MSc Project - Reflective Essay

Project Title:	Using Machine Learning Analysis to Predict Student Success in Higher Education
Student Name:	Hamza Noor Khan
Student Number:	180426352
Supervisor Name:	Chathura Kalpanee Sooriya Arachchi & Huan Zhang
Programme of Study:	MSc Computer Science

Taking on this research project about predicting student success in higher education was both an eye-opening and thought-provoking experience. It began with exploring factors influencing academic achievement and quickly evolved into a deeper reflection on the complexities of education systems, data analysis and machine learning modelling. This reflective essay aims to showcase the behind-the-scenes experiences that shaped my research, offering insights into the decision-making processes, obstacles, and personal growth that would not typically be included in the research paper. Through this reflection, I hope to illustrate how this project has not only enhanced my understanding of educational success but also expanded my perspective on the broader implications of research in higher education.

People often wonder why I chose to focus my research on predicting student success when there are many other innovative and exciting projects available. However, this topic holds personal significance for me for several reasons.

The inspiration for this research began with a conversation I had with a director on my team at work. We discussed the potential applications of AI and predictive modelling in the education sector, and as I shared my research ideas, she became increasingly interested. She confided that she had struggled in her educational journey and ultimately dropped out. She emphasised that if there had been some form of early intervention, her experience could have been vastly different. This insight resonated with me deeply.

My research is dedicated to developing models that predict students' final grades as early as possible, allowing us to identify those who may be at risk, much like the director on my team. By doing so, we can provide timely support and potentially transform their educational experiences for the better. This approach not only aims to increase graduation rates and reduce dropout rates but also to enhance the overall higher education experience, ensuring that students feel valued and supported by their universities.

Looking further at the project and analysing the strengths and weaknesses of the project are also an important step.

One of the key aspects of the project is the dataset used, which serves as both a strength and a potential weakness. The dataset's strength lies in its richness and the unique composition of the target variable, which includes 'Dropout,' 'Graduate,' and 'Enrolled' categories. This categorisation is particularly intriguing, as it's uncommon in the literature I read and provides a fascinating opportunity to explore how different models predict these outcomes and identify the variables that influence each category.

Additionally, the dataset encompasses a wide range of variables, such as the economic climate during the students' studies and their parents' occupation and education levels. This diversity allowed me to test several interesting hypotheses. I also appreciated the dataset's inclusion of students from various courses, which enabled me to compare features and assess their importance across different modules—a comparison that is often overlooked in other research.

However, the dataset also has some weaknesses, particularly concerning the number of entries for specific variables. For instance, certain modules are more popular than others, leading to significant differences in enrolment numbers. This disparity can result in overfitting in machine learning models, which is something I encountered during the modelling process.

Another one of the project's key strengths of this research is its integration of various machine learning models and the comparative analysis conducted between them. The research paper provides valuable insights into leveraging machine learning to predict student success in higher education. Several predictive models were applied to the student learner profile dataset, with their accuracy outcomes presented in Figure 1.

Fig. 1. Accuracy of Predictive Models Using Student Learner Profile Dataset

Model	Accuracy
Logistic Regression	0.75
XGBoost	0.81
Decision Tree	0.62

I would like to note that the only two directly comparable models in the results above are Logistic Regression and XGBoost, as they were both trained on the same dataset. Although the Decision Tree model used the same features, its results are based on a multi-class classification task, unlike the binary classification used for Logistic Regression and XGBoost.

XGBoost emerged as the most accurate model, closely followed by Logistic Regression, with Decision Tree performing the least effectively.

The performance of XGBoost over Logistic Regression can be attributed to its advanced boosting algorithm, which builds and refines multiple weak learners (typically decision trees) to optimise predictive accuracy. XGBoost works well at capturing complex, non-linear relationships in the data, making it particularly effective in scenarios where the dataset contains intricate patterns and interactions. In contrast, Logistic Regression is a linear model that may not fully capture such complexities, leading to its slightly lower accuracy in this context.

In addition to the learner profile dataset, two models were also tested on a dataset containing student grades to assess their effectiveness in predicting student success. For this analysis, Linear Regression and Logistic Regression models were utilised, with the key performance metrics detailed in Figure 2.

Fig. 2. Showing accuracy of each model using dataset containing student grades

Model	Accuracy	MSE	R-Squared
Linear Regression		7.65	0.72
Logistic Regression	0.65		

One of the paper's weaknesses is its emphasis on understanding student learner profiles to predict success in higher education, while placing less emphasis on utilising student grades. The dataset that included student grades—comprising prior grades, and grades and evaluations from the first and second semesters—was relatively limited. The model results indicate that this set of variables may not be optimal. The analysis could have been enhanced by incorporating a broader range of variables related to student grades.

- Presentation of possibilities for further work

While my research has yielded promising results and identified potential avenues for further investigation, there are still opportunities to expand the scope of predicting student success within the broader context of education.

Although this topic extends somewhat beyond the traditional boundaries of computer science, an intriguing direction for future research would involve integrating predictive machine learning models with educational psychology. This could begin by reviewing existing literature on how psychological factors influence student performance and identifying key drivers of academic success. Understanding these factors would inform the selection of relevant variables for inclusion in predictive datasets.

Furthermore, while predicting student success is crucial, it represents only part of the equation. An equally important challenge lies in identifying the most effective interventions to support students at risk of dropping out. My research and discussions with educational professionals suggest that different students require tailored support. Developing predictive models that recommend specific interventions for individual students could streamline the process and reduce the need for manual intervention, thereby enhancing the efficiency of educational support systems.

Although my current study focuses on students in higher education, another valuable area for exploration is predicting student success at the secondary and sixth form/college levels. These stages are foundational to university admission, with sixth form in the UK being particularly challenging due to its impact on university offers. Investigating whether performance during sixth form can predict university success could provide valuable insights.

Finally, a more ambitious area of research could involve developing a comprehensive scoring system that accounts for a wide range of factors, including grades, family circumstances, and educational special needs. Such a system would enable institutions to make more informed decisions by evaluating students based on a holistic set of criteria rather than relying solely on academic performance.

- Work that you would have conducted if you had more time.

The research conducted a comprehensive analysis to understand how students' learner profiles and academic performance can be leveraged to predict their success in higher education. However, there are several additional areas of exploration that could have been pursued with more time.

Firstly, the study utilised a dataset from a Portuguese database, specifically from the years 2009 to 2018. Given more time, the first step would have been to incorporate a more recent dataset from the Queen Mary University of London database. The primary reason this was not initially feasible was the potential three-month approval process required by the board. An intriguing avenue for future research would be to compare how students from different countries are impacted and to identify which variables most significantly influence the prediction of student success.

In addition to using an updated dataset, incorporating additional variables could further enhance the analysis of their impact on the target variable. Examples of such variables include student attendance, employment status (whether the student holds a full-time or part-time job), coursework performance, and class participation.

At the beginning of the research, I also considered developing a survey to be distributed to students and professors, aimed at gathering insights on how artificial intelligence could be used to predict student success and whether this would be beneficial. While creating predictive models that deliver strong results is valuable in theory, it is important to assess whether these models are needed and would be beneficial, a determination that could be made by engaging directly with students and professors. However, after consulting with my supervisor, it became apparent that this approach might be time-consuming, and there were concerns about low response rates from students and professors, possibly due to time constraints or lack of interest in participating in such a survey

During my course, I had the opportunity to explore the subject of Ontology, which piqued my interest in its practical applications. After discussing potential projects with my supervisor, we decided that building an ontology to predict student success in higher education could be a valuable. Using OWL, I could develop a structured framework that contains key concepts such as student characteristics, academic factors, institutional support, and external influences. This ontology would serve as a foundation for understanding the complex interplay of factors that contribute to academic outcomes, ultimately aiding in the development of predictive models and personalised interventions to support student achievement. However, after beginning the project and understanding the time I had it didn’t seem feasible to fit this in while building the machine learning models.

• Awareness of Legal, Social Ethical Issues and Sustainability

There are significant legal and ethical considerations involved in predicting student success in higher education. Before initiating this project, I encountered an important issue regarding the dataset provided by Queen Mary University of London: student names were removed to ensure that no individual could be identified, in compliance with legal and ethical standards. This also helped eliminate potential bias, particularly if I were acquainted with any of the students. While this anonymisation might pose challenges for external entities wishing to use the dataset, the data is primarily intended for internal institutional use. Institutions can utilise their own databases to support their students, and if external analysis is required, student names could be replaced with unique identifiers to maintain confidentiality.

From a sustainability perspective, this project is highly adaptable and could be implemented across various countries and universities. Although grading systems and evaluation techniques differ globally, minor adjustments would enable widespread application. Most institutions possess vast amounts of data that, when properly extracted with the appropriate variables, can be used to predict student success in higher education. This approach has the potential to greatly benefit both institutions and students, leading to higher pass rates, improved academic performance, and ultimately, a positive impact on the broader economy.

• Critical analysis of the relationship between theory and practical work produced

At the outset of the project, I identified several key steps to guide my approach. The most significant challenge I anticipated was related to the dataset acquisition. I expected that sourcing the right data would be difficult, particularly because I needed a specialised dataset that included student learner profiles, economic circumstances during their studies, and academic performance.

Initially, this proved to be a considerable hurdle, as the dataset requirements were quite specific. Fortunately, I discovered a pre-processed dataset from a Portuguese database that met my criteria and was in a suitable format. Without this dataset, the project might have been unfeasible, as obtaining similar data from my university would have required security clearance and the extraction of information from multiple databases.

If I were to replicate this approach with more recent data from my university, it would likely pose significant challenges. There would be numerous steps and potential obstacles between requesting access and actually obtaining the data—obstacles I might only fully understand if I pursued that route.