LD7083
Computing and Digital Technologies Project Dissertation

Predicting and Enhancing Students' Academic Performance using Machine Learning Algorithms

by

Kelechi Uzoukwu

w22078766

Supervisor Name: Rose Fong

A Dissertation

Submitted to Department of Computer and Information Sciences

Northumbria University

In Partial Fulfilment of the Requirements

For Master of Science in Big Data and Data Science Technologies

May 14, 2024

Declaration

I declare the following:

1. that the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic or personal.
2. the Word Count of this Dissertation is 10,964 excluding the bibliography.
3. that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on Blackboard, if deemed appropriate, to allow future students the opportunity to see examples of past dissertations. I understand that if displayed on Blackboard it would be made available for no longer than five years and that students would be able to print off copies or download. The authorship would remain anonymous.
4. I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service. In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.
5. I have read the UNN/CEIS Policy Statement on Ethics in Research and Consultancy and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

Signature: Kelechi Uzoukwu

Date: May 13, 2024

Acknowledgements

I would like to extend my deepest gratitude to my supervisor Rose Fong, for her valuable and constructive feedback throughout the planning and execution of this dissertation. Secondly, I wish to thank all my lecturers at the University of Northumbria, for their guidance and support throughout my master's program. Their expertise and teachings contributed to the successful completion of this dissertation. Finally, I am extremely grateful to my family and friends for their continued love and support in all my endeavours.

Abstract

In recent years, technology has played a critical part in the transformation of the education sector. Although machine learning techniques have changed the way traditional teaching methods are conducted in education settings through improved student engagement and collaboration, academic failure remains a persistent issue amongst higher education students. This report considers a large dataset of 12,411 students and applies five machine learning algorithms including LR, RF, SVR, KNN and DT to predict students' academic performance. Following the successful prediction of students' academic performance, a feature importance analysis was conducted to identify the most influential variables contributing to the models' predictive performance. Findings reveal that Random Forest is the most accurate model for predicting students' academic grades with a prediction accuracy of 84% and 83% for both target variables.

Contents

Table of Figures

Table of Tables

1 Introduction

The demand for data-driven approaches to education are rising rapidly, as education providers seek to improve student academic performance and minimise failure across all levels (Jayaprakash, et al., 2018). The datasets collected by educational institutions and online learning platforms are often too large and unstructured for simple data analysis, leading to the emergence of data mining and machine learning methods specially designed for the education sector (Malik & Jothimani, 2023).

By employing machine learning techniques, this report empowers education providers with accurate predictions to enable them to customise their teaching strategies to address individual student needs and preferences.

1.1 Aims

The primary aim of this report is to contribute to the enhancement of students' educational experiences using machine learning techniques. By developing machine learning models capable of accurately predicting students' academic performance, this report aims to provide educators, administrators, and students with a proactive tool to identify potential challenges and implement early intervention methods.

1.2 Background

According to Times Higher Education (2023), approximately 1 in 37 students in the UK failed their university courses in 2023, with first-year students at a greater risk of discontinuing their studies. As a result, higher education providers in the UK are faced with the challenge of ensuring that academic underperformance and drop-out rates are minimised across all education levels. The implications of students' failure cannot be overemphasised, as increased academic failure and drop-out rates lead to significant economic and social costs for the students, universities and societies at large (Kamala & Thangaiah, 2019). High failure rates often lead to limited employment opportunities and earning potential for the students, reputational risk for the university and a shortage of skilled workers for the society.

The amount of data generated from educational institutions is incredibly complex and massive, and if properly utilised, can serve as a valuable source of information to

improve decision making and student outcomes. Although traditional learning analytics have successfully gained insights into some of the factors that influence student outcomes, their solutions lack prediction systems, limiting their ability to accurately predict future academic performance (Malik & Jothimani, 2023). To manage this issue, it is imperative that higher education institutions leverage advanced data mining techniques such as prediction and clustering, to identify poor-performing students, review academic strategies and implement effective intervention support for students at risk of failing their courses (Malik & Jothimani, 2023; Tatar & Düştegör, 2020).

Educational Data Mining, popularly known as EDM, is the application of data mining and machine learning techniques to extract meaningful information from educational data for improved decision-making capabilities (Badugu & Rachakatla, 2020). As a result of advances in EDM, education providers are leveraging data-driven insights to identify underperforming students and implement timely intervention methods that could improve students' performance. Several studies in this research domain have documented various insights extracted from EDM such as course recommendations, student performance predictions, personalised teaching and learning, identification of atypical learning patterns and enrolment management (Sarra, et al., 2019).

One of the most common outcomes of EDM addressed in existing literature is students' performance prediction, whereby machine learning techniques and algorithms are applied to predict whether a student will pass or fail their course (Sánchez, et al., 2022). These predictive models are particularly useful as they provide policy makers and educators with valuable information that enable them to implement timely interventions and support students at risk of failing (Said, et al., 2023). In this study by Obsie & Adem (2018), various machine learning algorithms were applied to predict the grades of 134 students of Hawassa University. Similarly, these researchers utilized several machine learning techniques including Naïve Bayes, Artificial Neural Network and Logistic regression, to correctly predict binary grades, such as Pass or Fail, with a prediction accuracy of 90% (Naseer, et al., 2020; Nawai, et al., 2021). Another relevant outcome of EDM is the reduction of drop-out rates and subsequent increase in student retention. Using a support vector machine learning model, Chui et al. (2020) predicted students at risk of dropout from a virtual learning program. Similarly, some studies

developed a hybrid model using Decision Trees, Random Forests and the Confidence-Weighted Fusion Voting Classifier (CWFVC) algorithm to improve prediction capabilities and reduce dropout rates of the student population (Malik & Jothimani, 2023; Tatar & Düştegör, 2020).

1.3 Motivation for Study

A large number of studies on the prediction of students' academic performance using machine learning exists in the literature. However, these studies often utilise small datasets and fail to consider non-academic factors that significantly impact a student's academic performance such as family dynamics and socio-economic factors.

This report addresses this issue by analysing extensive datasets beyond academic records and developing a robust predictive model that considers the dynamic and unique characteristics of each student's learning experience. Specifically, this study considers a case of 12,411 students studying towards their professional engineering exams in Colombia. This present work proposes a prediction model that operates from the final semester of secondary school to enable the early identification of students at risk of failure in the final year of their professional engineering program. The early prediction of students' grades may help educators to evaluate students' capabilities, customise their learning plans to the students' needs and support the students in actualising their full potential. Timely predictions may also guide the university administrators to efficiently allocate resources among departments in need of prioritisation.

The contribution of this report is three-fold. First, it utilises a larger dataset to provide a more comprehensive representation of the student population. With 44 variables, this report considers a wide range of student characteristics including demographic information and socio-economic backgrounds, making it adaptable across different education settings. Finally, six different machine learning models are utilised to assess the performance and consistency of predictions across different machine learning models.

Furthermore, this report attempts to answer the following research questions to achieve its aims:

1. Can machine learning models accurately predict students' academic performance?

2. What academic and behavioural factors significantly influence students' academic performance?

3. What are the policy implications of predicting students' academic performance?

1.4 Objectives

- To critically analyse the literature review surrounding the prediction of students' academic performance.
- To develop machine learning models capable of accurately predicting students' academic performance.
- To compare the performance of all models using appropriate evaluation metrics.
- To identify the variables that significantly influence students' academic performance using a feature importance analysis.
- To critically draw impactful conclusions from this research work.
- To explore the implications of performance prediction on policies and educational experiences.

1.5 Research Approach

This report aims to follow the deductive research approach where hypotheses are formulated based on existing theories, and these hypotheses are then systematically tested. This structured approach will ensure clarity in my research process, making it easier to design experiments and analyse results. Given the proposed research approach, this report will implement an experimental research method. The experimental method is appropriate for testing hypothesis and establishing causality between defined variables in a controlled environment (Edgar & Manz, 2017). The proposed outcome of this report will analyse historical data from the university's databases to predict the future academic performance of students.

1.6 Work Done and Results

This report presents the results of the prediction of students' academic performance using machine learning algorithms along with the results of a feature importance analysis. After training multiple models with both numeric and non-numeric variables,

Random Forest emerged as the most accurate model with a prediction accuracy of 83% and 84% for both target variables.

The results of the feature importance analysis revealed past grades, socio-economic factors and students' academic program as the top predictors of future academic performance.

1.7 Structure of the Report

The structure of this report is outlined below:

Chapter 1: Introduction

This chapter highlights the research report's aims, objectives and its' significance/contribution to the broader education sector and society.

Chapter 2: Literature Review

This chapter critically evaluates the existing literature within this report's research area.

Chapter 3: Research Methodology

This chapter outlines the proposed research approach/method used in this research report.

Chapter 4: Model Development

This chapter explores different pre-processing techniques and trains the machine learning models. It also presents the model prediction results using appropriate visualisations.

Chapter 5: Results and Discussion

This chapter presents the key findings from the predictive analysis and discusses results in alignment with the research objectives. It also states the

Chapter 6: Conclusions and Recommendations

This chapter concludes the research report and proposes recommendations to support future development in the research area.

2 Literature Review

This chapter critically evaluates recent studies in the research area and highlights the advantages and limitations of previous researchers' works. It also examines the various methods, approaches and techniques used to predict students' performance within education settings. Finally, this chapter discusses any observed gaps in the existing research methodologies and highlights the importance of conducting additional research to support modern education settings in students' performance predictions.

The background research surrounding students' performance prediction is multifaceted and can be explored through multiple perspectives including the different machine learning algorithms used to predict students' performance, investigating the key variables that influence the models' prediction abilities and the implementation of early intervention systems.

2.1 Predictive Modelling for Students' Academic Performance

Previous researchers have utilised different approaches to predict students' performance, identify underperforming students and facilitate timely intervention methods (Waheed, et al., 2020).

The study conducted by Marquez-Vera, et al. (2013) proposed the use of induction rules and decision trees to predict the likelihood of students failing their grades in secondary school. Although their study produced satisfactory prediction results, it ignores the inclusion of non-academic data that could affect students' performance such as socio-economic background and motivation. Their proposed model also failed to identify students at risk of failure in a timely manner, rendering it insufficient for implementing early intervention systems.

Khobragade & Mahadik (2015) presented a similar approach in their prediction of students' failure by using White-Box classifiers such as Naïve Bayes and Decision Trees. By analysing a diverse dataset with both academic and non-academic data, this study effectively captured the dynamic influence of socio-economic factors on students' performance. However, the integrity of their data source ought to be taken with caution as surveys and other self-reported information are open to recall bias.

Although the Naïve Bayes model achieved a prediction accuracy of 87.1%, more effort should be invested in the early prediction of students at risk of failure in order to implement timely corrective measures and prevent failure before it is too late.

Yukselturk, et al. (2014) predicted the possibility of students dropping out from their courses using Decision Tree, Naïve Bayes, Neural Network and K-Nearest Neighbors. Their data was collected through online questionnaires, and it included only non-academic data such as age, gender, occupation and educational level. Although the inclusion of such data in their model explains the influence of diverse characteristics on student performance, their work may have overlooked critical predictors of students' performance by excluding assessment grades and other similar scores. The study's heavy reliance on questionnaires and surveys also raises questions on the validity of their model, as such data sources are prone to recall bias and recency bias. Moreover, their model's inability to predict students at risk of dropout early in their course program renders it insufficient for early intervention strategies.

The study by Bydžovská (2016) addressed the issue of early intervention by predicting the final grades of students at the beginning of the semester and identifying students at risk of failure. They employed various algorithms such as Support Vector Machine, Random Forest, Decision Tree and Naïve bayes, and credited SVM as the best performing model due to its ability to correctly identify almost half of the unsuccessful students. The study also considered the influence of socio-economic variables and social behaviour in predicting students' performance, making it a strong early intervention predictive model. However, this work could benefit from detailed discussions on the reasons for selecting the machine learning algorithms employed in the analysis.

In this work by Chen & Zhai (2023), seven machine learning models were applied to three large datasets to predict students' likelihood of graduating and receiving an internship. They reported that Random Forest demonstrated the highest prediction accuracy across all datasets and is the most suitable for both binary and multi-classification predictions. Although this study demonstrated promising results in predicting graduation rates, it could benefit from discussing the criteria used to select the machine learning algorithms.

This study by Alturki & Alturki (2021) aimed to predict the final grade of 300 female students at Princess Nourah Bint Abdulrah- man University in Saudi Arabia using six machine learning models. They concluded that Naïve Bayes performed best in students' final grade predictions with accuracy of 67%, while Random Forest performed better in predicting honorary students with an accuracy of 90%. Although this study contributed valuable insights, the generalisation of its findings are only applicable to a specific context due to its failure to include male students. The study is also limited by its small dataset and failure to address the imbalance in grades' distribution.

Beaulac & Rosenthal (2019) applied Random Forest and Logistic Regression to predict the likelihood of 38,842 students completing their program at the University of Toronto. They reported that Random Forest model achieved the best accuracy at 78.84%. However, their study is limited by the absence of socio-economic and demographic variables, as the sample dataset only comprised of students' grades. In this study by (Hussain, et al., 2022), Decision Tree and Support Vector Machine algorithms were utilised to predict the grades of 520 students with accuracy results of 69% and 78% respectively. However, the study's small dataset and high number of features (29) may be the reason for the moderate prediction performance of both models.

Deploying deep learning techniques to predict students at risk of failure is a fairly new area of research, and it involves developing a multi-layered model to learn interpretations from raw data (Waheed, et al., 2020). In this study by (Olabanjo, et al., 2022), a Neural Network was developed to predict the academic grades of 1927 students in a Nigerian secondary school, achieving an accuracy of 86.59%. However, their proposed model was not used to forecast future grades of students, as the dataset contained grades from first year to final year. Their study also failed to provide justifications for the selected features used to train the model. Using the Bidirectional Long Short Term Model, Uliyan et al. (2021) identified students who were at risk of leaving the College of Computer Science and Engineering in Saudi Arabia. Although the model achieved a prediction accuracy of 90%, it could benefit from a broader set of features.

2.2 Key Predictors of Students' Academic Performance

Several researchers have conducted numerous studies aimed at understanding the pattern of variables that influence the prediction of students' academic performance in educational institutions.

Using different classifiers, Malini & Kalpana (2021) identified the most important features that improved the prediction of students' academic performance. They concluded that economic background attributes such as family size, parents' cohabitation status, student's health, parent's education, and parents' job are essential predictors of students' academic success. Although the study explored the influence of socio-economic factors on student performance, it could benefit from replicating the findings across different educational settings to ensure the model's generalisation.

In another study by Vaarma & Li (2024), three machine learning models were used to predict students at risk of dropout from a Finnish University. Data was collected from a learning management system (Moodle) and the results revealed "accumulated credits", "number of failed courses" and "Moodle activity count" as the most important features in the predictive models. This suggests that data from online learning platforms such as Moodle, has significant predictive power and should be supplemented with transcripts and demographic data when predicting students at risk of dropout.

Nachouki et al. (2023) employed a Random Forest model to determine the most significant variables that predicted 650 undergraduate computing students' course performance. They credited grade point average (GPA) and high school final grades as the top predictors of students' course grades. Similarly, Beckham et al. (2023) leveraged a Random Forest model and identified age and past academic grades as the most significant predictors of students' academic success. In this paper, Rashid & Aziz (2016) developed a Neural Network with 4 neurons and concluded that the most important predictive variables of students' performance are student's course and course tutor. These studies further emphasise the importance of early intervention because students with low GPA could be identified as high-risk and promptly assisted to prevent failure in their courses.

Similarly, Martins et al. (2023) conducted a feature importance analysis using Random Forest and Support Machines and credited cumulative GPA as an important feature in the prediction of students at risk of dropout. Using Random Forest, Matz et al. (2023) identified grade point average (GPA) and class attendance as the top predictors of students' dropout. Song et al. (2023) employed six different machine learning classifiers to predict students at risk of dropout at a South Korean university. They reported number of scholarships and tuition fees as the most important features to predict students' dropout. In this study by Kiss et al. (2019), extreme gradient boosting and Neural Network were utilised to predict students' dropout from a database of over 10,000 Hungarian university students. They credited age and GPA as the most important predictors of dropout from the university. Yu, et al. (2021) applied Naïve Bayes and Decision Tree classifiers to predict students' dropout from a US university and identified first-year academic grades as the top predictor of the students' final year academic performance. However, they stated that students' gender, minority status and financial status were not important features in dropout predictions. Additionally, Sassirekha & Vijayalakshmi (2022) applied several machine learning models to predict 500 students' academic progression status. They identified 10 variables that are significantly correlated with students' final grades including students' continuous assessment grades, allocated study time, attendance records and high school grades.

2.3 Implementation of Early Intervention Systems

Following the development of predictive models to identify students at risk of failure or dropout, early intervention systems are usually implemented to support students in improving their academic grades and possibly reduce dropout rates.

Researchers at Purdue University developed 'Course Signals', a predictive student success algorithm which makes predictions based on students' previous academic grades (High school GPA, SAT scores), students' characteristics (age, gender, residency) and students' interaction on their Learning Management System. Based on its predictions, the algorithm displays appropriate signals on each student's dashboard whereby a red light signifies a high likelihood to fail the course; yellow indicates a moderate potential to fail; green indicates a low likelihood to fail the course. The early signals enable tutors to implement intervention methods such as personalised learning plans, face-to-face meetings and additional teaching sessions to

support the students in achieving success at their course. The research results revealed a significant increase in grades among courses that implemented 'Course Signals' as well as a significant decrease in low grades and students' withdrawals from their courses. Specifically, it was reported that in a course with 220 students, 55% of those identified as "high-risk" after one course completion moved into "moderate risk" after interventions and 70% of students identified as "moderate risk" moved to "low risk" category. Additionally, students who enrolled in at least one class with Course Signals had significantly higher retention rates than students without Course Signals classes (Arnold & Pistilli, 2012). This study highlights the importance of implementing early intervention systems in education settings, as evidenced by the students' improved academic performance in their courses.

In this study, Bañeres et al. (2020) developed an early prediction system that correctly identifies students at risk of failing their courses and provides timely feedback. Their Gradual At-Risk (GAR) model makes predictions based on data from two different sources: Learning Management Systems (data on students' interaction, navigation and engagement within online learning spaces) and Institutional Warehouse Systems (data from CRM, ERP on students' enrolment, accreditation, assessments). Based on its predictions, the feedback system assigns four signals to each student and sends personalised messages or warnings to help them succeed in their subsequent assessments. A green signal indicates a successful student, triggering an automated congratulatory message to be sent; a yellow signal indicates a student at possible risk of failing in the future and recommendations are sent to such students. Red and Black signals indicate students who have failed the course and are likely to drop out. For one of the courses, the green signal correctly identified more than 77% of the students and the red and black signals were correctly assigned to 96.72% and 98.44% students, respectively. Additionally, the dropout rates decreased from 53% to 50% by the end of the semester. The findings in this study further emphasise that early prediction of student performance is able to significantly improve students' academic performance as well as prevent dropouts if intervention methods are implemented early. This study could benefit from exploring the long-term effects of implementing intervention systems by investigating its' participants over a prolonged period of time.

Jayaprakash et al. (2014) developed an early feedback system to correctly identify students at risk of not completing their course at Marist College, New York. Based on the model's ability to generate failure probability scores, students were categorised into different risk categories including "low risk" with a probability of 0-50%; "medium risk" and "high risk" with probabilities of 50% and above. In this study, students who were identified as moderate to high risk were subjected to two different intervention methods: Awareness Messaging, whereby affected students received an automated message that informed them of their progress, along with recommendations to improve their chances of success in the course; Online Academic Support, whereby students were encouraged to join the institution's online support portal with access to lecture materials, practice assessments, peer mentors and professional support staff. Although this study provided valuable insights into different intervention methods, further research ought to be invested in assessing the sustainability and scalability of such methods across different education settings.

In summary, this chapter has presented the results of several studies that have investigated the prediction of students' academic performance using different data mining techniques, with each paper proposing a different approach to solve the problem. This report differs from previous work by utilising a larger dataset of 12,411 students, to provide a more comprehensive representation of the student population. With 44 variables, this report considers a wide range of student characteristics including demographic information and socio-economic backgrounds, making it adaptable across different education settings. Additionally, this report employs six different machine learning models to assess the performance and consistency of predictions across different target variables.

3 Research Methodology

This chapter discusses the proposed research methods utilised in this report to predict students' academic performance. It explains the data collection process and the techniques used to pre-process the dataset. Furthermore, this chapter discusses the selection criteria for the machine learning algorithms and evaluation metrics employed in this report with justifications. The chapter concludes by presenting the time management and risk assessment strategies adopted by the author to ensure a successful completion of this project within the required deadline.

3.1 Data Collection

This report intended to utilise the University of Northumbria's student dataset, and after a careful consideration of data privacy and ethics concerns, a public dataset was recommended to mitigate privacy risks. Therefore, the data used in this report is secondary data downloaded from Mendeley, a public data repository. It is a large dataset with 12,411 rows and 44 variables, making it suitable for developing accurate predictive models that are indicative of students' academic performance. The selected dataset comprises of both academic, demographic and socio-economic attributes such as age, gender, assessment grades, study program and economic aid.

All data preparation and prediction tasks are performed using Python because it has powerful libraries like Pandas, NumPy and Scikit-learn which are great for data manipulation, modelling and analysis tasks. Pandas is a great choice for data cleaning because it offers DataFrames that make it easy to handle missing values, outlier and duplicates in a dataset (W3 Schools, 2022). The dataset is uploaded in Jupyter notebook and read as a Python Pandas data frame to begin pre-processing and analysis tasks.

3.2 Data Pre-processing

Raw data in its collected state is often inconsistent, incomplete and inaccurate, and failure to prepare it for further analysis minimises the generalization and prediction performance of supervised machine learning algorithms (Alexandropoulos, et al., 2019). The data is pre-processed using the techniques below:

3.2.1 Outlier Detection

An outlier is a data point that is abnormally distanced or significantly different from other observations in a random sample. Outliers can misrepresent the spread and central tendency of the dataset, leading to the wrongful detection of nearest neighbours in some classification models (Alexandropoulos, et al., 2019). They can also distort the performance of machine learning models like linear regression, leading to inaccurate predictions. There are several methods for detecting outliers, but this report will check for outliers by visually inspecting the dataset's distribution using histograms. Other methods include the neighbourhood based algorithm proposed by (Chen, et al., 2010) which weights data points based on their sum of distance from each other and identifies the points with the highest weights as outliers.

3.2.2 Handling Missing Values

Missing values in datasets lead to biased results, therefore it is a major problem that must be tackled during data pre-processing (Kantardzic, 2019). Although some machine learning algorithms like Naïve Bayes (NB) can perform well with missing values, others like k nearest neighbours (k-NN) require a careful handling of missing information in order to perform accurate predictions. Some of the common methods of handling missing values are discussed below:

- Highest occurring value: This involves filling the missing values with the most common value in a categorical or numerical attribute.
- Mean substitution: This involves imputing the missing values with the average value of a numerical attribute.
- Regression: This involves developing regression models to predict missing values based on the past values within the same numerical class (Kantardzic, 2019).

3.2.3 Encoding Categorical Variables

As discussed earlier, this report contains several features with different data types such as integer and object. Thus, to ensure compatibility with machine learning algorithms, 'label encoding' was applied to transform categorical variables into numeric format. The reason this report uses 'label encoding' instead of 'one-hot encoding' is because the latter significantly increases the number of feature

dimensions. Whereas label encoding directly transforms the features into exact numeric values that can be fit into regression models for machine learning.

3.2.4 Feature Scaling

Feature scaling is an important pre-processing step which involves rescaling each feature to have a mean of 0 and a standard deviation of 1 (Scikit-Learn, 2023). Large datasets, similar to the dataset used in this report, often contain features with large differences in units, range and magnitude, therefore, feature scaling is performed to ensure all feature values have the same degree of influence on the target variable's predictions. Feature scaling is also performed in this report to improve the prediction performance of models like Support Vector Machines (SVM) and k-NN algorithms, which cannot perform efficiently in the presence of unscaled features (Alexandropoulos, et al., 2019).

3.2.5 Feature Selection

Feature selection is the process of selecting and retaining only the relevant variables necessary for building accurate machine learning models (Scikit-Learn, 2023). There are 44 features in this report's dataset, including some redundancies which could distort the prediction results. As a result, feature selection is employed in this report to enhance the selected models' learning and prediction performance and provide insights into the features that mostly contribute to the performance of the model. Additionally, several machine learning algorithms such as k-NN, SVMs and neural networks do not perform well with irrelevant or redundant features, as it can render their training performance inefficient and impractical (Alexandropoulos, et al., 2019).

3.3 Machine Learning Algorithms

Machine learning is categorised as supervised learning, whereby the model learns from labelled data to find patterns in the dataset, and unsupervised learning whereby the model extracts hidden patterns from unlabelled data without explicit guidance or instruction. This report trains various supervised machine learning algorithms using a set of input variables to predict the value of students' final year grades. To achieve this, each model is fitted to the training data to learn patterns, then it makes predictions on the unseen test data and calculates the regression metrics for easier evaluation. Several regression algorithms are capable of predicting students' academic

grades; however, the literature review confirms that there is no single algorithm that works best in all prediction contexts. Table below summarises the characteristics of both regression and classification models.

Table 1: Comparison between model types

Characteristics	Classification Models	Regression Models
Target Variable	Discreet categorical	Continuous numerical
Input Variable	Categorical or numerical	Continuous numerical
Model Output	Class labels (e.g. Pass or Fail)	Numerical grades (e.g. 80%)
Algorithms	Naïve Bayes, Random Forest, Decision Trees, Support Vector Machines.	Linear Regression, XGBoost, Neural Networks, Random Forest Regression.
Benefits	• Suitable for predicting students who will pass or fail their courses. • Can handle imbalanced datasets • Suitable for datasets with distinct classes.	• Suitable for predicting students' academic grades. • Coefficients are easy to interpret. • Predictions are precise.
Limitations	• May not perform well when classes are not well defined. • Output is not as precise as regression.	• Sensitive to outliers • Assumes an existing linear relationship between variables.
Applications	Classify students' performance into two categories (Pass, Fail)	Predicting students' numerical grades.

Regression models are utilised in this report to determine the relationship between the target variables and independent features. They are also applied in this report because they are designed to predict numerical values, as opposed to classification models that predict categorical classes. Additionally, regression metrics such as Mean

Squared Error (MSE), Mean Absolute Error (MAE) and R-Squared are well-suited to evaluate the performance of regression models because they are easy to interpret, and they provide insights into the relationships between a target variable and its features. Therefore, the following regression models are applied in this report to accurately predict the final year grades of 12,411 students.

3.3.1 Linear Regression (LR)

Linear regression assumes a linear relationship between multiple input variables (features) and output (target variable) as shown below:

$$y = \alpha_1 x_1 + \alpha_2 x_2 \dots \dots \dots . + \alpha_n x_n + e$$

Where $\alpha_1, \alpha_2, \alpha_n$ are the coefficients of the input variables (x_1, x_2, x_n) respectively, y is the target variable and e is the error rate. Linear regression works by using pairs of inputs and outputs to find the line that best fits the known pairs, while minimising the error rate (Scikit Learn, 2023). Linear regression is applied in this report for several reasons including its simplicity and ability to handle large datasets with many features. They can also provide insights into the features that are important in predicting the target variable.

3.3.2 Decision Tree (DT)

A Decision Tree builds a model similar to the structure of a tree with decision nodes and leaf nodes. A decision node has two or more branches which represent a feature or attribute, and the branch represents the result of a decision with each leaf node representing the final decision on the numerical target variable (Tatar & Düştegör, 2020). This algorithm works by splitting the data into subsets by asking 'true' or 'false' questions until the model learns enough to make a single prediction. Decision Tree is applied in this report because it is easy to understand and implement, giving it a stronger advantage over algorithms. Specifically, this model will be trained on historical data such as past grades, attendance, socio-economic characteristics and these features will be used as decision nodes to predict the possible academic grades of students. This model was recently used in this study by (Matzavela & Alepis, 2021) to create dynamic tests for assessing student performance.

3.3.3 Support Vector Regression (SVR)

Support Vector Regression is built on the foundation of Support Vector Machines, and it works by finding the best fit line or hyperplane within a threshold value that can predict students' academic grades (Scikit Learn, 2024). SVR is applied in this report because it is robust to outliers and easy to implement. It also identifies complex patterns and generalises well to unseen data with high prediction accuracy. SVR was used in this study by (Cardonaa & Cudneya, 2019) to predict student retention and completion rates among STEM community college students.

3.3.4 K-Nearest Neighbours (KNN)

K-Nearest neighbour regression is a simple algorithm that works by predicting a target value based on the average value of its nearest neighbors. Where the nearest neighbours are found by identifying a predefined number of training data points with the closest distance to the new data point (Scikit Learn, 2024). Also referred as a lazy learning algorithm, KNN does not have a specific training process for modelling. KNN regressor is applied in this report because it is easy to understand and relatively resistant to outliers since it only considers the nearest data points that are closest to the predicted value (Scikit Learn, 2024)

3.3.5 Random Forest (RF)

Random Forest is an ensemble method of supervised learning that is widely used for non-linear regressions. The idea behind Random Forest models is that they rely on using multiple decision trees to predict output (Scikit Learn, 2024). The final predicted value is the average of the outputs from different decision trees in the random forest. Random Forest is applied in this report because it prevents overfitting by building smaller trees from random subsets of features and combines their decisions to create a more accurate decision (Scikit Learn, 2024). It is one of the most accurate machine learning algorithms which performs well with large datasets. Random Forest was recently applied in this study by (Huynh-Cam, et al., 2021) to predict poor performing students among a sample of 2407 first-year students.

3.4 Model Evaluation Metrics

Evaluation metrics play a key role in determining the machine learning models that perform well during training. The prediction performance of the models in this report are evaluated using the following regression metrics:

Mean Squared Error (MSE): This measures the squared differences between actual and predicted values. A lower MSE indicates a good fit model whereby the predicted values are closer to the actual values (Scikit Learn, 2024).

$$MSE = \frac{\sum_{i=0}^{N} |Y_i - \widehat{Y}_i|}{N}$$

Mean Absolute Error (MAE): This measures the average of the absolute differences between actual and predicted values. A lower MAE indicates a good fit model, whereby the predicted values are closer to the actual values (Scikit Learn, 2024).

$$MAE = \frac{\sum_{i=0}^{N} (Y_i - \widehat{Y}_i)^2}{N}$$

Root Mean Squared Error (RMSE): This is the square root of MSE, and it measures the average difference between actual and predicted values. Its output is reported in the same scale as the target variable, making it more interpretable than MSE (Scikit Learn, 2024).

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N} (Y_i - \widehat{Y}_i)}{N}}$$

R-Squared (R2): This metric measures the proportion of the predicted variable that is explained or influenced by its features. An R-squared of 1 indicates that the model perfectly predicts the target variable and an R-squared of 0 indicates the model does not explain any variability in the target variable (Scikit Learn, 2024).

$$R^2 = 1 - \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2}$$

Where Y_i is actual value; $\widehat{Y}_i$ is predicted value, $\bar{Y}$ is mean of actual values and N is number of data points (Scikit Learn, 2024).

3.5 Project Plan

Drafting a project plan using a Gaant Chart has been an instrumental step in this dissertation journey because it enables a seamless tracking of both academic and personal goals. The Gaant chart ensures that the author's strengths are strongly aligned with the demands of the dissertation, ensuring a strategic approach to research. The Gaant chart also helps the author manage their time effectively and ensures that each stage of the dissertation receives adequate amount of attention and resources. The Gaant chart contains 29 tasks in total to be completed by a self-imposed deadline of May 8th, 2024. The deadline ensures that the final dissertation chapters are completed and submitted on time.

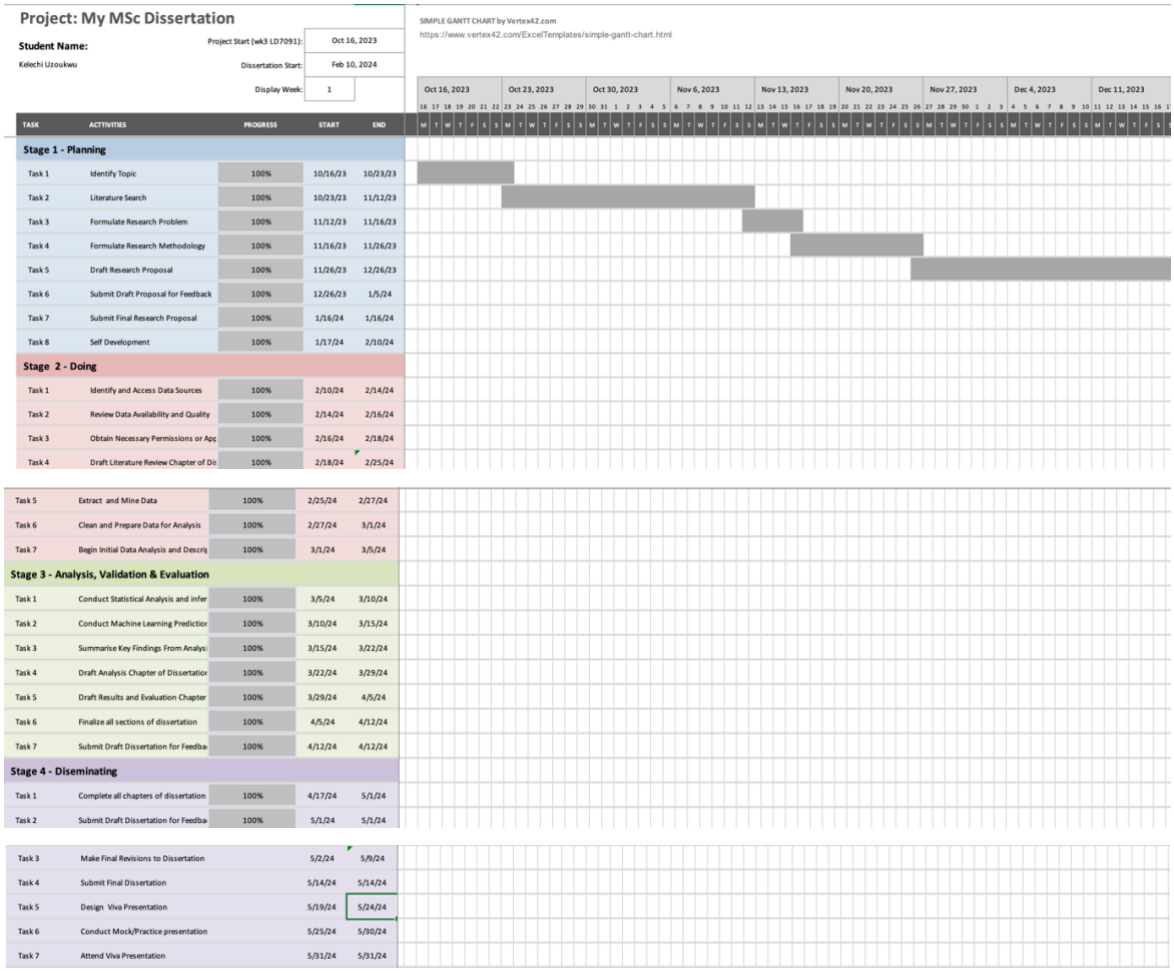


Figure 1: Gaant Chart

Risk Assessment Form

Northumbria University NEWCASTLE

Date: 27/12/2023	Student Name: Kelechi Uzoukwu
Module: LD7083	Dissertation Title: **Optimizing Educational Experiences through Machine Learning: Predicting and Enhancing Students' Academic Performance**

Item No.	Activity, Equipment, Materials, etc. required for dissertation	Hazard	Persons at risk	Severity	Likelihood	Risk Rating H 20-36 M 12-18 L 1-10	Control Measures Required	Final Result*
1	Data	Data Loss	Researcher	4	3	12 (M)	• I will regularly back up data and use reliable storage solutions.	4x3=12 (M)
2	Data	Poor quality or unreliable secondary data	Researcher	4	4	16 (M)	• Thoroughly assess the quality of selected data sources, validate data through cross-referencing, and document any limitations.	4*4 = 16 (M)
3	Data	Unavailability of key data points in selected datasets.	Researcher	3	3	9 (L)	• Verify data availability during the initial stages, plan for potential gaps, and explore supplementary sources.	3*3 = 9 (L)
4	Research Ethics Review	Violation of research ethics guidelines	Researcher, University	5	3	15 (M)	• I will adhere strictly to ethical guidelines and seek approval from the university.	5*3 = 15 (M)
5	Dissertation Submission	Failure to submit or late submission of dissertation.	Researcher, University	4	2	8 (L)	• I will establish a realistic timeline, set project deadlines and monitor progress regularly.	4*2 = 8 (L)
6	Laptop, Analysis and Programming Software	Technical Issues with software or tools	Researcher	3	4	12 (M)	• I will test technology regularly. • Update software requirements. • Acquire backup laptop	3*4 = 12 (M)
7	References and Citations	Plagiarism	Researcher, University	4	4	16 (M)	• Understand and respect intellectual property rights, seek permissions when necessary. • Use correct Harvard referencing when necessary.	4*4 = 16 (M)

UNN/H&S/RA/F/V2 Risk Assessment Form.docx

Risk Assessment Form

Northumbria University NEWCASTLE

8	Dissertation Scope	Expanding the research scope beyond initial plans.	Researcher	3	4	12 (M)	• Clearly define project scope, regularly review objectives	3*4 = 12 (M)
9	Communication Channel	Breakdown in communication with supervisors or peers	Researcher	5	4	20 (H)	• Schedule regular meetings with supervisor. • Seek feedback for clarity.	5*4 = 20 (H)
10	Health Challenges	Short-term illness (e.g. Covid 19)	Researcher	4	2	8 (L)	• Prioritize health, maintain work-life balance and seek support if needed.	4*2 = 8 (L)
11	Burnout	Exhaustion and Burnout	Researcher	3	3	9 (L)	• Manage workload, take breaks, prioritize self-care	3*3 = 9 (L)

Does this Risk Assessment Require Further Specific Risk Assessment:

Manual Handling: Y/N Please list reference No:	COSHH: Y/N? Please list reference No:	PUWER: Y/N? Please list reference No:	DSEAR: Y/N? Please list reference No:	Young Persons: Y/N? Please list reference No:	New & Expectant Mothers: Y/N? Please list reference No:

**Should reduce to the acceptable region (L: 1-10) after the control measures*

To be completed by the person undertaking the risk assessment

Student Name: Kelechi Uzoukwu Student ID: w22078766

Signature: *Kelechi Uzoukwu* Date: 27/12/2023

Figure 2: Risk Assessment Form

In summary, this chapter has presented the step by step research methodologies employed in the report. It also presented the time management and risk assessment strategies adopted by the author to ensure a successful completion of the project within the required deadline.

4 Model Development

This chapter discusses the steps taken in this report to predict students' academic performance. It begins by describing the dataset and explaining the various pre-processing techniques employed in this report. Then it discusses the training process of the learning models and concludes by evaluating the models' prediction performance using selected evaluation metrics.

4.1 Dataset

The dataset used in this report was downloaded from Mendeley data repository. It contains the academic, demographic and socio-economic information of 12,411 engineering students at the Technological University of Bolívar (UTB). The dataset has 44 variables and includes a combination of both categorical and numerical variables. The academic component of the dataset presents the results of standardised national assessments developed by the Colombian Institute for the Evaluation of Education (ICFES). The standardised tests, popularly known as Saber Test are recorded at two different periods in a student's life. Saber11 is taken during the final year or 11th grade of high school, and it consists of 5 tests: Mathematics, Critical Reading (CR_S11), English Language (ENG_S11), Natural Sciences and Citizenship Competencies. SaberPro is aimed at students in their final year of their professional engineering program. It assesses Critical Reading (CR_PRO), Quantitative Reasoning, English Language (ENG_PRO), Writing, Citizenship Competencies and Formulation of Engineering Projects (Mendoza-Mendoza, et al., 2023). These variables (final year PRO grades) will form the basis of the target variables to measure students' academic performance. the Table 1 displays a brief description of the numerical variables while Table 2 describes the categorical variables of the dataset.

All data preparation and prediction tasks are performed using Python because it has powerful libraries like Pandas, NumPy and Scikit-learn which are great for data manipulation, modelling and analysis tasks. Pandas is a great choice for data cleaning because it offers DataFrames that make it easy to handle missing values, outlier and duplicates in a dataset (W3 Schools, 2022). The dataset is uploaded in Jupyter

notebook and read as a Python Pandas data frame to begin pre-processing and analysis tasks.

Table 2: Description of numerical variables

Variable	Description	Data Type
MAT_S11	Mathematics	Integer
CR_S11	Critical Reading	Integer
CC_S11	Citizenship Competencies	Integer
ENG_S11	English	Integer
BIO_S11	Biology	Integer
QR_PRO	Quantitative Reasoning	Integer
CR_PRO	Critical Reading	Integer
CC_PRO	Citizenship Competencies	Integer
ENG_PRO	English	Integer
WC_PRO	Written Communication	Integer
FEP_PRO	Formulation of Engineering Projects	Integer
G_SC	Global Score	Integer
PERCENTILE	Percentile	Integer
SECOND DECILE	Second Decile	Integer
QUARTILE	Quartile	Integer
SEL	Socioeconomic Level	Integer
SEL_IHE	Socioeconomic Level of The Institution of Higher Education	Integer

Note: S_11 refers to secondary school test and S_PRO refers to the professional test

Table 3: Description of categorical variables

Variable	Description	Data Type
GENDER	Student's gender	Object
EDU_FATHER	Father's education	Object
EDU_MOTHER	Mother's education	Object
OCC_FATHER	Father's occupation	Object
OCC_MOTHER	Mother's occupation	Object
STRATUM	Income Stratum/Level	Object

SISBEN	Economic aid granted to low-income families	Object
PEOPLE_HOUSE	Number of people living in the household	Object
INTERNET	Internet	Object
TV	Tv	Object
COMPUTER	Computer	Object
WASHNG_MCH	Washing machine	Object
MIC_OVEN	Microwave oven	Object
CAR	Car	Object
DVD	DVD	Object
FRESH	Fresh	Object
PHONE	Phone	Object
MOBILE	Mobile phone	Object
REVENUE	Household revenue	Object
JOB	Job	Object
SCHOOL_NAME	Name of High school	Object
SCHOOL_NAT	Nature of high school	Object
SCHOOL_TYPE	Type of high school	Object
COD_SPRO	Code for Saber Pro test	Object
UNIVERSITY	University name	Object
ACADMIC_PROGRAM	Academic university program	Object
COD_S11	Code for Saber 11 test	Object

The variables DVD, PHONE, MOBILE, COMPUTER, MIC_OVEN, WASHING_MCH, INTERNET, TV, CAR, DVD, FRESH, refer to the presence of such appliances in the household and are identified by Yes / No.

4.2 Data Pre-processing

Data pre-processing refers to the techniques applied to a dataset to improve its quality, consistency and representation (Alexandropoulos, et al., 2019). By employing the most widely used methods of data pre-processing, this report aimed to improve the generalisation performance of its supervised machine learning algorithms.

```
# read the dataset as a pandas dataframe with latin1 encoding type

encoding_type = 'latin1'
student = pd.read_csv ('/Users/kelechiuzoukwu/Downloads/Research Methods/Disso/data_academic_performance.csv',
                       encoding=encoding_type)
student.shape
```

```
(12411, 45)
```

Figure 3: Read dataset as pandas dataframe

```
In [8]: student.sample(5)
```

Out[8]:

	COD_S11	GENDER	EDU_FATHER	EDU_MOTHER	OCC_FATHER	OCC_MOTHER	STRATUM	SISBEN	PEOPLE_HOUSE	Unnamed: 9	...
2561	SB11201220334047	M	Complete professional education	Complete professional education	Auxiliary or Administrative	Home	Stratum 3	It is not classified by the SISBEN	Four	NaN	...
2141	SB11201220300992	M	Complete professional education	Complete professional education	Executive	Executive	Stratum 3	It is not classified by the SISBEN	Three	NaN	...
10321	SB11201320453229	M	Complete professional education	Complete professional education	Executive	Executive	Stratum 2	It is not classified by the SISBEN	Four	NaN	...
11666	SB11201420058436	F	Incomplete technical or technological	Postgraduate education	Executive	Executive	Stratum 4	It is not classified by the SISBEN	Five	NaN	...

Figure 4: sample dataset

As seen in figure above, there are 12,411 rows and 45 columns. However, an irrelevant column 'Unnamed: 9' was identified and subsequently removed.

4.2.1 Deleting Unwanted Columns

The column 'Unnamed: 9' was erroneously created when reading the dataframe and it was subsequently deleted because it was empty and not needed for this report's analysis.

```
# Delete unwanted columns

student.drop(columns='Unnamed: 9', inplace=True)

student.shape
```

```
(12411, 44)
```

Figure 5: delete unwanted column

Following the removal of the unwanted column, there are now 21,411 rows and 44 columns in the student dataset.

4.2.2 Outlier Detection

Outliers can misrepresent the spread and central tendency of the dataset. They can also distort the performance of machine learning models like linear regression, leading

to inaccurate predictions. This report used histograms to check for outliers and upon visual inspection, there appears to be no outliers as seen below.

```
# Set the style for better visualization
sns.set(style="whitegrid")

numerical_variables = ['MAT_S11',
'CR_S11',
'CC_S11',
'ENG_S11',
'BIO_S11',
'QR_PRO',
'CR_PRO',
'CC_PRO',
'ENG_PRO',
'WC_PRO',
'FEP_PRO',
'G_SC'
]

# Create subplots for distribution plots
fig, axes = plt.subplots(nrows=5, ncols=4, figsize=(20, 20))
fig.subplots_adjust(hspace=0.5)

# Plot histograms for numerical variables
for i, var in enumerate(numerical_variables):
    row = i // 4
    col = i % 4
    sns.histplot(student[var].dropna(), ax=axes[row, col], kde=True, color='skyblue', bins=30)
    axes[row, col].set_title(f'Distribution of {var}')

# Show the plots
plt.show()
```

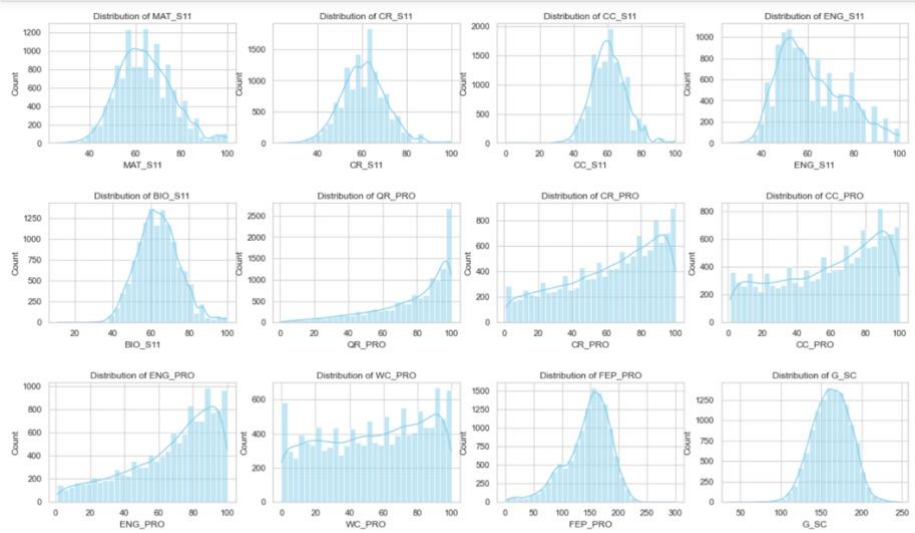


Figure 6: Outlier detection using histograms

4.2.3 Missing Values and Duplicates

Handling missing values is an important step in data-processing and subsequent modelling because they can skew statistical results and produce inaccurate predictions.

```
# check for missing values

missing_values = student.isnull().sum().sum()
print("Number of Missing Values:", missing_values)
```

```
Number of Missing Values: 0
```

Figure 7: Check for missing values

```
In [9]: # check for duplicates

        num_duplicates = student.duplicated().sum()
        print("Number of duplicated rows:", num_duplicates)
```

```
Number of duplicated rows: 0
```

Figure 8: Check for duplicates

As seen in figure 7, there are no missing values in the dataset. To prevent compromised prediction results and maintain data integrity, the dataset was checked for duplicated values, and there was none found.

4.2.4 Data Transformation

The values of these columns (EDU_FATHER, EDU_MOTHER, OCC_FATHER, OCC_MOTHER) were inconsistent. To simplify the columns for further analysis, they were grouped into two categories: 'Yes' signified parents who were educated or had jobs, and 'No' signified parents without any education or jobs.

```
# group education and occupation into 2 categories (yes and no)

student['EDU_FATHER']=student['EDU_FATHER'].apply(lambda x: 'No' if x == '0' else 'Yes')
student['EDU_MOTHER']=student['EDU_MOTHER'].apply(lambda x: 'No' if x == '0' else 'Yes')
student['OCC_FATHER']=student['OCC_FATHER'].apply(lambda x: 'No' if x == '0' else 'Yes')
student['OCC_MOTHER']=student['OCC_MOTHER'].apply(lambda x: 'No' if x == '0' else 'Yes')


print("EDU_FATHER:", student['EDU_FATHER'].unique())
print("EDU_MOTHER:", student['EDU_MOTHER'].unique())
print("OCC_FATHER:", student['OCC_FATHER'].unique())
print("OCC_MOTHER:", student['OCC_MOTHER'].unique())
```

```
EDU_FATHER: ['Yes' 'No']
EDU_MOTHER: ['Yes' 'No']
OCC_FATHER: ['Yes' 'No']
OCC_MOTHER: ['Yes' 'No']
```

Figure 9: Data Transformation

4.2.5 Encoding Categorical Data

Many machine learning algorithms such as Linear regression require only numerical input, so label encoding was applied to convert categorical variables into a numerical format, ensuring uniformity of data types across the dataset.

```
# encode categorical variables
cols_encoded = ['GENDER','EDU_FATHER','EDU_MOTHER','OCC_FATHER','OCC_MOTHER','STRATUM','PEOPLE_HOUSE','INTERNET',
                'TV','COMPUTER','WASHING_MCH','MIC_OVEN','CAR','DVD','FRESH','PHONE','MOBILE','REVENUE','JOB','SCHOOL
                'SCHOOL_TYPE','ACADEMIC_PROGRAM','SISBEN']

label_encoder=LabelEncoder()

for col in cols_encoded:
    student[col + '_encoded']=label_encoder.fit_transform(student[col])
```

Figure 10: Encoding using LabelEncoder

4.2.6 Target Variable Selection

The dataset consists of 6 potential target variables, which is excessive. To select appropriate target variables, a correlation analysis was performed to evaluate the strength and direction of relationships between them.

```
# correlation analysis for target variables
subset=student[['CC_PRO','CR_PRO','FEP_PRO','ENG_PRO','QR_PRO','WC_PRO']]
corr_subset = subset.corr()

plt.figure (figsize =(10,8))
sns.set(font_scale = .8)
sns.heatmap(corr_subset,annot=True, cmap='coolwarm', fmt='.2f', linewidths=.5)
plt.title('Correlation Matrix', fontsize=16)
plt.xticks(rotation=45,ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()

### wc_pro and fep_pro are the least correlated with other target variables
```

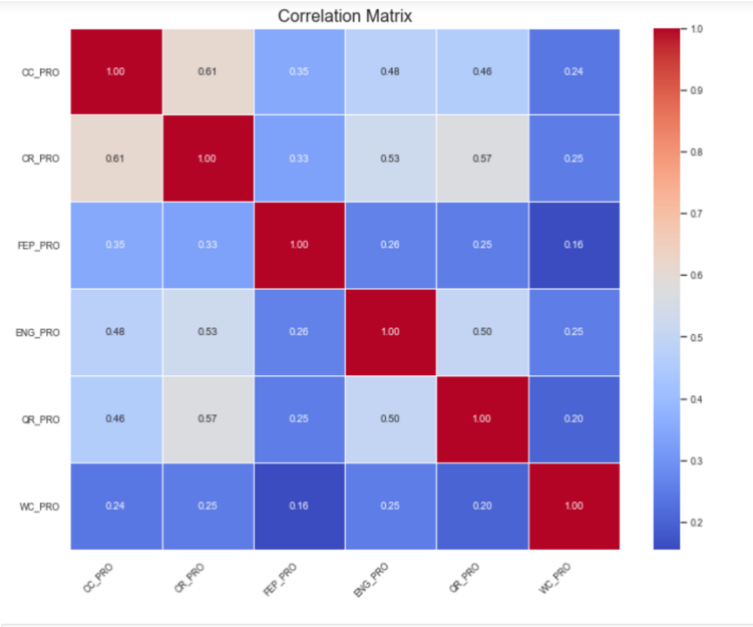


Figure 11: Correlation matrix

CR_PRO and CC_PRO were selected as target variables because they were highly correlated with each other and other variables (features). An additional correlation analysis was performed to determine which features had the strongest and weakest correlations with the target variables.

Features with Strong Correlation to the Target Variables

```
print ('Top 5 most correlated to target variable CR_PRO')
corr_matrix['CR_PRO'].sort_values(ascending=False).head(6)
```

```
Top 5 most correlated to target variable CR_PRO

CR_PRO        1.000000
PERCENTILE    0.791534
G_SC          0.786409
2ND_DECILE    0.761176
QUARTILE      0.745124
CC_PRO        0.607908
Name: CR_PRO, dtype: float64
```

```
print ('Top 5 most correlated to target variable CC_PRO')
corr_matrix['CC_PRO'].sort_values(ascending=False).head(6)
```

```
Top 5 most correlated to target variable CC_PRO

CC_PRO        1.000000
G_SC          0.757822
PERCENTILE    0.755583
2ND_DECILE    0.723866
QUARTILE      0.709715
CR_PRO        0.607908
Name: CC_PRO, dtype: float64
```

Figure 12: Correlation analysis

As seen in figure 12, the numerical variables in the dataset are highly correlated with the target variables, suggesting that a student's performance in one subject could explain their performance in another subject.

4.2.7 Data Splitting

The dataset was split into two subsets using the 'train-test-split' technique, where 70% of it was allocated for training the model and learning patterns in the data (x1_train, x2_train). The remaining 30% was used to test the model's performance and evaluate how well it generalises to unseen data (x1_test, x2_test). Due to the large size of the dataset, a 30% testing allocation is sufficient to provide a reliable estimate of the model's performance. Before splitting the dataset, the target variables were clearly defined as (y_cr, y_cc). As discussed previously, the target variables are all the final year engineering examinations indicated by the suffix 'PRO'. Therefore, these variables were dropped from the features list.

```
# Define Features

x_cr = student.drop(columns=['CR_PRO', 'CC_PRO', 'QR_PRO', 'ENG_PRO','FEP_PRO', 'WC_PRO'])
x_cc = student.drop(columns=['CC_PRO', 'CR_PRO', 'QR_PRO', 'ENG_PRO','FEP_PRO', 'WC_PRO'])


# Define Target Variables

y_cr = student['CR_PRO']
y_cc = student['CC_PRO']


# Data Splitting

x1_train,x1_test,x2_train,x2_test,y1_train,y1_test,y2_train, y2_test = train_test_split
(x_cr,x_cc, y_cr,y_cc,test_size =0.3, random_state=42)
```

Figure 13: Feature selection & data split

4.2.8 Feature Scaling

The target variables in the dataset have a large scale ranging from 1 – 100, while gender has been encoded to a smaller scale between 1 – 0. This means that the target variables could have a dominant influence on the model's predictions, therefore feature scaling was applied to ensure all features contribute equally to the model's learning process (Scikit-Learn, 2023). Similarly, distance-based machine learning models such as K-NN and Support Vector Machines are sensitive to the scales of variables and require feature scaling to ensure uniformity across all features (Scikit-Learn, 2023). 'StandardScaler' is widely used by other researchers to perform feature scaling, and it was employed in this report due to its efficiency.

```
# Feature Scaling
scaler = StandardScaler()
x1_train_scaled = scaler.fit_transform(x1_train)
x1_test_scaled = scaler.transform(x1_test)

x2_train_scaled = scaler.fit_transform(x2_train)
x2_test_scaled = scaler.transform(x2_test)

x3_train_scaled = scaler.fit_transform(x3_train)
x3_test_scaled = scaler.transform(x3_test)
```

Figure 14: Feature scaling

4.2.9 Feature Selection

The dataset has 44 features, and according to (Cardonaa & Cudneya, 2019), models with fewer features are less prone to overfitting issues and poor generalisations. As a result, a feature selection technique called SelectKBest was employed to remove redundant features and retain the optimal features impacting students' academic performance.

```
# feature selection

k_best = SelectKBest(score_func=f_regression, k='all')

x1_train_sel = k_best.fit_transform(x1_train_scaled, y1_train)
x1_test_sel = k_best.transform(x1_test_scaled)

x2_train_sel = k_best.fit_transform(x2_train_scaled, y2_train)
x2_test_sel = k_best.transform(x2_test_scaled)

x3_train_sel = k_best.fit_transform(x3_train_scaled, y3_train)
x3_test_sel = k_best.transform(x3_test_scaled)
```

Figure 15: Feature selection

4.3 Model Training and Evaluation

Machine learning techniques are employed in this report because they offer several benefits over traditional statistical methods. These benefits include flexibility, allowing the machine learning algorithms to learn complex and non-linear patterns between target variables and independent features without imposing strict assumptions (Jin, 2023). Machine learning algorithms also have built-in feature selection functions which are capable of identifying the most relevant variables in a predictive model. Another benefit is their ability to leverage validation techniques such as cross-validation to ensure that findings are authentic and resistant to overfitting issues (Jin, 2023). Finally, machine learning models are robust to multicollinearity which occurs when independent variables are highly correlated with one another. Multicollinearity leads to biased prediction results and error metrics.

The following machine learning algorithms were chosen to predict the academic performance of students in the dataset. The rationale for selecting these algorithms is discussed in great detail in chapter 3.

- Linear Regression (LR)
- Decision Tree Regressor (DT)
- Random Forest (RF)
- Support Vector Regressor (SVR)
- K Nearest Neighbors Regressor (KNN)

A custom function was built to train the selected regression models and evaluate their prediction performance using specified regression metrics (Mean Squared Error, Mean Absolute Error, Root Mean Squared Error). Each model is fitted to the training data (x1_train_sel, y1_train) to learn patterns, then it makes predictions on the unseen test data (x1_test_sel) and calculates the regression metrics for easier evaluation.

4.3.1 Target 1 - CR_PRO (Critical Reading)

```
models = [LinearRegression(),
          DecisionTreeRegressor(),
           RandomForestRegressor(),
           SVR(),
           KNeighborsRegressor(),
           GradientBoostingRegressor()
          ]

for model in models:
    model.fit(x1_train_sel, y1_train)
    y1_pred = model.predict(x1_test_sel)
    mse = mean_squared_error (y1_test,y1_pred)
    mae = mean_absolute_error (y1_test, y1_pred)
    r2 = r2_score (y1_test, y1_pred)
    rmse = np.sqrt(mse)


    submit = pd.DataFrame()
    submit['Actual CR_PRO Grades'] = y1_test
    submit['Predicted CR_PRO Grades'] = y1_pred
    submit = submit.reset_index()

    print(f"\nModel:{type(model).__name__}")
    print(f"MSE: {mse:.2f}, MAE: {mae:.2f}, R-squared: {r2:.2f}, RMSE: {rmse:.2f}")
    print (submit.head(5))
```

```
Model:LinearRegression
MSE: 192.92, MAE: 10.01, R-squared: 0.75, RMSE: 13.89
   index  Actual CR_PRO Grades  Predicted CR_PRO Grades
0   9985                    97                91.682868
1   7476                    88                81.097469
2   6867                    84                77.157233
3   5350                    85                77.405530
4   9669                    38                35.431339

Model:DecisionTreeRegressor
MSE: 278.16, MAE: 12.46, R-squared: 0.64, RMSE: 16.68
   index  Actual CR_PRO Grades  Predicted CR_PRO Grades
0   9985                    97                      98.0
1   7476                    88                      82.0
2   6867                    84                      96.0
3   5350                    85                      95.0
4   9669                    38                       4.0

Model:RandomForestRegressor
MSE: 128.61, MAE: 8.60, R-squared: 0.83, RMSE: 11.34
   index  Actual CR_PRO Grades  Predicted CR_PRO Grades
0   9985                    97                    93.53
1   7476                    88                    84.69
2   6867                    84                    87.60
3   5350                    85                    84.95
4   9669                    38                    19.62

Model:SVR
MSE: 265.84, MAE: 12.36, R-squared: 0.65, RMSE: 16.30
   index  Actual CR_PRO Grades  Predicted CR_PRO Grades
0   9985                    97                94.744878
1   7476                    88                73.907171
2   6867                    84                85.399906
3   5350                    85                83.713604
4   9669                    38                17.861325

Model:KNeighborsRegressor
MSE: 350.38, MAE: 14.61, R-squared: 0.54, RMSE: 18.72
   index  Actual CR_PRO Grades  Predicted CR_PRO Grades
0   9985                    97                      94.6
1   7476                    88                      72.0
2   6867                    84                      88.0
3   5350                    85                      75.8
4   9669                    38                      16.2
```

Figure 16: Training models for CR_PRO

Comparative Analysis

Several scatterplots were plotted to visually compare the predictive performance of the trained machine learning models. The charts below also show that Random Forest model appears to produce the most accurate predictions, as majority of the predicted values fit closely around the line of perfect prediction.

```
# Train Random Forest for CR_PRO
rf_model = RandomForestRegressor()
rf_model.fit(x1_train_sel, y1_train)
y1_pred_rf = rf_model.predict(x1_test_sel)

# Plot actual vs predicted values for Random Forest model
plt.figure(figsize=(10, 6))
plt.scatter(y1_test, y1_pred_rf, color='blue', label='Actual vs Predicted (Random Forest)')
plt.plot([min(y1_test), max(y1_test)], [min(y1_test), max(y1_test)], color='red', linestyle='--', label='Perfect Pre
plt.xlabel('Actual CR_PRO')
plt.ylabel('Predicted CR_PRO')
plt.title('Actual vs Predicted Values for CR_PRO')
plt.legend()
plt.grid(True)
plt.show()
```

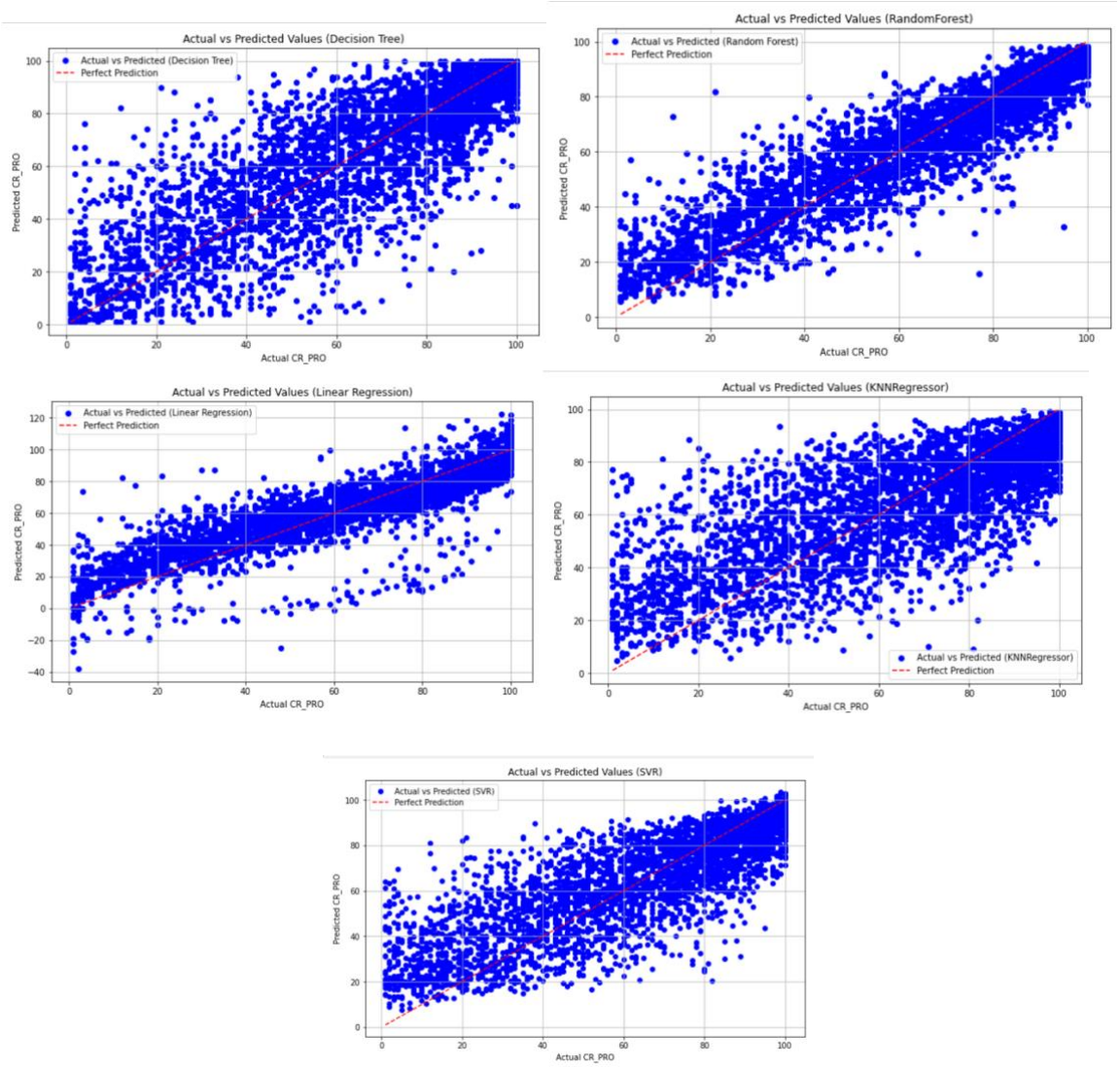


Figure 17: Comparison of models' performance for CR_PRO

4.3.2 Target 2 – CC_PRO (Citizenship Competencies)

```
models = [LinearRegression(),
          DecisionTreeRegressor(),
           RandomForestRegressor(),
           SVR(),
           KNeighborsRegressor(),
           GradientBoostingRegressor()
          ]

for model in models:
    model.fit(x2_train_sel, y2_train)
    y2_pred = model.predict(x2_test_sel)
    mse = mean_squared_error (y2_test,y2_pred)
    mae = mean_absolute_error (y2_test, y2_pred)
    r2 = r2_score (y2_test, y2_pred)
    rmse = np.sqrt(mse)
    max_err =max_error(y2_test, y2_pred)

    submit = pd.DataFrame()
    submit['Actual CC_PRO Grades'] = y2_test
    submit['Predicted CC_PRO Grades'] = y2_pred
    submit = submit.reset_index()

    print(f"\nModel:{type(model).__name__}")
    print(f"MSE: {mse:.2f}, MAE: {mae:.2f}, R-squared: {r2:.2f}, RMSE: {rmse:.2f}")
    print (submit.head(5))
```

```
Model:LinearRegression
MSE: 204.59, MAE: 10.21, R-squared: 0.76, RMSE: 14.30
   index  Actual CC_PRO Grades  Predicted CC_PRO Grades
0   9985                    75                79.436617
1   7476                    87                78.242262
2   6867                    95                82.080715
3   5350                    70                68.496528
4   9669                    20                30.935922

Model:DecisionTreeRegressor
MSE: 310.53, MAE: 12.91, R-squared: 0.63, RMSE: 17.62
   index  Actual CC_PRO Grades  Predicted CC_PRO Grades
0   9985                    75                      85.0
1   7476                    87                      67.0
2   6867                    95                      94.0
3   5350                    70                      81.0
4   9669                    20                      18.0

Model:RandomForestRegressor
MSE: 131.73, MAE: 8.75, R-squared: 0.84, RMSE: 11.48
   index  Actual CC_PRO Grades  Predicted CC_PRO Grades
0   9985                    75                    88.26
1   7476                    87                    79.94
2   6867                    95                    92.90
3   5350                    70                    77.53
4   9669                    20                    21.60

Model:SVR
MSE: 304.40, MAE: 13.03, R-squared: 0.64, RMSE: 17.45
   index  Actual CC_PRO Grades  Predicted CC_PRO Grades
0   9985                    75                90.656014
1   7476                    87                69.194677
2   6867                    95                88.008648
3   5350                    70                81.335429
4   9669                    20                18.999462

Model:KNeighborsRegressor
MSE: 430.98, MAE: 16.15, R-squared: 0.49, RMSE: 20.76
   index  Actual CC_PRO Grades  Predicted CC_PRO Grades
0   9985                    75                      94.8
1   7476                    87                      60.8
2   6867                    95                      71.0
3   5350                    70                      78.0
4   9669                    20                      16.4
```

Figure 18: Training models for CC_PRO

Comparative Analysis

Several scatterplots were plotted to visually compare the predictive performance of the trained machine learning models. The charts below also show that Random Forest model appears to produce the most accurate predictions, as majority of the predicted values fit closely around the line of perfect prediction.

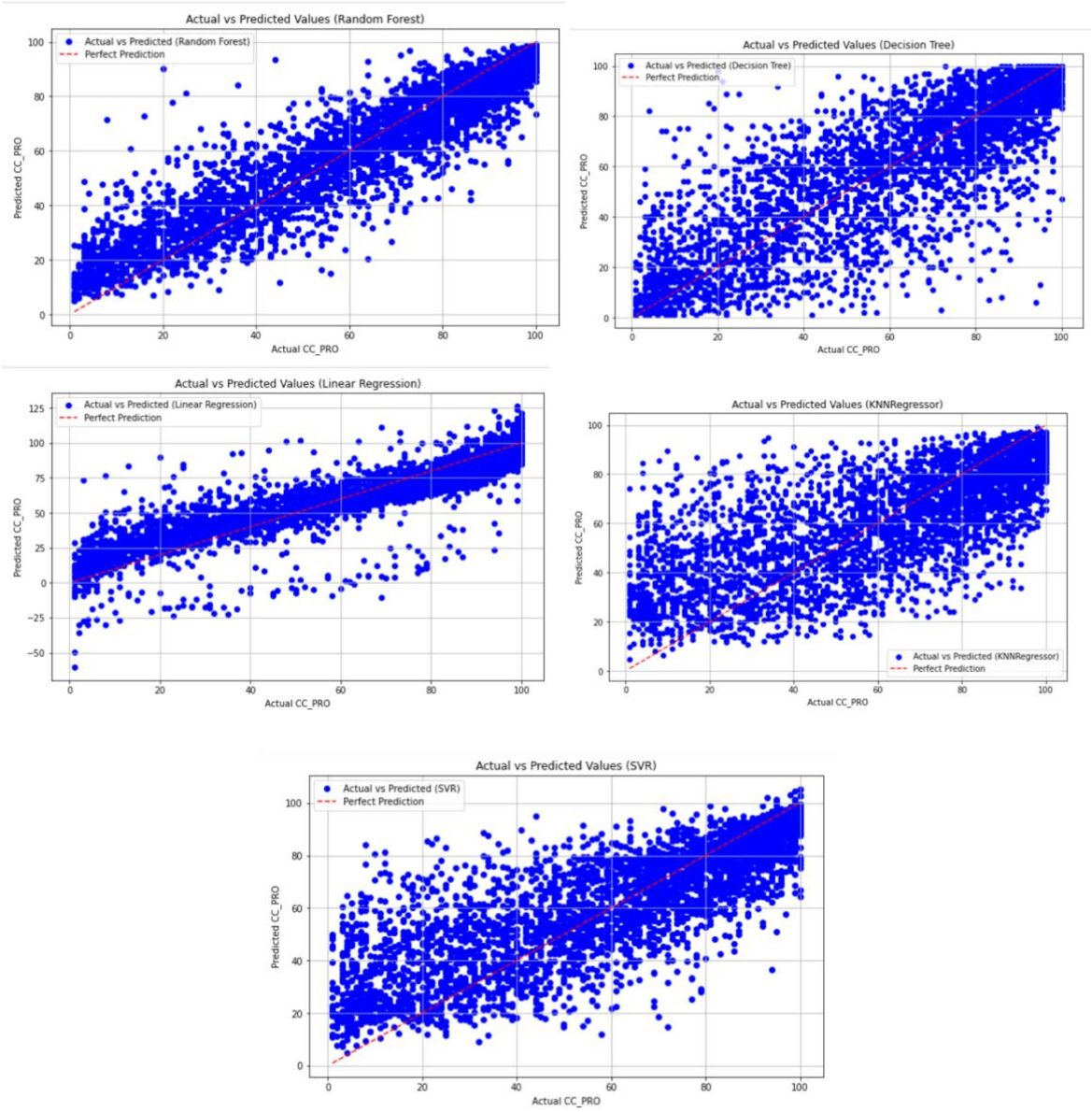


Figure 19: Comparison of models' performance for CC_PRO

4.4 Feature Importance Analysis

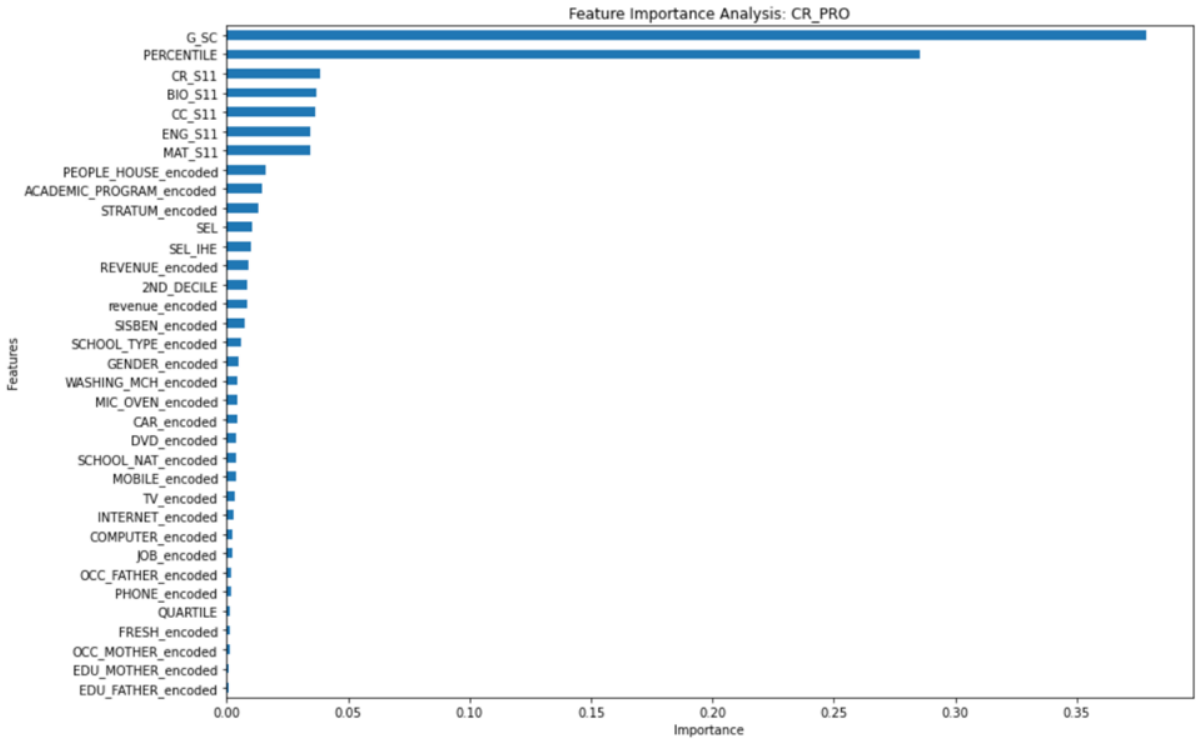
The choice of input features is as important as the choice of machine learning algorithms. The Random Forest model is the most accurate in terms of r-squared value and error metrics, therefore feature importance analysis was conducted to identify the relative influence of the input variables in predicting the target variables (CR_PRO, CC_PRO). After training the Random Forest model, the feature importance scores are extracted using the 'feature_importances_' function of the trained model. A high feature importance score implies the variable is influential in contributing to the prediction of the target variable, while a low score implies it has less impact on the model's predictive performance (AlSagri & Ykhlef, 2020).

```
# view the feature importance scores : CR_PRO

feature_scores = pd.Series(rf_model.feature_importances_, index=x1_train.columns).sort_values(ascending=False)

feature_scores
```

```
# Plot the bar chart
plt.figure(figsize=(14, 10))
feature_scores.sort_values().plot(kind='barh')  # Using 'barh' for horizontal bar chart
plt.title('Feature Importance Analysis: CR_PRO')
plt.xlabel('Importance')
plt.ylabel('Features')
plt.show()
```



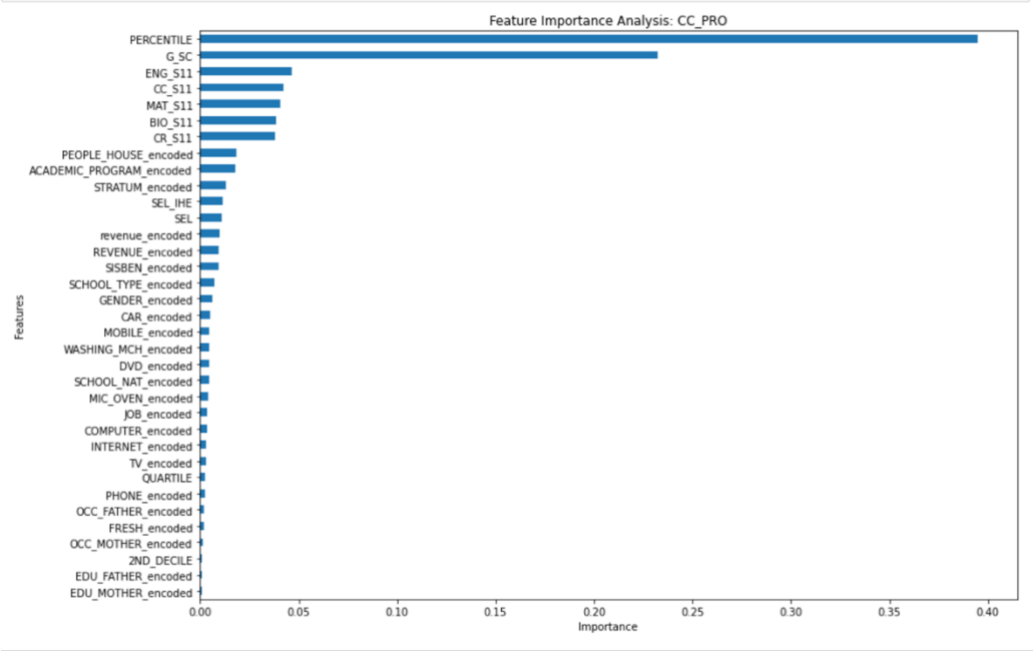


Figure 20: Feature importance analysis

In summary, this chapter discussed the steps taken by the author to predict students' academic performance. It began by describing the dataset and explaining the various pre-processing techniques which were applied in the report. Then it provided justifications for the chosen machine learning algorithms and discussed the training process of the learning models. It also evaluated the models' prediction performance using selected evaluation metrics.

5 Results and Discussion

In this chapter, the results of the machine learning predictions will be analysed and evaluated in line with the research questions of this report. This report followed a two-fold analysis where the final year grades of students in their professional engineering program were predicted using five machine learning models. Following the successful prediction of students' academic performance, a feature importance analysis was conducted to identify which features had the most influence on the predicting the target variables (CR_PRO, CC_PRO).

RQ1: Can machine learning models accurately predict students' academic performance?

This report applied five machine learning models to predict students' academic grades: Random Forest, Linear Regression, K-Nearest Neigbours Regression, Support Vector Regression and Decision Tree. Based on the objectives of this report, two target variables were predicted and their results are given below.

Table 4: Summary of prediction results: CR_PRO

Model	R-Square	MSE	MAE	RMSE
Random Forest	0.83	128.61	8.60	11.34
Linear Regression	0.75	192.92	10.01	13.89
Decision Tree	0.64	278.16	12.46	16.68
KNN Regression	0.54	350.38	14.61	18.72
SVR	0.65	265.84	12.36	16.30

Random Forest outperforms other models in predicting CR_PRO and CC_PRO. An R-square of 0.83 indicates the model explains 83% of the variance in CR_PRO. Compared to other models, Random Forest also has the lowest error metrics indicating that the model's predicted values are closer to the actual values.

Linear regression shows moderate performance with an R-square value of 0.75 indicating that the model explains 75% of the variance in Critical Reading grade (CR_PRO). The MAE and MSE are significantly low suggesting that the model's predicted values are close to the actual values.

Decision tree achieves a moderate prediction performance with an R-square value of 0.64, indicating the model explains 64% of the variance in CR_PRO.

KNN Regression performs poorly in predicting students' academic performance with its high error metrics and low R-square value of 0.49. This implies the KNN model only explains 49% of the variance in the target variable. It could benefit from additional pre-processing or parameter tuning.

SVR demonstrates moderate prediction performance with a R-square value of 0.65. However, its high error metrics suggest there is room for improvement.

Table 5: Summary of prediction results: CC_PRO

Model	R-Square	MSE	MAE	RMSE
Random Forest	0.84	131.73	8.75	11.48
Linear Regression	0.76	204.59	10.21	14.30
Decision Tree	0.63	310.53	12.91	17.62
KNN Regression	0.49	430.98	16.15	20.76
SVR	0.64	304.40	13.03	17.45

Random Forest outperforms other models again with the lowest error metrics and highest R-square value of 0.84. The high R-squared value demonstrates that Random Forest is best suited to model complex relationships within the dataset and produce accurate predictions.

Linear Regression shows moderate performance with an r-square of 0.76 indicating that 76% of the variance in CC_PRO is explained by the model. Its low error metrics also suggest that linear regression is able to make reasonable accurate predictions on the target variable.

Decision tree did not perform as well as linear regression due to its lower R-square value (0.63) and higher error metrics. This suggests the model may not generalise well to new data, given the large size of the dataset.

SVR's performance is similar to decision tree with a moderate R-square value (0.64) and larger errors, suggesting that its predictions are less accurate and unsuitable for this dataset.

KNN regression demonstrates the lowest prediction performance among other models with an r-square value of 0.49 and higher error metrics.

The main aim of this report is to develop a robust machine learning model capable of accurately predicting students' academic performance. After training multiple models with both numeric and non-numeric variables, Random Forest emerged as the most accurate model with a prediction accuracy of 83% and 84% for CR_PRO and CC_PRO respectively. Random Forest model also reported the lowest error metrics (MSE, MAE, RMSE) indicating the predicted values are closer to the actual values when compared with other models. Thus, we can conclude that Random Forest model is effective and capable of accurately predicting students' academic performance.

RQ2: What academic and behavioural factors significantly influence students' academic performance?

The results of the feature importance analysis revealed valuable insights about the influence of academic and behavioural factors on students' academic performance. Specifically, this analysis focused on the Random Forest results because it emerged as the most accurate model with a prediction accuracy of 83% and 84% for both target variables. The results revealed past grades as the top predictors of CR_PRO and CC_PRO. The data on past grades was collected at the end a student's high school education and indicated by S_11. Conversely, the same students were assessed after five years during their final year professional engineering exams. These grades are denoted by S_PRO in the dataset. For both target variables, the analysis results show CR_S11, CC_S11, BIO_S11, ENG_S11 and MAT_S11 as the top variables contributing to the prediction of students' academic performance. Thus, the identification of past academic grades as the top predictors of future academic performance emphasises the importance of early prediction of at-risk students to implement intervention methods and prevent failure. The analysis also revealed socio-economic and demographic variables such as 'gender', 'people_house', 'stratum', 'SEL', 'revenue' as influential predictors of students' final grades. Where stratum and SEL refer to income level and socio-economic level respectively. The results further highlight the influential role of socio-economic factors in determining students' academic performance.

The results also revealed the impact of household amenities such as 'washing machine', 'tv', 'mobile' and 'DVD' in predicting student outcomes. This highlights how students' study habits at home could influence their performance in school. Similarly,

the identification of 'academic_program' reveals that the design of a course curriculum or teaching methods could influence students' academic performance. Surprisingly, the results of the feature importance analysis revealed that the education and occupation of parents had very limited influence on students' academic performance.

RQ3: What are the policy implications of predicting students' academic performance?

Predicting student performance is important for policy implementations and social development. This report provides concrete evidence that past performance and socio-economic factors are key predictors for future students' performance. Therefore, the analysis and prediction results in this report can provide guidelines for implementing timely interventions for students at risk of failure. Upon the successful prediction of students' academic performance, early intervention systems such as personalised lecture materials, additional teaching sessions and family support programs could be provided to students to improve their chances of success in their respective academic programs. Several studies such as (Arnold & Pistilli, 2012) highlight the success of 'Course Signals' in the early identification of underperforming students at Purdue University and the subsequent improvement in their academic performance. The study also revealed an increase in student retention following the implementation of 'Course Signals' in the university. Similarly, the feature importance analysis conducted in this report revealed 'Academic Program' as one of the top predictive features of students' academic performance. Therefore, these results could guide policy administrators and educators in allocating valuable resources to academic departments with the most underperforming students.

5.1 Limitations

Acknowledging the potential limitations of this report's findings is critical for future development. A major limitation of this report is that predictive models generally establish correlation between target and independent variables, however, they do not accurately establish the underlying causal influence. As a result, the insights in this report should be interpreted with caution and supplemented with additional research methods to show causal relationships.

Another limitation is the inability of predictive models to forecast outcomes beyond the range of data used to train them, resulting to erroneous extrapolations. Therefore, extra caution must be exercised when applying this model's predictions to a different student population where underlying relationships may be different. Thus, this report could benefit from collecting data from multiple sources, to provide a comprehensive conclusion of the predicted target variables.

Finally, this report originally intended to utilise real-world data from the University of Northumbria students' database, but after a careful consideration of data privacy and ethics concerns, a public dataset was recommended to mitigate privacy risks. As a result, this report's findings may not be directly applicable to the unique characteristics and environment of the University of Northumbria.

5.2 Future Development

This report contributes to the education sector through the early prediction of underperforming students along with the identification of the significant features that enable a student's optimal academic performance. Overall, the results highlight the effectiveness of machine learning techniques in evaluating the early prediction of students' academic performance. As a result, the findings in this report may assist education providers in the formulation of predictive learning analytics framework for better decision-making.

In the future, the author intends to explore advanced machine learning algorithms such as Neural Networks to improve the prediction results and model the complex relationships between features and academic performance. Similarly, the current dataset will be supplemented with additional data sources including student

engagement metrics and student information from learning management systems to provide a complete understanding of student performance and behaviour. Finally, early intervention systems and other personalised strategies should be developed to support underperforming students in a timely manner. Such strategies are built on the foundation of predictive analytics, and they dynamically provide targeted resources, feedback and support, based on students' individual needs and preferences.

By pursuing these further developments, this work will contribute to the advancement of educational data mining and ultimately assist education providers in implementing data-driven strategies to support students' academic performance.

This chapter discussed the steps taken by the author to predict students' academic performance. It began by describing the dataset and explaining the various pre-processing techniques which were applied in the report. Then it provided justifications for the chosen machine learning algorithms and discussed the training process of the learning models. It also evaluated the models' prediction performance using selected evaluation metrics and presented the limitations of the report.

6 Conclusion

This report presents the results of the prediction of students' academic performance using machine learning techniques, as well as the identification of significant factors that influence the prediction of students' academic performance. This report differs from previous work by analysing an extensive dataset beyond academic records and developing a robust predictive model that considers the dynamic and unique characteristics of each student's learning experience. Specifically, this report considers a case of 12,411 students studying towards their professional engineering exams in Colombia. It develops a prediction model that operates from the final semester of secondary school to enable early identification of underperforming students in the final year of their professional engineering program. The early prediction of students' grades may help educators to identify weak students, customise the teaching plans to the students' needs and support them in achieving academic success. Timely predictions may also guide the university administrators to efficiently allocate resources among departments in need of prioritisation.

The contribution of this report is three-fold. First, it utilises a larger dataset to provide a more comprehensive representation of the student population. With 44 variables, this report considers a wide range of student characteristics including demographic information and socio-economic backgrounds, making it adaptable across different education settings. Finally, six different machine learning models are employed to assess the performance and consistency of predictions across different target variables. Overall, this report highlights the effectiveness of the Random Forest model in the prediction of students' grades, development of data-driven policies and consequently aiding the institution in maintaining their academic career.

Throughout the course of this project, the author has gained relevant knowledge and experiences that have contributed to both their personal and professional growth as a research student. Upon reflecting on this journey, some notable lessons and challenges stand out. One of the key challenges encountered by the author was accessing high-quality students' data that reflected the unique experiences of higher education students in the UK. This project originally intended to utilise real-world data from the University of Northumbria students' database, but after a careful consideration of data

privacy and ethics concerns, my request was denied in order to mitigate privacy risks. I overcame this challenge by exploring alternative data sources such as Mendeley's public data repository which had a large collection of high-quality labelled data. Another challenge I encountered was managing my time effectively to complete the project within the required timeline and meet the demands of my part-time job simultaneously. To address this challenge, I utilised strategic time management tools such as personal development plans and Gaant charts to ensure that each stage of the dissertation received an adequate amount of attention and resources.

In overcoming these challenges with the support of my supervisor, I gained a deeper understanding of machine learning techniques and the nuances involved in applying them to education data. Each setback motivated me to explore deeply into the field of machine learning and refine my problem-solving skills. More specifically, I gained valuable insights about developing predictive models along with the significant factors that influence a student's academic performance. As I continue to explore data science more deeply, I am inspired by the endless possibilities of the applications of machine learning in education and other sectors. The knowledge and lessons learned during this project will assist me to develop early warning intervention dashboards and data-driven insights capable of shaping the future of education. I also intend to apply this newly acquired knowledge in upcoming personal projects such as the prediction of possible loan applicants who are at risk of defaulting on their loans.

7 Bibliography

Agrawal, S., Vishwakarma, S. K. & Sharma, A. K., 2017. Using data mining classifier for predicting student's performance in UG level. *International Journal of Computer Applications,* 172(8), pp. 39-44.

Ahmed, A. & Elaraby, I., 2014. Data mining: a prediction for student's performance using classification method. *World Journal of Computer Application and Technology,* 2(2), pp. 43-47.

Alexandropoulos, S.-A. N., Kotsiantis, S. B. & Vrahatis, M. N., 2019. Data preprocessing in predictive data mining. *The Knowledge Engineering Review,* 34(1), pp. 1-33.

Alghamdi, A. S. & Rahman, A., 2023. Data mining approach to predict success of secondary school stu- dents: A saudi arabian case study. *Education Sciences,* 13(3).

AlSagri, H. & Ykhlef, M., 2020. Quantifying Feature Importance for Detecting Depression using Random Forest. *International Journal of Advanced Computer Science and Applications,* 11(5), pp. 628-635.

Al-Shabandar, R. et al., 2017. *Machine learning approaches to predict learning outcomes in Massive open online courses.* s.l., International Joint Conference on Neural Networks (IJCNN).

Alturki, S. & Alturki, N., 2021. Using educational data mining to predict students' academic perfor- mance for applying early interventions. *Journal of Information Technology Education: Innovations in Practice,,* Volume 20, pp. 121-137.

Analytics Vidhya, 2023. *Random Forest vs Decision Tree.* [Online] Available at: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ [Accessed 15 January 2024].

Anderson, H., Boodhwani, A. & Baker, R., 2019. *Predicting graduation at a public R1 university.* s.l., 9th International Learning Analytics and Knowledge Conference.

Arnold, K. E. & Pistilli, M. D., 2012. *Course Signals at Purdue: Using Learning Analytics to Increase Student Success.* Purdue, 2nd International Conference: Learning Analytics.

Badugu, S. & Rachakatla, B., 2020. Student's performance prediction using machine learning approach.. In: *Data Engineering and Communication Technology.* Singapore: Springer, pp. 333-240.

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E. & Karadeniz, A., 2020. An Early Warning System to Detect At-Risk Students in Online Higher Education. *Applied Sciences,* 10(3), pp. 249-263.

B

eaulac, C. & Rosenthal, J. S., 2019. Predicting university students' academic success and major using random forests.. *Research in Higher Education,* 60(7), pp. 1048-1064.

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P. & Moniaga, J. V., 2023. Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science,* Volume 216, p. 597–603.

Belachew, E. B. & Gobena, F. A., 2017. Student Performance Prediction Model using Machine Learning Approach: The case of Wolkite University. *International Journal of Advanced Research in Computer Science and Software Engineering,* 7(2), pp. 46-50.

Bydžovská, H., 2016. *A Comparative Analysis of Techniques for Predicting Student Performance.* Raleigh, North Carolina, International Conference on Educational Data Mining (EDM), pp. 306-311.

Cardonaa, T. A. & Cudneya, E. a., 2019. Predicting Student Retention Using Support Vector Machines. *Procedia Manufacturin,* Volume 39, pp. 1827-1833.

Chen, R. & DesJardins, S. L., 2010. Investigating the Impact of Financial Aid on Student Dropout Risks: Racial and Ethnic Differences. *The Journal of Higher Education ,* Volume 81, pp. 179-208 .

Chen, Y., Miao, D. & Zhang, H., 2010. Neighborhood outlier detection. *Expert Systems with Applications,* 37(12), p. 8745–8749.

Chen, Y. & Zhai, L., 2023. A comparative study on student performance prediction using machine learning. *Education and Information Technologies,* Volume 28, p. 12039–12057.

Costa, E. B. et al., 2017. Evaluating the effec- tiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior,* Volume 73, pp. 247-256.

Datacamp, 2023. *Decision Tree Classification in Python Tutorial.* [Online] Available at: https://www.datacamp.com/tutorial/decision-tree-classification-python [Accessed 15 January 2024].

Datatron, 2023. *What is a Support Vector Machine?.* [Online] Available at: https://datatron.com/what-is-a-support-vector-machine/ [Accessed 15 January 2024].

Daud, A. et al., 2017. *Predicting student performance using advanced learning analytics.* Geneva, Switzerland:, Proceedings of the 26th international conference on world wide web (WWW '17).

Delen, D., 2010. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems,* 49(4), p. 498–506.

Edgar, T. W. & Manz, D. O., 2017. Experimental research designs. In: *Research Methods for Cyber Security.* Cambridge: Elsevier Science, pp. 215-249.

Huang, H., Lin, J., Chen, C. & Fan, M., 2006. Review of outlier detection. *Application Research of Computers,* Volume 8, pp. 2006-2008.

Hussain, A., Khan, M. & Ullah, K., 2022. Student's performance prediction model and affecting factors using classifcation techniques. *Education and Information Technologies,* 27(6), p. 8841–8858.

Hussain, M., Zhu, W. & Zhang, W., 2019. Using machine learning to predict student difficulties from learning session data.. *Artificial Intelligence Review,* Volume 52, pp. 381-407.

Hussain, S. et al., 2019. Prediction Model on Student Performance based on Internal Assessment using Deep Learning. *International Journal of Emerging Technologies in Learning,* 14(8), pp. 4-22.

Huynh-Cam, T.-T., Chen, L.-S. & Le, H., 2021. Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms,* 14(11), p. 318.

Jayaprakash, S., E, B. & Chandar, V., 2018. *Predicting Students Academic Performanceusing Naive Bayes Algorithm,* Ghana: Oxford.

Jayaprakash, S. M. et al., 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics,* 1(1), pp. 6-47.

Jin, X., 2023. Predicting academic success: machine learning analysis of student, parental, and school efforts. *Asia Pacifc Education Review,* pp. 1-22.

Kamala, R. & Thangaiah, R., 2019. An improved hybrid feature selection method for huge dimensional datasets. *IAES International Journal of Artificial Intelligence (IJ-AI),* 8(1), pp. 77-86.

Kantardzic, M., 2019. Preparing the Data. In: *Data Mining.* s.l.:Wiley, pp. 45-46.

Keshtkar, F., 2018. *Predicting Risk of Failure in Online Learning Platforms Using Machine Learning Algorithms for Modeling Students' Academic Performance,* Missouri: Missouri State University.

Khobragade, L. & Mahadik, P., 2015. Students' academic failure prediction using data mining. *International Journal of Advanced Research in Computer and Communication Engineering,* 4(11), pp. 290-298.

Kiss, B., Nagy, M., Molontay, R. & Csabay, B., 2019. *Predicting dropout using high school and first-semester academic achievement measures.* Starý Smokovec, Slovakia, 17th International Conference on Emerging eLearning Technologies and Applications (ICETA).

Lau, E. T., Sun, L. & Yang, Q., 2019. Modelling, prediction and classification of studentacademic performance using artificial neural networks. *Applied Sciences,* 1(9), pp. 982-992.

Malik, S. & Jothimani, K., 2023. Enhancing Student Success Prediction with FeatureX: A Fusion Voting Classifier Algorithm with Hybrid Feature Selection. *Education and Information Technologies,* Volume 39, pp. 1-51.

Malini, J. & Kalpana, Y., 2021. Investigation of factors affecting student performance evaluation using education materials data mining technique. *Materials Today: Proceedings,* Volume 47, pp. 6105-6110.

Marquez-Vera, C., Morales, C. R. & Soto, S. V., 2013. Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologias del Aprendizaje,* Volume 8, pp. 7-14.

Martins, M., Baptista, L., Machado, J. & Realinho, V., 2023. Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education. *Applied Sciences,* 13(8), p. 4702.

Mastour, H., Dehghani, T., Moradi, E. & Eslami, S., 2023. Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. *Heliyon,* Volume 9, pp. 1-17.

Matzavela, V. & Alepis, E., 2021. Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. *Computers and Education: Artificial Intelligence,* Volume 2, pp. 1-12.

Matz, S. et al., 2023. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics,. *Scientific Reports,* 13(5705).

Mendoza-Mendoza, A., Hoz-Domínguez, E. D. L. & Visbal-Cadavid, D., 2023. Classification of industrial engineering programs in Colombia based on state tests. *Heliyon,* Volume 9, pp. 1-12.

Nachouki, M., Mohamed, E. A., Mehdi, R. & Naaj, M. A., 2023. Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education ,* Volume 33, pp. 1-7.

Naseer, M., Zhang, W. & Zhu, W., 2020. Early prediction of a team performance in the initial assessment phases of a software project for sustainable software engineering education.. *Sustainability,* 12(11).

Nawai, S., Saharan, S. & Hamzah, N., 2021. *An analysis of students' performance using CART approach.* s.l., AIP Publishing LLC.

Obsie, E. Y. & Adem, S. A., 2018. Prediction of Student Academic Performance using NeuralNetwork, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications,* 180(40), pp. 39-47.

Ofori, F., Maina, E. & Gitonga, R., 2020. Using machine learning algorithms to predict students performance and improve learning outcome: a literature based review. *Journal of Information Technology,* 4(1), pp. 33-55.

Olabanjo, O. A., Wusu, A. S. & Manuel, M., 2022. A machine learning prediction of academic performance of secondary school students using radial basis function neural network. *Trends in Neuroscience and Education,* Volume 29, pp. 100-190.

Oloruntoba, S. A. & Akinode, J. L., 2017. Student academic performance prediction using support vector machine. *International Journal of Engineering Sciences and Research Technology,* 6(12), pp. 588-597.

Rajalaxmi, R., Natesan, P., Krishnamoorthy, N. & Ponni, S., 2019. Regression Model for Predicting Engineering Students Academic Performance. *International Journal of RecentTechnologyand Engineering,* 7(653), pp. 71-75.

Rashid, T. A. & Aziz, N. K., 2016. Student Academic Performance Using Artificial Intelligence. *Journal of Pure and Applied Sciences,* 28(2), pp. 56-69.

Sánchez, C. J. P., Calle-Alonso, F. & Vega-Rodríguez, M. A., 2022. Learning analytics to predict students' performance: A case study of a neurodidactics-based collaborative learning platform. *Education and Information Technologies ,* Volume 27, p. 12913–12938.

Said, M. B., Kacem, Y. H., Algarni, A. & Masmoudi, A., 2023. Early prediction of Student academic performance based on Machine Learning algorithms: A case study of bachelor's degree students in KSA. *Education and Information Technologies,* pp. 1-24.

Sarra, A., Fontanella, L. & Zio, S. D., 2019. Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. *Social Indicators Research,* 146(1), pp. 41-60.

Sassirekha, M. S. & Vijayalakshmi, S., 2022. Predicting the academic progression in student's standpoint using machine learning. *Automatika,* 63(4), pp. 605-617.

Scikit Learn, 2023. *Linear Models.* [Online] Available at: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

Scikit Learn, 2023. *Linear Models.* [Online] Available at: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

Scikit Learn, 2024. *Ensembles: Gradient boosting, random forests, bagging, voting, stacking.* [Online] Available at: https://scikit-learn.org/stable/modules/ensemble.html#random-forests-and-other-randomized-tree-ensembles

Scikit Learn, 2024. *Metrics and scoring: quantifying the quality of predictions.* [Online] Available at: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

Scikit Learn, 2024. *Nearest Neighbors.* [Online] Available at: https://scikit-learn.org/stable/modules/neighbors.html

Scikit Learn, 2024. *Support Vector Machines.* [Online] Available at: https://scikit-learn.org/stable/modules/svm.html#support-vector-machines

Scikit-Learn, 2023. *Importance of Feature Scaling.* [Online] Available at: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html [Accessed 20 March 2024].

Sekeroglu, B., Dimililer, K. & Tuncal, K., 2019. *Student performance prediction and classification using machine learning algorithms.* s.l., Proceedings of the 2019 8th International Conference on Educational and Information Technology.

Shahiri, A. M., Husain, W. & Rashid, N. A., 2015. The Third Information Systems International Conference: A Review on Predicting Student's Performance using Data Mining Techniques. *Procedia Computer Science,* Volume 72, pp. 414-422.

Shaikh, M. K. & Shah, T., 2022. Factors Affecting Computer Science Student's Academic Performance During Covid-19. *Journal of Engineering Education Transformations ,* 36(2), pp. 2349-2473.

Shorfuzzaman, M. et al., 2019. Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Computers in Human Behavior,* Volume 92, pp. 578-588.

Sivasakthi, M. & Padmanabhan, K. R., 2023. Prediction of Students Programming Performance Using Naïve Bayesian and Decision Tree. In: *Soft Computing for Security Applications. Advances in Intelligent Systems and Computing.* Singapore: Springer, pp. 97-143.

Song, Z., Sung, S.-H., Park, D.-M. & Park, B.-K., 2023. All-year dropout prediction modeling and analysis for university students. *Applied Sciences,* 13(2), p. 1143.

Tatar, A. E. & Düştegör, D., 2020. Prediction of academic performance at undergraduate gradua- tion: course grades or grade point average?. *Applied Sciences,* 10(14), pp. 49-67.

Times Higher Education, 2023. *More than 40,000 students drop out of UK university courses.* [Online] Available at: https://www.timeshighereducation.com/news/more-40000-students-drop-out-uk-university-courses[Accessed 6 March 2024].

Uliyan, D. et al., 2021. Deep learning model to predict students retention using BLSTM and CRF. *IEEE Access,* Volume 9, pp. 135550-135558.

Vaarma, M. & Li, H., 2024. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society,* Volume 76, pp. 1-10.

Vinod, K. P. & Bhatt, V. K. K., 2019. Performance Prediction for Post Graduate Students using Artificial Neural Network. *International Journal of Innovative Technology and Exploring Engineering,* 8(7S2), pp. 446-454.

Vries, T. d. & Chawla, S., 2009. *Density-Preserving Random Projection with Application to Local Outlier Detection,* s.l.: Research Gate.

Waheed, H. et al., 2020. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior ,* Volume 104, pp. 106-189.

Waheed, H. et al., 2020. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior,* Volume 104, pp. 106-189.

Yanagiura, T., Yano, S., Kihira, M. & & Okada, Y., 2023. Examining Algorithmic Fairness for First- Term College Grade Prediction Models Relying on Pre-matriculation Data. *Journal of Educational Data Mining,* 15(3), pp. 1-25.

Yukselturk, E., Ozekes, S. & Türel, Y. K., 2014. Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning,* 17(1), pp. 1027-5207.

Yu, R., Lee, H. & Kizilcec, R., 2021. *Should college dropout prediction models include protected attributes?.* Germany, Proceedings of the Eighth ACM Conference on Learning @ Scale ACM, Virtual Event .

Appendix A – Ethics Approval Form

Student Project Approval Form

LD7083 Computing and Digital Technologies Project

You should use this document if you intend to use one of the existing module level approval ethics applications. Please complete this document and discuss your study with your supervisor before you collect any data. *Failure to complete this document and have all aspects signed off and approved by your supervisor risks a notable deduction in your grade and may risk a case of Academic misconduct. Please see the module Bb site for more details.*

Supervisor sign off	
Ethics form complete	☒
Ethical concerns acknowledged	☒
Research tool(s) checked	☒
All relevant forms included (consent etc.)	☒
Is not high risk	☒

Please ensure that your project meets the conditions of the existing ethics application (available on Module Bb site). ***If it does not, then you will need to submit a full ethics application instead***.

Student Name:	Kelechi Uzoukwu
Project Title:	Predicting and Enhancing Students' Academic Performance using Machine Learning Techniques.
Supervisor Name:	Rose Fong
Ethics application you are amending (check box):	☐ Low-risk Lab-based research ☐ Low Risk Secondary Data Science project (yes) ☒ Medium Risk Secondary Data Science project from the private domain required membership ☐ Questionnaire/ survey Study ☐ Interview Study or other Usability Study

Introduction to the project: *Treat like an introduction to the study. Why is your proposed study important? What has already been done on the topic? How does your proposed study 'fit' with the current literature and what does it add? What is the aim of the proposed study? Make reference to appropriate studies.*

There is a growing demand for data-driven approaches to education, as educators and policymakers seek to improve student outcomes and optimize resource allocation. Traditional educational systems often fail to consider non-academic factors that significantly impact a student's academic performance such as family dynamics and socio-economic disparities. As a result, interventions are

often generic as they may not accurately address the underlying reasons for underperformance (Malik & Jothimani, 2023).

Educational institutions and online learning platforms collect large amounts of relevant data that are often too large and complex for simple data analysis, which has led to the development of data mining and machine learning techniques specially designed for the education sector.

Evidently, different machine learning models and algorithms produce different prediction results, and this could be as a result of differences in non-academic variables that were not taken into consideration. Variables like socio-economic factors, family dynamic, personal circumstances and other environmental variables have been opted out of empirical research in this domain. Therefore, this paper recognises the methodological gap and will consider the non-academic factors when predicting academic performance in its study. This paper addresses this issue by analysing extensive datasets beyond academic records and developing a robust predictive model that considers the dynamic and unique characteristics of each student's learning experience.

By employing machine learning techniques, this paper empowers educators with accurate predictions so they can tailor their teaching strategies to address individual student needs. Additionally, this paper seeks to encourage efficient resource allocation by identifying factors that influence academic performance, leading to an equitable distribution of educational benefits.

Methodology: *Please complete the table below, using the following info to guide you. Write this as a future tense method. Describe the* ***participants*** *that you will recruit, how many you are going to recruit, and indicate if you have any additional exclusion criteria. Include the* ***research design*** *(e.g. randomised/repeated measures/quantitative/qualitative/case study etc) and detail of your proposed* ***procedures*** *(i.e., how are you collecting the data?). Include information on all of the equipment you plan to use. If this is a low-risk study, outline how you will extract data and list the criteria you will use to do this. Somebody should be able to read this and replicate it. Describe all planned* ***data analysis*** *for both quantitative (e.g. t-tests, ANOVA, correlation etc.) and qualitative (content analysis, thematic analysis etc.) data. If doing a low-risk study explain how you intend to analyse the data you have collected. Use literature to justify your method.*

1. Is this a low-risk secondary data or lab-based study? If Yes please go to questions 6 and 7.	☒ No
2. Who are your participants and what is the inclusion criteria?	Not applicable, it is medium risk data science project
3. How many will you recruit and from where?	Not applicable, it is medium risk data science project

4. Are there any exclusion criteria (reasons why people should not participate)?	Not applicable, it is medium risk data science project
5. Research design:	Quantitative Analysis
6. Procedures (describe what you will do to collect data, include all equipment/methods you plan to use).	Anonymised data will be downloaded/collected from university's database.
7. Data analysis methods:	Inferential statistics, Hypothesis Testing, Machine learning algorithms (Classic and Ensemble)
8. Additional information:	NIL

Health and Safety: *Relevant risk assessments are listed in the ethics application. If your project needs additional risk assessments, then you will need to submit a new ethics application. Please identify the elements of the listed risk assessment that are relevant for your study and the risk assessment(s) you are working with.*

Please check the relevant boxes*:

☐ HL_RISK_173 Testing in an external environment

☐ HL_RISK_722 face to face interview

☐ HL_RISK_727 Group interview

Areas of potential risk *Please indicate how you will eliminate, or as a minimum ameliorate, the following areas of potential risks throughout the processes of research design, data generation, data analysis and dissemination*		
Area of risk	**Questions relating to this risk**	**How will you mitigate against this risk?**
Avoiding harm to all involved in or potentially affected by the research	How will you ensure that your participants/ respondents come to no harm (psychological; emotional; physical). e.g. not subjecting them to questioning about sensitive issues without advance agreement?	N/A
	How will you ensure your own safety (beyond just physical) in undertaking the Enquiry?	N/A
Ensuring the anonymity of all	How will you ensure anonymity in collecting/generating data	N/A

participants/respondents	How will you ensure anonymity in reporting the data?	N/A
Gaining informed consent from all participants / respondents	How will you ensure respondent/participant consent in advance? You should provide a copy of the necessary consent form/s with this document	N/A
	(How) might participants/respondents be able to withdraw their data?	N/A
Avoiding deception	How will you promote accuracy in recording, analysis, reporting of the data/findings?	N/A
Data storage and destruction	How will you transport and store your data securely (e.g. password protected; cloud storage)	Data will be kept in my laptop which is password protected.
	How will you destroy the data and when?	Following the completion of the research, any data will be retained and stored on the student's University OneDrive folder in accordance with the Data Protection Act. Data will be destroyed after a maximum of 3 years following the conclusion of the study. All data not being presented in the research will be erased from all mentioned storage devices and thoroughly digitally disposed of.
Secondary data sets	*Is your data set(s) from a domain requires membership?*	no
	Does this data set can be used for educational or academic research purpose?	yes

Please check this box after you have read and understood ethics and health and safety information.

☒ I confirm I have read the University's health and safety policy and ethics policy. I have read and understood the requirement for the mandatory completion of risk assessments and that my study does not deviate from the module level approval ethics forms on Blackboard.

Further information (add below, if applicable)

- Consent forms
- Participant information sheet
- Debrief form

- Recruitment materials
- Permission letters
- Data collection tools

Student's Name and sign *Kelechi Uzoukwu* **Kelechi Uzoukwu**	**Date 13/03/2024**
Supervisor's name and sign **Rose Fong** **(Name) Rose Fong**	**Date** **13 Mar. 24**

Appendix B – Meeting Logs

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	February 21st. 3pm – 5pm
Meeting Number:	1
Mean of the meeting:	Online

Brief Summary of Discussion (200 words max):

We shared our dissertation topics and the reasons for choosing them. She also explained the importance of showing originality in our dissertation work. We were asked to draft the introduction chapter of our dissertation before the next meeting.

Agreed Actions:

-Draft the introduction chapter of our dissertation before the next meeting.

-Identify a key literature for your research topic for discussion next week

Student signature: ……*Kelechi Uzoukwu*……

Supervisor signature: ……

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	February 28th. 3pm – 5pm
Meeting Number:	2
Mean of the meeting:	Online

Brief Summary of Discussion (200 words max):

We shared our research objectives, and Rose explained the criteria for drafting SMART research objectives. We discussed literature themes and how to integrate them into our literature reviews. Rose asked us to finalise our research objectives and review the mind map we completed for our proposal module to ensure the themes matched our research questions.

Agreed Actions:

-Draft outline of literature review chapter
-Identify literatures about different research methods for next meeting discussion.
-Identify three themes for literature review.

Student signature: ……*Kelechi Uzoukwu*……

Supervisor signature: ……

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	March 6th. 3pm – 5pm
Meeting Number:	3
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): We discussed the checklist for a good literature review and shared the drafts with each other and our supervisor.	
Agreed Actions: -Draft outline of literature review chapter -Identify literatures about different research methods for next meeting discussion. -Identify three themes for literature review.	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	March 13th. 3pm – 5pm
Meeting Number:	4
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): Rose shared the checklist/requirements for a good research methodology and asked us to expand on our methodology chapter in our proposal. I sent her my draft of chapters 1-2 for review.	
Agreed Actions: -Draft research methodology chapter	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	March 20th. 3pm – 5pm
Meeting Number:	5
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): We discussed our progress with the interim report and shared the drafts amongst ourselves for peer feedback. We also reviewed the checklist for the interim report. We were reminded to submit the report on March 21st. We completed and submitted the research application form on Blackboard.	
Agreed Actions: -Finalise interim report (chapters 1-3) and submit for feedback. -We were reminded that 1-1 meetings were to begin after the easter break.	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	April 10th. 11am – 11:30m
Meeting Number:	6
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): Today was my first 1-1 supervisory meeting and we discussed the feedback/corrections on my submitted interim report. Rose advised me to amend the literature review chapter and perform a more detailed critical analysis. We discussed my progress with chapter 4 (model development). I asked questions regarding data collection, and Rose provided a list of websites to obtain public datasets.	
Agreed Actions: -Draft model development (data analysis) chapter -Implement corrections made by supervisor on interim report	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	April 17th. 11am – 11:30m
Meeting Number:	7
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): We discussed the elements of a results and discussion chapter. Rose advised me that chapter 5 should be dedicated for discussing analysis results, while chapter 6 (conclusion) should be dedicated to reflecting on the lessons and challenges I encountered while completing the project. I agreed to send chapters 1-5 to her before the next meeting.	
Agreed Actions: -Draft chapter 5 (results and discussion) and submit for feedback	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	May 1st. 11am – 11:30m
Meeting Number:	8
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): We discussed the feedback made on my submitted chapter 5 and I agreed to send the completed dissertation report before our next meeting.	
Agreed Actions: -Draft chapter 6 (conclusion) and submit report for feedback	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	

Record of Supervisory Meeting

Student Name: Kelechi Uzoukwu	Programme: Big Data and Data Science Technology
Supervisor: Rose Fong	

The minimum number of formal contacts between students and Supervisor(s) will normally be 6 hours. However, this contact may be maintained in part via video conferencing or email where necessary. Formal supervisory contact meetings and their outcomes must be recorded using the template below (copy the table below for each meeting)-and attached in the appendix of your dissertation report.

Date & starting/ ending time of Meeting:	May 8th. 11am – 11:30m
Meeting Number:	9
Mean of the meeting:	Online
Brief Summary of Discussion (200 words max): At our last meeting today, Rose provided final feedback on my report and advised me to include the meeting logs in the appendix.	
Agreed Actions: -Attach meeting logs to appendix - Make final edits and submit dissertation	
Student signature: ……*Kelechi Uzoukwu*…… Supervisor signature: ……	