Using Machine Learning Analysis to Predict Student Success in Higher Education

Hamza Noor Khan
Student Number: 180426342
Supervisor: Chathura Kalpanee Sooriya Arachchi & Huan Zhang
Student at Queen Mary University of London
London, United Kingdom
Student Email: h.n.khan@se19.qmul.ac.uk
Computer Science MSc

***Abstract* – This research paper addresses the critical challenge of predicting student success in higher education, particularly focusing on identifying students at risk of dropping out, while also looking at what differentiates students who dropout from those who graduate. By analysing a dense dataset that includes social and economic factors beyond students' control, this study employs advanced machine learning models and boosting techniques to predict student success. The aim is to develop comprehensive predictive models that can accurately identify students who may face difficulties in completing their education. By leveraging state-of-the-art machine learning methods, this research provides a holistic approach to understanding the multifaceted factors influencing student success in the current decade. The findings have the potential to significantly reduce dropout rates and enhance graduation outcomes. Thereby contributing to the broad goal of improving higher education retention and success rates.**

INTRODUCTION

Understanding the importance of student success is a topic that should resonate with everyone, not just students and educational institutions. A student's performance profoundly influences their future trajectory, including the lifestyle they will lead, the type of job they will secure, and even the social circles they will associate with. For institutions, fostering student success is crucial for attracting talent in the years to come. A good reputation enables an institution to attract the best students, leading to new research and studies that benefit the university.

Moreover, improving student success can have a significant positive impact on the economy. Higher graduation rates mean more individuals entering the workforce with higher earning potential, contributing to the economic health of the country by boosting GDP and reducing unemployment. Fewer students dropping out can also reduce the likelihood of crime. Christopher paper [1], highlights that dropping out of college is positively related to increased crime throughout one's life.

The importance of higher education today is immense, with most job applications requiring some form of higher education qualification. However, significant issues within the system need to be addressed. Having recently navigated the higher education system myself; I have observed that the primary focus is often on the final grade. While the final grade is undoubtedly important, it should not be considered the sole indicator of student success. The factors contributing to this final grade warrant closer examination.

Higher education institutions enrol thousands of diverse students and employ teaching staff with varied instructional styles. One significant challenge these institutions face is identifying students who struggle with their modules or are at risk of failing and providing them with effective support solutions. I personally experienced difficulties with some modules and did not receive the necessary assistance from any teaching staff member. While I understand that instructors manage thousands of students, implementing a machine learning model that could be run periodically, such as every six weeks or adjusted to the institution's needs, to identify students at risk of failing and offer additional resources could significantly improve student success and reduce dropout rates.

My personal experience, particularly spending over 80% of my higher education at home due to the COVID-19 pandemic, has highlighted the impact of external factors on academic performance. This research aims to delve deeper into these factors, particularly those beyond students' control, and to utilise advanced machine learning techniques to better predict and understand student success. By doing so, we hope to provide a more comprehensive understanding of what influences academic outcomes, ultimately helping to reduce dropout rates and improve graduation rates in higher education.

While my experience and that of many students during COVID-19 was exceptional, research consistently highlights the importance of external factors in contributing to student success. For instance, an article by HEPI [2], the UK's only independent think tank devoted to higher education, determined that "seven-in-ten students have considered dropping out of higher education since starting their degree, with nearly two-fifths citing rising living costs as the main reason." Similarly, Nutmalitasari's paper [3] notes that "31% of

the respondents were busy working because of financial difficulties."

By addressing these external factors through targeted interventions and predictive models, higher education institutions can better support their students, leading to improved academic outcomes and reduced dropout rates. This, in turn, has far-reaching benefits, including enhancing institutional reputation, contributing to economic growth, and reducing crime rates, thereby fostering a more prosperous and stable society.

The aims of this project are to answer the following questions.

1) Can learner profile of a student be used to determine the student's success in higher education?
2) Can student's previous qualification and semester one grade be used to determine the student's semester two grade and finally outcome of higher education?
3) Looking at student's success across multiple courses and providing a comparison plus looking for patterns.

This will be done by achieving the following objectives:

1) Finding a dataset that contains key data points about students.
2) Building, training and testing minimum 2 machine learning models using different methodologies.
3) Applying boosting methodologies alongside cross validation to confirm and test the results.

DISCLAIMER ABOUT THE DATASET

Mónica V. Martins' [4] utilises the same dataset used for this study. Due to the nature and time constraints of this project, the above dataset has been selected as it contains the crucial variables which need to be analysed.

Initially, I aimed to obtain a dataset from my institution, Queen Mary University of London. However, the approval process and preparation of the dataset would have been too time-consuming, making the project infeasible. The dataset used throughout this research paper is indicative of the type of data available at all universities and is readily accessible.
More comprehensive datasets can be created from university student databases for future research.

One limitation of Martins' study is the lack of detail regarding the classification of students and the differentiation of various groups. The dataset includes students from a range of courses, which is both rare and valuable. This research paper aims to run predictive models across different courses to determine if the predictions vary by course, along with exploring additional factors. Incorporating this perspective into my research will not only enhance the existing body of knowledge but also provide practical insights that can be applied across various educational institutions to improve student success outcomes.

III. RELATED WORK

Taking into consideration, this research topic is one that has been extensively studied. Thus, before starting this project, it was crucial to review previous work in this field. To guide this paper's analysis, the following research questions were applied.

- How can this paper's research differ against others and importantly add value in the provided research area?
- Have any issues in this area been encountered before?
- If implementations exist with similar problems, which Machine Learning Models were used?

The first paper of interest focuses on using Bayesian Belief Networks (BBNs) to predict student performance and progress in higher education institutions [5]. The key goal of the study was to develop a framework that leverages BBNs to forecast students' future GPAs based on their performance in initial semesters. This paper provided a practical application of BBNs in predicting academic outcomes, which was pivotal for demonstrating how early performance data can be used to make accurate predictions about future success.

The framework presented in the paper differs from traditional performance tracking methods (which I will be using) by integrating probabilistic modelling and machine learning techniques to offer more precise and actionable insights into student progress. Unlike simpler predictive models, this BBN-based approach allows institutions to provide early advisement and targeted interventions, which can significantly enhance student retention and graduation rates.

Although this research paper's approach represents a sophisticated use of Bayesian Belief Networks for predictive modelling, aiming to explore whether simpler machine learning frameworks could yield comparable results. This is motivated by the desire to develop models that are more accessible and understandable to a broader audience. Simplifying the models could make them easier to apply and interpret, potentially broadening their utility and effectiveness in educational settings.

The next paper is particularly pertinent to the research carried out in this paper as it offers a thorough review of approximately 70 recent studies on predicting student performance through machine learning techniques [6]. The review encompasses a range of methodologies, including supervised learning, unsupervised learning, collaborative filtering, recommender systems, and artificial neural networks. Significantly, the paper notes that around 70% of the reviewed studies concentrate on predicting student performance. Additionally, it identifies early dropout prediction as a promising area for further investigation and suggests that exploring new predictive techniques could be advantageous.

The research applied throughout this paper aligns with this focus, emphasising the need to understand dropout risks from an early stage, even prior to university enrolment. It also examines how performance in the first and second semesters influences final outcomes.

Al-Bahri Mahmood [7] serves as a foundational reference for my research perspective. Mahmood begins by emphasising the widespread adoption of AI across various industries and the critical need for its integration into the education sector. Given the vast amounts of data available on students, AI can be effectively utilised to predict academic success. I concur with Mahmood's assertion that universities possess extensive data on students, including their previous and current grades, as well as demographic information, all of which can be leveraged to forecast student performance.

Mahmood references a pioneering study by Professor Sotiris Kotsiantis in 2003, which highlighted the underutilisation of available data and the untapped potential within the field. Although significant advancements have been made since that study, there is still room for further progress.

In his research, Mahmood used a dataset comprising high school grades in Mathematics, Arabic, Social Studies, and English from students enrolled in Computer Science and Physics degrees, predicting their first-year grades for each semester. In contrast, the dataset used in this study is considerably more comprehensive and detailed, enabling me to test additional hypotheses that were beyond the scope of Mahmood's study.

DATASET

The dataset utilised in this research required minimal preprocessing. Firstly, the numerical categories were reverse-engineered back to their original text labels to aid a better understanding of the data. After this, the original data dataset was used to run the machine learning algorithms.

Secondly, given that the dataset used was in a numerical format, testing for any linear relationships between the variables and the target variable was necessary. Upon further investigation, variables like parents' occupations and education levels lacked meaningful numerical order. This called for mapping seeking to reorder these variables and assign significance to the numbers. Ultimately, allowing for the variables to be ranked from lowest to highest for significant testing.

An example of this processing is shown in Fig 1 below, which displays the ranking of mothers' qualifications.

Fig. 1. Table shows a sample of variable 'Mothers Qualification' where variable numerical input has been changed

New Number	Original Key	Qualification
1	34	Unknown
2	35	Can't read or write
3	36	Can read without having a 4th year of schooling

DATASET INVESTIGATION

The dataset used contains 4,424 students and 36 feature variables. These variables include details about students' parents' jobs and qualifications, students' first and second semester grades, and the economic climate during their studies.

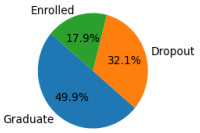
A central question of this paper is whether a student's learner profile can predict their success in higher education.

The dataset used contains several key features that enable the attribution of specific characteristics to students. To test this paper's hypothesis, a subset of the original dataset was created with the following columns: 'Marital status', 'Application mode', 'Application order', 'Course', 'Daytime/evening attendance', "Mother's qualification", "Father's qualification", "Mother's occupation", "Father's occupation", 'Nationality', 'Displaced', 'Educational special needs', 'Debtor', 'Tuition fees up to date', 'Gender', 'Scholarship holder', 'Age at enrolment', and 'International'.

A key question explored in this paper is whether students' prior qualifications, along with their performance in the first and second semesters, can be predictive of their success in higher education. To examine this hypothesis, a subset of the dataset was constructed, including the variables 'Previous Qualification (Grade)', 'Curricular Units 1st Semester (Evaluations)', 'Curricular Units 1st Semester (Approved)', 'Curricular Units 1st Semester (Grade)', 'Curricular Units 2nd Semester (Evaluations)', 'Curricular Units 2nd Semester (Approved)', and 'Curricular Units 2nd Semester (Grade)'.

Based on my research and prior knowledge, I know that more students graduate than drop out, leading to an imbalance in the data. Taking a deeper dive into the Target variable in my dataset we can see this more clearly.

Fig. 2. Pie Chart of imbalance of my Target Variable



This issue arises because the minority classes, which are of primary interest, tend to be overlooked by machine learning algorithms that focus more on the majority class—in this case, Graduates—while ignoring the minority classes such as Dropouts and Enrolled students.

Class imbalance can be addressed through oversampling and Undersampling techniques. Oversampling involves generating new samples for the under-represented class. However, a drawback of this method is that it often duplicates existing entries in the minority class, failing to introduce new data points. To mitigate this, methods like SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are employed. SMOTE generates synthetic samples by interpolating between selected instances

and their nearest neighbours uniformly. In contrast, ADASYN focuses on the more challenging minority instances to classify, assigning weights based on classification difficulty to produce more targeted synthetic samples.

On the other hand, Undersampling reduces the number of samples from the over-represented class. This technique is less commonly used because it can result in the loss of potentially valuable information.

To identify the most effective solution for addressing class imbalance, two research papers focused on this issue were reviewed to determine the best technique for application to the dataset.

One key paper by Muhammad Arham Tariq [8] examines the prediction of students' academic performance using imbalanced multi-class educational datasets. The study compares ten resampling methods (including SMOTE and ADASYN) and nine classification models (Logistic Regression, XGBoost, and Random Forest). For simplicity, this review concentrates on the performance of SMOTE and ADASYN. Upon evaluating the results presented by the paper, SMOTE consistently produced superior outcomes compared to ADASYN.

The subsequent paper [9], focused on predictive models for classifying students' academic performance as pass or fail, emphasizes the importance of addressing imbalanced data. To mitigate this issue, the authors employed SMOTE, which resulted in improved model accuracy.

Upon reviewing both papers and analysing the differences in sampling techniques utilized in similar research, SMOTE has been selected as the appropriate technique for application to the dataset in this study.

Shifting focus from class imbalance to other features in the dataset used in this paper, with a particular interest in exploring the module breakdown. One fascinating aspect of this dataset is its richness in terms of the number of courses it includes. From an initial overview, we can compare the rates of Dropout, Enrolled, and Graduate students. This table is useful for identifying courses with the highest dropout rates relative to graduates. However, this research paper aims to address involve understanding the impact of specific courses on the target variable. Do machine learning algorithms detect significant relationships between courses and student outcomes? If so, why do certain courses have higher dropout rates? What types of students enrol in these courses, and are there any common characteristics among them?

Fig. 3. Overview of the 'Module' variable, detailing the distribution of students across 4 modules in terms of dropout rates, graduation rates, and current enrolments

	Dropout	Enrolled	Graduate
Management (evening attendance)	136	54	78
Management	134	108	138
Nursing	118	100	548
Journalism and Communication	101	34	196

An intriguing aspect of this dataset pertains to the detailed insights it offers regarding the marital status of students. Based on my research and prior knowledge, it is generally understood that most students tend to be single. However, the rise of online universities and institutions offering evening classes has led to an increase in the number of mature students pursuing higher education. Consequently, this paper's focus warranted it worthwhile to explore this data further to determine if the model identifies marital status as a significant variable in predicting the target outcome.

Fig 4 provides a comprehensive analysis of this aspect.

Fig. 4. Overview of the 'Married' variable, illustrating the distribution of students across different marital statuses with respect to dropout rates, graduation rates, and current enrolments

	Dropout	Enrolled	Graduate	Dropout (%)	Enrolled (%)	Graduate (%)
Divorced	42	16	33	46.0	18.0	36.0
Facto Union	11	3	11	44.0	12.0	44.0
Legally Separated	4	1	1	67.0	17.0	17.0
Married	179	52	148	47.0	14.0	39.0
Single	1184	720	2015	30.0	18.0	51.0
Widower	1	2	1	25.0	50.0	25.0

VIII. METHODOLGOIES

To develop machine learning models for predicting student success in higher education, the dataset was partitioned into two subsets: the student learner profile and student grades. For the student learner profile dataset, Logistic Regression, XGBoost, and Decision Tree algorithms were applied. Additionally, Random Forest Classifier was run with the specific objective of predicting student success across various courses and assessing the importance of different features for each course. Subsequently, the student grades dataset, Linear Regression and Logistic Regression models were applied. These models underwent iterative testing and refinement. The methodologies employed are detailed in the following sections.

Below is the breakdown of the Machine Learning Models used for the dataset containing information about student learner profile.

Given that the target variable comprises three categories, Multinomial Logistic Regression seemed an appropriate initial method due to its simplicity. Additionally, this approach allows for a comparison with more complex models, helping to determine whether a simpler model might be more effective for this problem.

A. Logistic Regression

1) Pre-processing: Since the dataset came pre-processed, there was not much preprocessing required, however, spot checks were carried out to make sure that everything was in line. Part of the pre-processing however, I applied SMOTE to my dataset, to overcome the class imbalance between my target variables.

2) Training: Prior to training, the dataset was divided into features (X) and the target variable (y). The data was further split into training and test sets using an 80:20 ratio.

As part of the logistic regression implementation, I employed a preprocessing pipeline that included a StandardScaler to standardise the features. The model was run twice: once with balanced data and once without, to assess the impact of the sampling technique on the results. Additionally, GridSearch was applied for hyperparameter tuning to evaluate if it enhanced the model's performance.

While running Multinomial Logistic Regression is a good starting point it does have its limitations. Although it provides a good foundation it may not fully capture the complex relationships within the data, whereas a model like XGBoost performs well in modelling nonlinear interactions and handling a vast array of features through its gradient boosting framework. Which is why I felt it was the right next step.

B. XGBoost Model

1) Pre-processing: Separating the features and the target variable and encoding the target variable for classification.
2) Training: Before initiating the model training, the dataset was organised and split into distinct categories to facilitate effective training and evaluation:
   - Feature Data: The features were divided into training and testing sets with an 80:20 ratio. This ensured that the model was trained on a substantial portion of the data while retaining a separate subset for testing its generalisation ability.
   - Target Data: The target variable, indicating student success, was also split into training and testing sets with an 80:20 ratio. This consistent splitting allowed for an unbiased assessment of the model's performance.

In the preceding section, the use of student learner profiles to predict success in higher education is explored. The dataset includes detailed course breakdowns, encompassing a diverse range of students across various courses. Based on existing research and prior knowledge, it is evident that the calibre of students varies significantly between different courses, leading to variations in student success rates. To address this, a course-specific predictive model was developed for student success and assess its effectiveness. For this purpose, I employed the Random Forest Classifier.

C. Decision Tree Model

1) Pre-processing: No processing required.
2) Training: The dataset was divided into features (X) and the target variable (y). The data was further split into training and test sets using an 80:20 ratio.

The subsequent section is dedicated exclusively to analysing student grades. The dataset employed in this analysis is notably limited, encompassing only 2 to 6 features. This limitation is intentional, as the objective is to assess the feasibility of predicting semester 2 grades based on prior academic performance, including grades from semester 1 and other evaluations. By integrating these predictors, the goal is to forecast final grades and identify students at high risk of underperformance. Early identification of such students would facilitate timely interventions aimed at preventing academic failure.

The final section, the use of student learner profiles to predict success in higher education was explored. The dataset includes detailed course breakdowns, encompassing a diverse range of students across various courses. Based on existing research and prior knowledge, it is evident that the calibre of students varies significantly between different courses, leading to variations in student success rates. To address this, a course-specific predictive model was developed for student success and assess its effectiveness. For this purpose, I employed the Random Forest Classifier.

A. Random Forrest Classifier

1) Pre-processing: No pre-processing required & no sampling techniques too.
2) Training: Like the above models, there is an 80:20 split between training and test sets.

The model was executed across three distinct iterations. The Random Forest classifier, which possesses an inherent sampling function, was examined to assess its impact. The first iteration was conducted without any data balancing. In the second iteration, data balancing techniques were applied. For the final iteration, a combination of balanced data and hyperparameter tuning was utilised, specifically employing Randomised Search Cross-Validation to optimise the model.

Below is the breakdown of the Machine Learning Models used for the dataset containing information about student grades.

In this phase of my research, the objective was to determine whether incorporating previous academic performance alongside semester one results could effectively predict semester two outcomes. Given that semester two grades are continuous variables, Linear Regression was selected as the most appropriate model for this analysis. Despite the relatively small dataset, the rationale behind this approach was to assess whether these variables could serve as early indicators of students at risk of receiving low grades in semester two, thereby enabling timely intervention strategies.

A. Linear Regression

1) Pre-processing: No pre-processing required.
2) Training: There is an 80:20 split between training and test sets. Along with this, normalisation was performed on x_train and x_test and a feature for bias was added.

The final model employed in this section was Logistic Regression. This analysis utilized previous grades, along with semester one and semester two results, to predict students' final grades. The underlying approach was like that of the Linear Regression model, aiming to identify whether students are at risk of dropping out as early as the first year. This early identification would enable the implementation of timely interventions to reduce dropout rates.

B. Logistic Regression

1) Pre-processing: No pre-processing required.
2) Training: Like the Linear Regression.

IX. RESULTS

Below are the model results for the dataset containing student learner profile.

A. Logistic Regression

Overall, the model demonstrates a decent performance in predicting student success in higher education, with an accuracy of 76%. When analysing the classification report, it can be shown that the model is particularly adept at predicting students who dropout (class 0), as evidenced by a higher recall of 0.83 and an F1-score of 0.79. This indicates that the model is proficient in identifying actual dropouts but also suggests a susceptibility to classifying students as dropouts even when they may graduate, as reflected by the lower precision of 0.75 for class 0.

Conversely, the model exhibits some challenges in accurately predicting graduates (class 1). Although the precision for class 1 is relatively high at 0.79, indicating a strong ability to identify true positives among predicted graduates, the recall is comparatively lower at 0.69. This discrepancy results in an F1-score of 0.73, suggesting that a significant portion of actual graduates are being misclassified as dropouts.

Fig. 5. Logistic Regression Classification Report Results

	Precision	Recall	F1-Score	Support
0 (Dropout)	0.75	0.83	0.79	464
1 (Graduate)	0.79	0.69	0.73	420
Accuracy			0.76	884
Macro Avg	0.77	0.76	0.76	884
Weighted Avg	0.76	0.76	0.76	884

B. XGBoost

The XGBoost model shows robust performance in predicting both student dropouts and graduates, with a slightly higher recall for graduates (0.84) compared to dropouts (0.78). This indicates that the model is more effective at identifying students who will graduate. However, it still maintains a good balance with a high precision for both classes (0.81 for both dropouts and graduates). The overall accuracy of 81% signifies that the model is reliable in classifying student success in higher education, and the performance metrics suggest that it is well-suited for this predictive task. This performance improvement over the logistic regression model highlights the effectiveness of the XGBoost algorithm in handling complex classification problems.

Fig. 6. XGBoost Classification Report Results

	Precision	Recall	F1-Score	Support
0 (Dropout)	0.81	00.78	0.80	420
1 (Graduate)	0.81	0.84	0.82	464
Accuracy			0.81	884
Macro Avg	0.81	0.81	0.81	884
Weighted Avg	0.81	0.81	0.81	884

C. Decision Tree

The decision tree model's performance, as indicated by the classification report, demonstrates a balanced and consistent level of effectiveness across the three classes. The precision, recall, and F1-scores for each class (0, 1, and 2) are all closely aligned, ranging from 0.61 to 0.65. This suggests that the model is equally capable of identifying true positives and minimizing false positives across all classes.

With an overall accuracy of 62%, the model achieves moderate performance. The macro and weighted averages for precision, recall, and F1-score all stand at 0.62, indicating that no particular class disproportionately influences the model's performance.

In summary, while the model performs consistently across the board, the 62% accuracy suggests that there may be room for improvement in the model's predictive capabilities, potentially through further tuning or the use of more complex algorithms.

Fig. 7. Decision Tree Classification Report Results

	Precision	Recall	F1-Score	Support
0 (Dropout)	0.61	0.60	0.61	444
1 (Graduate)	0.62	0.61	0.62	443
2 (Enrolled)	0.64	0.65	0.64	439
Accuracy			0.62	1326
Macro Avg	0.62	0.62	0.62	1326
Weighted Avg	0.62	0.62	0.62	1326

While the previous models concentrated on identifying the key variables influencing the target variable and determining if we can predict the target variable, this section shifts focus to evaluating differences between various courses and assessing the ability to predict student success on a per-course basis. The dataset used in this study is notably rich in

diverse course offerings, facilitating a comparative analysis. To predict student success in this study, a Random Forest Classifier model was applied.

A. Random Forest Classifier

The Random Forest Classifier model not only forecasts the target variable but also generates a list of key variables for each course. This analysis allows for an examination of how different variables affect each course.

Notably, courses 33 and 9556 have insufficient student numbers, which may lead to model overfitting for these courses and potentially unreliable results. Consequently, the findings for these two courses should be interpreted with caution.

Due to space constraints, detailed classification reports for each course are available in the accompanying notebook.

The model demonstrates robust performance, accurately predicting outcomes for approximately 80% of the courses, while encountering challenges with a few others.

Below are the results for the dataset containing information about students' previous qualifications and year 1 grades + evaluation.

A. Linear Regression

The Linear Regression Model was executed, yielding an R-squared (R^2) value of 0.732. This indicates that approximately 73.2% of the variance in the dependent variable can be explained by the independent variables included in the model. This relatively high R-squared value suggests a strong goodness-of-fit and that the model offers a reasonable explanation of the observed data variation.

B. Logistic Regression

Next, logistic regression using prior grades, first-semester grades, and second-semester grades from the first year to predict final grades in higher education. The model achieved an overall accuracy rate of 65.31%. The classification report reveals the following performance metrics:

Class 0: Precision was 0.83, indicating a high proportion of true positives among the predicted positives. However, the recall was lower at 0.54, reflecting that the model missed a considerable number of actual instances of Class 0.

Class 1: The model demonstrated relatively strong performance for this class, with a precision of 0.73 and a higher recall of 0.81, suggesting effective identification of Class 1 instances.

Class 2: The model had difficulties with Class 2, achieving a precision of 0.32 and a recall of 0.46, indicating challenges in accurately predicting this class.

Overall, while the model performs reasonably well, particularly for Class 1, there is potential for improvement, especially in increasing recall for Class 0 and precision for Class 2.

Fig. 8. Logistic Regression Classification Report Results

	Precision	Recall	F1-Score	Support
0 (Dropout)	0.83	0.54	0.65	316
1 (Graduate)	0.73	0.81	0.77	418
Accuracy			0.65	885
Macro Avg	0.63	0.60	0.60	885
Weighted Avg	0.70	0.65	0.66	885

Now that model performances have been shown throughout this paper. The hypothesis questions enlisted at the start if this paper can be addressed.

Hypothesis: Students with educational needs and those who are displaced are more likely to struggle in higher education, resulting in an increased risk of dropping out.

The primary hypothesis this paper investigated was whether students with educational needs or those who are displaced are more likely to drop out. The rationale behind this hypothesis is straightforward. Students with educational special needs may face challenges such as difficulties in revising, stress during examinations, problems retaining educational information, and limited attention spans, all of which can increase their likelihood of dropping out. Similarly, displaced students, who are seeking refuge in a new country, often experience additional stress related to their situations back home and may find it difficult to adapt to new environments. This stress and difficulty in adaptation can lead to decreased focus on their education, thereby increasing their risk of dropping out.

In the Logistic Regression model the significance of the variables "Educational Special Needs" and "Displaced" is particularly noteworthy. A feature importance analysis revealed that "Educational Special Needs" has a coefficient of 0.024501, while "Displaced" has a coefficient of -0.139673.

The positive coefficient for "Educational Special Needs" suggests that having educational special needs slightly increases the likelihood of graduating. Specifically, for each unit increase in the "Educational Special Needs" variable, the log odds of graduating increase by 0.024501. However, this coefficient is very small and has minimal impact on predicting the target variable compared to other factors.

Conversely, the negative coefficient for "Displaced" indicates a reduced likelihood of graduating. For each unit increase in the "Displaced" variable, the log odds of graduating decrease by 0.139673. In simpler terms, displaced students are less likely to graduate compared to their non-displaced counterparts.

Overall, it was quite surprising how little impact "Educational Special Needs" and "Displaced" had on the Logistic Regression Model, therefore, to make sure that the

above was accurate another model was run to cross validate the results received from the Logistic Regression.

In addition to the Logistic Regression model, this paper sought to determine the key features influencing the prediction of the target variable in the XGBoost model. Specifically, aiming to understand the significance of the "Educational Special Needs" and "Displaced" variables. In the XGBoost model, which outperformed the Logistic Regression model, "Educational Special Needs" ranked lowest in feature importance with a score of 8.0, while "Displaced" had a higher importance score of 89.0.

Overall, considering both models, it is evident that "Educational Special Needs" and "Displaced" have minimal impact on predicting the target variable. This finding is surprising and contradicts my initial hypothesis.

Hypothesis: Students who didn't perform well in their previous education and didn't receive good grades in semester one is likely to get a low score in semester two.

In the previous analysis, how student learner profiles might influence success in higher education were primarily examined. Acknowledging the significant role of grades, the impact of prior qualifications and first-year grades on second-semester performance were assessed. This approach aims to determine the predictive value of these factors for final grades in higher education. By segmenting the analysis into these components, the objective was to evaluate if early predictions regarding student success or potential dropout could be made based on previous grades. Such insights could enable timely interventions from the second year onwards to mitigate dropout rates and enhance graduation rates.

The analysis aimed to determine the predictive value of students' previous educational performance and first-year academic outcomes on their final grade in higher education. The study was divided into two parts: examining the impact of previous qualifications and year one grades on semester two outcomes, and subsequently using these factors to predict final grades.

1. Impact of Grades: Semester two grades, particularly, emerged as a critical factor across all classes. High performance in the second semester strongly correlated with the likelihood of graduation, while low performance was a key indicator of dropout risk.

2. Mixed Effects of Previous Qualifications: The inclusion of previous qualifications and first-semester results provided some predictive power but did not significantly enhance the model's ability to accurately forecast final outcomes. This indicates that while early academic performance is important, it alone is insufficient for reliable long-term predictions.

3. Need for Additional Data: The relatively moderate accuracy and performance variability across classes suggest that other factors beyond grades and qualifications are likely influencing student outcomes. These may include personal circumstances, engagement levels, support systems, and other non-academic factors.

Given these findings, early identification of at-risk students based solely on grades may not be sufficient. A more comprehensive approach that includes additional data sources—such as student engagement metrics, socio-economic background, and psychological factors—might be necessary to improve prediction accuracy. This would allow institutions to intervene earlier and more effectively, potentially reducing dropout rates and increasing graduation rates.

Overall, while academic performance in the first year provides some insight into student outcomes, the model's limitations highlight the need for a more holistic approach to predicting and supporting student success in higher education.

XII. CONCLUSION

This project successfully employed multiple machine learning models—Logistic Regression, XGBoost, Random Forest Classifier, Decision Tree, and Linear Regression—to classify and predict student success in higher education. The results demonstrated the efficacy of these models in accurately forecasting student outcomes, thereby underscoring the potential of machine learning as a valuable tool in educational settings.

The insights gained from this research provide a solid foundation for the future integration of machine learning techniques in academic institutions. By harnessing these predictive models, educators and administrators can proactively identify at-risk students, tailor interventions, and ultimately enhance the overall educational experience. The findings also open avenues for further research to refine these models and explore their application across diverse educational contexts.

In conclusion, this study contributes meaningfully to the growing body of knowledge on the intersection of education and machine learning, paving the way for data-driven approaches to improving student success in higher education.

XIII. ACKNOWLEDGEMENTS

Many thanks to Chathura Kalpanee Sooriya Arachchi & Huan Zhang for their invaluable guidance and expertise throughout the project. Their insights were crucial to the

success of this project, and it would not have been possible without them.

REFERENCES

[1] Dropping out of college and dropping into crime by Christopher R. Dennison
[2] Seven in ten students consider dropping out – How can universities fix this gloomy statistic? By Leo Hanna
[3] Factors Influencing Dropout Students in Higher Education by Nurmalitasari Nurmalitasari, Zalizah Awang Long and Faizuddin Mohd Noor
[4] Early Prediction of student's Performance in Higher Education: A Case Study by Mónica V. Martins, Daniel Tolledo, Jorge Machado, Luís M. T. Baptista and Valentim Realinho
[5] Employing Bayesian Belief Networks for Predicting Student Performance by Gil, Martins, Moro, and Costa (2020)
[6] Analysing and Predicting Students' Performance by Means of Machine Learning: A Review" by Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido, and Arturo Durán-Domínguez (2020)
[7] Machine Learning Based Predicting Student Academic Success by Al-Bahri Mahmood (2020)
[8] Comparing Different Oversampling Methods in Predicting MultiClass Educational Datasets Using Machine Learning Techniques by Muhammad Arham Tariq, Allah Bux Sargano, Muhammad Aksam Iftikhar and Zulfiqar Habib2
[9] Enhancing Prediction of Student Success: Automated Machine Learning Approach Hassan ZeineddineHassan, Udo C. Braendle and Assaad Farah