Noma-Osaghae. "Data mining approach to predicting the performance of first year student in a university using the admission requirements". In: *Education and Information Technologies* 24 (Mar. 2019), pp. 1527–1543.

[3] Muhammad Adnan et al. "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models". In: *IEEE Access* PP (Jan. 2021), pp. 1–1. DOI: 10.1109/ACCESS.2021.3049446.

[4] Charu Aggarwal and Chandan Reddy. *DATA CLUSTERING Algorithms and Applications*. Aug. 2013.

[5] Deepti Aggarwal, Sonu Mittal, and Vikram Bali. "Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques". In: *International Journal of System Dynamics Applications* 10 (July 2021), pp. 38–49. DOI: 10.4018/IJSDA.2021070103.

[6] Everaldo Aguiar et al. "Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. LAK '15. Poughkeepsie, New York, NY: Association for Computing Machinery, 2015, 93–102. ISBN: 9781450334174. DOI: 10.1145/2723576.2723619. URL: https://doi.org/10.1145/2723576.2723619.

[7] Alireza Ahadi et al. "Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance". In: *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*. ICER '15. New York, NY, USA: Association for Computing Machinery, 2015, 121–130. ISBN: 9781450336307. DOI: 10.1145/2787622.2787717. URL: https://doi.org/10.1145/2787622.2787717.

[8] Sarah Alwarthan, Nida Aslam, and Irfan Khan. "Predicting Student Academic Performance at Higher Education Using Data Mining A Systematic Review". In: *Applied Computational Intelligence and Soft Computing* 2022 (Sept. 2022). DOI: 10.1155/2022/8924028.

[9] Grigoris Antoniou, Emmanuel Papadakis, and George Baryannis. "Mental Health Diagnosis: A Case for Explainable Artificial Intelligence". In: *International Journal on Artificial Intelligence Tools* 31.03 (2022), p. 2241003. DOI: 10.1142/S0218213022410032.

[10] Ghada Badr et al. "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department". In: *Procedia Computer Science* 82 (2016). 4th Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia, pp. 80–89. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.04.012. URL: https://www.sciencedirect.com/science/article/pii/S1877050916300266.

[11] Maria Barron Estrada et al. "Sentiment Analysis in an Affective Intelligent Tutoring System". In: July 2017, pp. 394–397. DOI: 10.1109/ICALT.2017.137.

[12] George Baryannis, Kyriakos Kritikos, and Dimitris Plexousakis. "A specification-based QoS-aware design framework for service-based applications". In: *Service Oriented Computing and Applications* 11.3 (2017), pp. 301–314. ISSN: 1863-2394. DOI: 10.1007/s11761-017-0210-4.

[13] George Baryannis and Dimitris Plexousakis. "Fluent Calculus-based Semantic Web Service Composition and Verification using WSSL". In: *9th International Workshop on Semantic Web Enabled Software Engineering (SWESE2013), co-located with ICSOC 2013*. Ed. by A.R. Lomuscio et al. Vol. 8377. Lecture Notes in Computer Science. Springer International Publishing Switzerland, 2014, pp. 256–270. DOI: 10.1007/978-3-319-06859-6_23.

[14] George Baryannis and Dimitris Plexousakis. "WSSL: A Fluent Calculus-Based Language for Web Service Specifications". In: *25th International Conference on Advanced Information Systems Engineering (CAiSE 2013)*. Ed. by Camille Salinesi, Moira C. Norrie, and Óscar Pastor. Vol. 7908. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 256–271. ISBN: 978-3-642-38708-1. DOI: 10.1007/978-3-642-38709-8_17.

[15] George Baryannis, Przemyslaw Woznowski, and Grigoris Antoniou. "Rule-Based Real-Time ADL Recognition in a Smart Home Environment". In: *Rule Technologies. Research, Tools, and Applications: 10th International Symposium on Rules and Rule Markup Languages for the Semantic Web (RuleML 2016)*. Ed. by Jose Julio Alferes et al. Vol. 9718. Lecture Notes in Computer Science. Springer International Publishing, 2016, pp. 325–340. DOI: 10.1007/978-3-319-42019-6_21.

[16] George Baryannis et al. "A Trajectory Calculus for Qualitative Spatial Reasoning Using Answer Set Programming". In: *Theory and Practice of Logic Programming* 18.3-4 (2018), 355—371. DOI: 10.1017/S147106841800011X.

[17] Sotiris Batsakis et al. "Legal Representation and Reasoning in Practice: A Critical Comparison". In: *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018*. 2018, pp. 31–40. DOI: 10.3233/978-1-61499-935-5-31.

[18] Alisa Bilal et al. "Predicting Students' Academic Performance Based on Enrolment Data". In: (Oct. 2020), pp. 54–61.

[19] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists : 50+ essential concepts using R and Python*. 2020.

[20] Siti Bujang, Ali Selamat, and Ondrej Krejcar. "A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning". In: *IOP Conference Series: Materials Science and Engineering* 1051 (Feb. 2021), p. 012005. DOI: 10.1088/1757-899X/1051/1/012005.

[21] Susan Bull and Judy Kay. "Open Learner Models". In: *Advances in Intelligent Tutoring Systems*. Ed. by Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 301–322. ISBN: 978-3-642-14363-2.

[22] Fuyuan Cao, Jiye Liang, and Liang Bai. "A new initialization method for categorical data clustering". In: *Expert Syst. Appl.* 36 (2009), pp. 10223–10228.

[23] Wei-Lun Chao. "Machine Learning Tutorial". In: 2012.

[24] Muhammad Ali Chaudhry and Emre Kazim. "Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021". In: *AI and Ethics* 2.1 (Feb. 2022), pp. 157–165. ISSN: 2730-5961.

[25] Tu Chengsheng, Liu Huacheng, and Xu Bing. "AdaBoost typical Algorithm and its application research". In: *MATEC Web of Conferences* 139 (Jan. 2017), p. 00222. DOI: 10.1051/matecconf/201713900222.

[26] Virginia Dignum. "The role and challenges of education for responsible AI". In: *London Review of Education* 19 (Jan. 2021). DOI: 10.14324/LRE.19.1.01.

[27] Mucong Ding et al. "Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses". In: *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019).

[28] Ivanna Dronyuk, Volodymyr Verhun, and Eleonora Benova. "Non-academic factors impacting analysis of the student's the qualifying test results". In: *Procedia Computer Science* 155 (2019). The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology, pp. 593–598. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.08.083. URL: https://www.sciencedirect.com/science/article/pii/S1877050919309974.

[29] OECD (Organisation for Economic Co-operation and Development). "Recommendation of the Council on Artificial Intelligence". In: (May 2019).

[30] Stefan Edelkamp and Stefan Schrödl. *Heuristic Search - Theory and Applications.* Jan. 2012. ISBN: 978-0-12-372512-7. DOI: 10.1016/C2009-0-16511-X.

[31] Bindhia K. Francis and Suvanam Sasidhar Babu. "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach". In: *J. Med. Syst.* 43.6 (June 2019), 1–15. ISSN: 0148-5598. DOI: 10.1007/s10916-019-1295-4. URL: https://doi.org/10.1007/s10916-019-1295-4.

[32] Zoubin Ghahramani. "Unsupervised learning". In: *Advanced Lectures on Machine Learning*. Springer-Verlag, 2004, pp. 72–112.

[33] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press, 2016.

[34] Siobhan Greatorex-Voith and A. Anand. "A Data-Driven Framework for Identifying High School Students at Risk of Not Graduating on Time [Extended". In: 2015.

[35] Geoff Hayward and Michael Hoelscher. "The Use of Large-Scale Administrative Data Sets to Monitor Progression from Vocational Education and Training into Higher Education in the UK: Possibilities and Methodological Challenges". In: *Research in Comparative and International Education* 6 (Sept. 2011), p. 316. DOI: 10.2304/rcie.2011.6.3.316.

[36] Jiazhen He et al. "Identifying At-Risk Students in Massive Open Online Courses". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. Austin, Texas: AAAI Press, 2015, 1749–1755. ISBN: 0262511290.

[37] HEFCE. "Differences in degree outcomes: Key findings". In: 2014.

[38] Neil T. Heffernan and Cristina Lindquist Heffernan. "The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching". In: *International Journal of Artificial Intelligence in Education* 24.4 (Dec. 2014), pp. 470–497. ISSN: 1560-4306. DOI: 10.1007/s40593-014-0024-x.

[39] Sumyea Helal et al. "Predicting academic performance by considering student heterogeneity". In: *Knowledge-Based Systems* 161 (2018), pp. 134–146. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2018.07.042. URL: https://www.sciencedirect.com/science/article/pii/S0950705118303939.

[40] Arto Hellas et al. "Predicting academic performance: a systematic literature review". In: *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (2018).

[41] Communities & Local Government Minister of Housing. "The English Indices of Deprivation 2019 (IoD2019)". In: 2019.

[42] Joshua Zhexue Huang. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values". In: *Data Mining and Knowledge Discovery* 2 (2004), pp. 283–304.

[43] Zhexue Huang. "CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES". In: (1997).

[44] University of Huddersfield. *Key Facts*. https://www.hud.ac.uk/about/keyfacts/. 2022.

[45] Sadiq Hussain et al. "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA". In: *Indonesian Journal of Electrical Engineering and Computer Science* 9 (Feb. 2018), pp. 447–459. DOI: 10.11591/ijeecs.v9.i2.pp447-459.

[46] Shah Hussain and Muhammad Qasim Khan. "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning". In: *Annals of Data Science* (June 2021). DOI: 10.1007/s40745-021-00341-0.

[47] B.S. Jaber et al. "Predicting success: A comparative analysis of student performance on the surgical clerkship and the NBME surgery subject exam". In: *Surgery Open Science* 1 (Aug. 2019). DOI: 10.1016/j.sopen.2019.07.002.

[48] Matthew C. Jadud. "Methods and Tools for Exploring Novice Compilation Behaviour". In: ICER '06. New York, NY, USA: Association for Computing Machinery, 2006, 73–84. ISBN: 1595934944. DOI: 10.1145/1151588.1151600.

[49] Suhang Jiang et al. "Predicting MOOC performance with Week 1 Behavior". In: *EDM*. 2014.

[50] Botond Kiss et al. "Predicting Dropout Using High School and First-semester Academic Achievement Measures". In: Nov. 2019, pp. 383–389. DOI: 10.1109/ICETA48886.2019.9040158.

[51] Thomas Kotter et al. "Perceived Medical School stress of undergraduate medical students predicts academic performance: an observational study". In: *BMC Medical Education* 17.1 (2017), p. 256. ISSN: 1472-6920. DOI: 10.1186/s12909-017-1091-0. URL: https://doi.org/10.1186/s12909-017-1091-0.

[52] Byoung-Suk Kweon et al. "The link between school environments and student academic performance". In: *Urban Forestry & Urban Greening* 23 (Feb. 2017). DOI: 10.1016/j.ufug.2017.02.002.

[53] E. Lau, L. Sun, and Qingping Yang. "Modelling, prediction and classification of student academic performance using artificial neural networks". In: *SN Applied Sciences* 1 (Aug. 2019). DOI: 10.1007/s42452-019-0884-7.

[54] Eric L. Lee, Tsung-Ting Kuo, and Shou de Lin. "A Collaborative Filtering-Based Two Stage Model with Item Dependency for Course Recommendation". In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2017), pp. 496–503.

[55] Hengxuan Li, Collin Lynch, and Tiffany Barnes. "Early Prediction of Course Grades: Models and Feature Selection". In: *ArXiv* abs/1812.00843 (2018).

[56] Andy Liaw and Matthew Wiener. "Classification and Regression by RandomForest". In: *Forest* 23 (Nov. 2001).

[57] Wei-Yin Loh. "Classification and Regression Trees". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (Jan. 2011), pp. 14 –23. DOI: 10.1002/widm.8.

[58] Rosemary Luckin and Benedict Du Boulay. "Reflections on the Ecolab and the Zone of Proximal Development". In: *International Journal of Artificial Intelligence in Education* 26 (Nov. 2015). DOI: 10.1007/s40593-015-0072-x.

[59] Lonia Masangu, Ashwini Jadhav, and Ritesh Ajoodha. "Predicting Student Academic Performance Using Data Mining Techniques". In: *Advances in Science, Technology and Engineering Systems Journal* 6 (Jan. 2020), pp. 153–163. DOI: 10.25046/aj060117.

[60] Hanan Mengash. "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems". In: *IEEE Access* PP (Mar. 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.2981905.

[61] Ahmed Mueen, Bassam Zafar, and Umar Manzoor. "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques". In: *International Journal of Modern Education and Computer Science* 11 (Nov. 2016), pp. 36–42. DOI: 10.5815/ijmecs.2016.11.05.

[62] Jovial Niyogisubizo et al. "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization". In: *Computers and Education: Artificial Intelligence* 3 (2022), p. 100066. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2022.100066. URL: https://www.sciencedirect.com/science/article/pii/S2666920X22000212.

[63] Emmanuel Papadakis, Song Gao, and George Baryannis. "Combining Design Patterns and Topic Modeling to Discover Regions That Support Particular Functionality". In: *ISPRS International Journal of Geo-Information* 8.9 (Sept. 2019), p. 385. ISSN: 2220-9964. DOI: 10.3390/ijgi8090385. URL: http://dx.doi.org/10.3390/ijgi8090385.

[64] Nestor Pereira. "Using multiple linear regression based on principal component analysis (Factor analysis of mixed data FAMD) for predicting the final score of secondary students from Portugal". In: (Dec. 2019).

[65] Sarah Petersen and Mari Ostendorf. "A machine learning approach to reading level assessment". In: *Computer Speech & Language* 23 (Jan. 2009), pp. 89–106. DOI: 10.1016/j.csl.2008.04.003.

[66] Riccardo Poli, James Kennedy, and Tim Blackwell. "Particle Swarm Optimization: An Overview". In: *Swarm Intelligence* 1 (Oct. 2007). DOI: 10.1007/s11721-007-0002-0.

[67] Howard R. Pollio, W. Lee Humphreys, and James A. Eison. "Patterns of Parental Reaction to Student Grades". In: *Higher Education* 22.1 (1991), pp. 31–42. ISSN: 00181560, 1573174X. URL: http://www.jstor.org/stable/3447152.

[68] Sandra W. Pyke and Peter M. Sheridan. "Logistic Regression Analysis of Graduate Student Retention". In: 23 (Aug. 1993), 44–64.

[69] Carlos Felipe Rodríguez-Hernández et al. "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation". In: *Computers and Education: Artificial Intelligence* 2 (2021), p. 100018. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2021.100018. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000126.

[70] Kerry Schwanz et al. "College Students' Perceptions of Relations with Parents and Academic Performance". In: *American Journal of Educational Research* 2 (Jan. 2014), pp. 1 3 –17.

[71] Scikit-Learn. *Scikit-Learn Documentation*. https://scikit-learn.org/stable. (Accessed on 29/12/2021). 2021.

[72] Bjarne Strøm, Torberg Falch, and Päivi Lujala. "Geographical constraints and educational attainment". In: 11811 (Sept. 2011).

[73] Oliver Sutton. "Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction". In: 2012.

[74] Nguyen Thai-Nghe et al. "Recommender system for predicting student performance". In: *Procedia Computer Science* 1.2 (2010). Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010), pp. 2811–2819. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2010.08.006.

[75] Tamara Thiele et al. "Predicting students' academic performance based on school and socio-demographic characteristics". In: *Studies in Higher Education* 41.8 (2016), pp. 1424–1446. DOI: 10.1080/03075079.2014.974528.

[76] Ruangsak Trakunphutthirak, Yen Cheung, and Vincent Lee. "A Study of Educational Data Mining: Evidence from a Thai University". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 734–741. DOI: 10.1609/aaai.v33i01.3301734.

[77] The Complete University. *University of Huddersfield*. https://www.thecompleteuniversityguide.co.uk/universities/university-of-huddersfield. 2022.

[78] Ralph Utzman, Daniel Riddle, and Dianne Jewell. "Use of Demographic and Quantitative Admissions Data to Predict Performance on the National Physical Therapy Examination". In: *Physical therapy* 87 (Oct. 2007), pp. 1181–93. DOI: 10.2522/ptj.20060222.

[79] Nelis J. de Vos. *kmodes categorical clustering library*. https://github.com/nicodv/kmodes. 2015–2021.

[80] Hajra Waheed et al. "Predicting academic performance of students from VLE big data using deep learning models". In: *Computers in Human Behavior* 104 (2020), p. 106189. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2019.106189. URL: https://www.sciencedirect.com/science/article/pii/S0747563219304017.

[81] Alvin Wang and Michael Newlin. "Characteristics of students who enroll and succeed in psychology Web-based classes". In: *Journal of Educational Psychology* 92 (Mar. 2000), pp. 137–143. DOI: 10.1037/0022-0663.92.1.137.

[82] Christopher Watson, Frederick W. B. Li, and Jamie L. Godwin. "No tests required : comparing traditional and dynamic predictors of programming success." In: *Proceedings of the 45th ACM Technical Symposium on Computer Science Education.* Ed. by J. D. Dougherty et al. New York: Association for Computing Machinery (ACM), Jan. 2014, pp. 469–474. URL: http://dro.dur.ac.uk/19224/.

[83] Darrell Whitley. "A Genetic Algorithm Tutorial". In: *Statistics and Computing* 4 (Oct. 1998). DOI: 10.1007/BF00175354.

[84] Michalis Xenos. "Prediction and Assessment of Student Behaviour in Open and Distance Education in Computers Using Bayesian Networks". In: 43.4 (Dec. 2004), 345–359. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2003.09.005. URL: https://doi.org/10.1016/j.compedu.2003.09.005.

[85] Jie Xu, Kyeong Moon, and Mihaela Schaar. "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs". In: *IEEE Journal of Selected Topics in Signal Processing* PP (Apr. 2017), pp. 1–1. DOI: 10.1109/JSTSP.2017.2692560.

[86] Stephen J.H. Yang et al. "Predicting Students & Academic Performance Using Multiple Linear Regression and Principal Component Analysis". In: *Journal of Information Processing* 26 (2018), pp. 170–176. DOI: 10.2197/ipsjjip.26.170.

[87] Nick Z. Zacharis. "Predicting Student Academic Performance in Blended Learning Using Artificial Neural Networks". In: *International Journal of Artificial Intelligence & Applications* 7 (2016), pp. 17–29.

[88] Li Zhang and Huzefa Rangwala. "Early Identification of At-Risk Students Using Iterative Logistic Regression". In: June 2018, pp. 613–626. ISBN: 978-3-319-93842-4. DOI: 10.1007/978-3-319-93843-1_45.

[89] Yijun Zhao et al. "Predicting Student Performance in a Master's Program in Data Science using Admissions Data". In: International Educational Data Mining Society, 2020.