Addressing Class Imbalance in Predicting Student Academic Outcomes: A Comparative Study of Resampling Techniques with Machine Learning Classifiers in Higher Education

Tri Rochmadi[1,*], Anantian Mahendra Tirta Saputra[2]

[1]Information System Department, Universitas Alma Ata, Indonesia

ABSTRACT

The increasing availability of educational data presents a significant opportunity for higher education institutions to proactively identify and support students at risk of academic failure or dropout. However, datasets in this domain are often characterized by a severe class imbalance, where successful students vastly outnumber those who drop out or struggle, posing a substantial challenge for standard predictive modeling techniques. This study addresses this issue by conducting a comprehensive, comparative analysis of machine learning classifiers and data resampling techniques to accurately predict student academic outcomes—categorized as Graduate, Dropout, or Enrolled. Using a dataset of 4,424 undergraduate students from a Portuguese institution, we evaluate six distinct classifiers, including Logistic Regression, Random Forest, and Support Vector Machines. The models are first trained on the original, imbalanced data to establish a performance baseline, which reveals a significant weakness in identifying the minority 'Enrolled' class. Subsequently, we implement a suite of oversampling, undersampling, and hybrid resampling techniques, such as SMOTE, ADASYN, and RandomUnderSampler, to balance the training data. The results demonstrate that data resampling, particularly oversampling, provides a significant performance improvement across all models. The combination of a Random Forest classifier with the ADASYN technique emerged as the most effective approach, achieving the highest macro-averaged F1-score of 0.7081. Crucially, this method substantially improved the model's ability to correctly classify the underrepresented 'Enrolled' students. This research validates a robust methodology for handling imbalanced data in educational analytics and underscores the necessity of such techniques for building fair and effective early-warning systems. The findings provide a clear pathway for institutions to leverage AI for more equitable and targeted student support, ultimately fostering higher retention and success rates.

Keywords Class Imbalance, Educational Data Mining, Learning Analytics, Predictive Modeling, Student Retention

Submitted 2 July 2025
Accepted 29 July 2025
Published 1 September 2025

*Corresponding author
Tri Rochmadi,
trirochmadi@almaata.ac.id

Additional Information and Declarations can be found on page 110

DOI: 10.63913/ail.v1i3.32

Introduction

In today's economy, the significance of higher education is multifaceted, serving not only as a critical pathway for individual advancement but also contributing significantly to societal progress. Higher education equips students with essential knowledge and skills while fostering a culture of lifelong learning and innovation. Student success in this arena is paramount, yet dropout rates remain alarmingly high. The problem of student attrition poses significant challenges for educational institutions, impacting their reputation and sustainability, and ultimately leading to economic repercussions for society at large, including lost

How to cite this article: T. Rochmadi and A. M. T. Saputra, "Addressing Class Imbalance in Predicting Student Academic Outcomes: A Comparative Study of Resampling Techniques with Machine Learning Classifiers in Higher Education," *Artif. Intell. Learn.*, vol. 1, no. 3, pp. 211-227, 2025.

productivity and increased educational costs due to incomplete programs [1].

The issue of student dropout and academic failure is compounded by various factors, including socioeconomic status, academic support, and mental health issues. High dropout rates can predispose students to long-term socioeconomic disadvantages, perpetuating cycles of inequality and affecting their overall well-being [2]. As such, effective interventions aimed at improving student retention and success are critically required. However, accurately predicting student academic outcomes—such as dropout, enrollment status, and graduation—presents significant challenges. This complexity is further exacerbated by class imbalance within datasets used for prediction, where the number of samples from minority classes—such as students who drop out—often underrepresents the majority classes, negatively impacting model performance [3], [4].

The challenge of accurately predicting student academic outcomes revolves around the limitations posed by class imbalance, which can lead to biased and less effective predictive models. In many instances, predictive models trained on imbalanced datasets tend to ignore the minority class, which often consists of students who are at risk of dropping out or failing. This oversight results in high false-negative rates, meaning that many students who require intervention may be overlooked, ultimately leading to higher dropout rates [5]. Citing the need for developing robust prediction models, research [6] emphasize the necessity for accurate early-warning systems that can effectively identify at-risk students, thereby facilitating timely interventions to increase retention.

The pressing need for early identification of at-risk students cannot be overstated. With timely and appropriate intervention, educational institutions can significantly enhance student success and retention rates. Utilizing machine learning (ML) techniques represents a promising avenue for improving prediction accuracy, allowing for the processing of large and complex datasets more effectively than traditional statistics [7], [8]. As highlighted by Amare and Šimonová [4], the integration of ML into educational analytics can lead to the development of predictive models that consider a wide variety of factors influencing student success.

Moreover, to address the challenges of class imbalance, various resampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) have been proposed. These methods augment the minority class by generating synthetic samples, thus creating a more balanced dataset for training predictive models. This approach has shown promise in enhancing the predictive performance of models on minority classes while also improving overall model robustness [9], [10]. By harnessing these ML frameworks and adaptively mitigating class imbalance, educational institutions can foster environments that significantly curtail dropout rates and promote academic success.

The significance of addressing student dropout rates can profoundly impact individual lives and shape broader societal outcomes. The challenges associated with predicting academic success, particularly the class imbalance in datasets, necessitate innovative approaches such as machine learning combined with effective resampling techniques. The integration of these methodologies into educational frameworks is essential for improving student outcomes, thereby enhancing retention rates and fostering a more equitable educational landscape.

The primary aim of this research is to investigate and systematically compare

the effectiveness of various data resampling techniques in improving the performance of machine learning classifiers for predicting student academic outcomes. This study specifically addresses the common challenge of class-imbalanced datasets in higher education, where the number of successful students often far exceeds those who are at-risk. The central goal is to identify a robust modeling strategy that can accurately and equitably classify students into 'Graduate', 'Dropout', and 'Enrolled' categories, thereby enhancing the potential for effective and timely institutional intervention.

To achieve this aim, the study is structured around several key objectives. The first objective is to thoroughly preprocess and prepare the student dataset for machine learning modeling, which includes handling categorical data through one-hot encoding and scaling numerical features. The second is to establish a clear performance baseline by training and evaluating a selection of machine learning classifiers on the original, imbalanced dataset. Following this, the core objectives are to apply a range of oversampling, undersampling, and hybrid resampling techniques to the training data, and subsequently train and evaluate the classifiers on these newly balanced datasets. The final objectives are to rigorously compare the performance of models trained with and without resampling, using evaluation metrics appropriate for imbalanced data, and to identify which resampling techniques are most effective for improving prediction accuracy for each student outcome category, especially the underrepresented ones.

This research is guided by three central questions designed to probe the core of the problem. First, how does the inherent class imbalance within the dataset quantitatively affect the performance of standard machine learning classifiers in predicting student outcomes? Second, which specific resampling techniques—whether oversampling, undersampling, or hybrid approaches—lead to a significant improvement in the predictive performance for the minority classes ('Enrolled' and 'Dropout')? Finally, what is the overall impact of these techniques on the classifiers' ability to reliably distinguish between the three distinct student outcomes, and what combination of model and technique yields the most balanced and accurate results?

The scope of this investigation is intentionally delimited to ensure a focused and rigorous analysis. The study is conducted on a dataset from a single Portuguese higher education institution, which includes students from a specific set of degree programs. The analysis is confined to a selection of six widely-used machine learning classifiers and seven distinct resampling techniques. Consequently, while the findings provide deep insights and a robust methodological template, their direct generalizability is constrained to contexts with similar institutional and data characteristics. This defined scope allows for a detailed and controlled comparison, providing a strong foundation for future, broader research in the field.

Literature Review

Student Success and Dropout Prediction

Student academic success and dropout rates are influenced by a myriad of factors, including socio-economic background, academic performance, familial

support, and psychological well-being. Research has shown that students with low academic performance, high levels of stress, or lack of familial support are more likely to drop out [11], [12]. Srairi [11] identifies that socio-cultural status and family education levels are critical predictors of dropout rates, linking economic stability to educational attainment and, consequently, student retention.

Machine learning has increasingly been utilized to address dropout prediction by identifying at-risk students through various predictive modeling techniques. Studies have highlighted the effectiveness of diverse algorithms, such as decision trees, support vector machines (SVM), and neural networks. For instance, Silva and Roman [13] conducted a systematic review and found that decision trees are frequently employed due to their interpretability and efficiency in handling diverse datasets. Further research by Rincón et al [14] focused on rural education contexts, underscoring the applicability of machine learning in predicting dropout in virtual settings.

Predictive models for student dropout often leverage a combination of academic performance metrics, demographic data, and behavioral indicators. Common features include GPA, attendance rates, social integration, mental health status, and external commitments, such as part-time employment [15], [16]. When it comes to algorithms, logistic regression remains popular for its simplicity, while ensemble methods such as Random Forests are favored for their robustness against overfitting and their ability to handle large datasets [4]. Moreover, Zheng et al [17] introduced a FWTS-CNN model to integrate time-series data with dropout prediction attributes, showcasing the evolution and diversification within these predictive methodologies.

Machine Learning in Education

The broader context of AI and ML in educational settings is profoundly transformative, impacting teaching methods, curriculum design, and administrative efficiency. Educational institutions are increasingly employing AI-driven tools for personalized learning experiences, aiming to enhance student engagement and retention. Nonetheless, as AIED adoption grows, ethical considerations surrounding data privacy, student surveillance, and algorithmic bias become increasingly pertinent. Transparent model interpretations and equitable access to educational resources are necessary to mitigate potential harms [18], [19].

Class imbalance is defined as a scenario where the number of instances in one class is significantly different from the others within a classification problem. This poses a significant challenge in dropout prediction contexts, where students who drop out constitute a minority class compared to those who complete their education [4]. Class imbalance can lead to biased predictions, where models may inaccurately predict the majority class, resulting in high false negatives for the minority class, which is particularly detrimental in educational contexts where timely intervention is crucial [9].

Techniques for Handling Imbalanced Data

Effective methods to manage the common challenge of class imbalance in machine learning can be broadly categorized into two primary approaches: data-level and algorithmic-level techniques. Data-level strategies focus on modifying the dataset itself to create a more balanced distribution before training a model.

These methods, collectively known as resampling, directly alter the composition of the training data. In contrast, algorithmic-level techniques do not change the data but instead modify the learning algorithms to make them more robust to the imbalanced distribution, often by imposing a higher cost for misclassifying the minority class. This study focuses primarily on data-level resampling due to its wide applicability and model-agnostic nature.

Oversampling represents a popular data-level approach where the goal is to increase the representation of the minority class. Techniques like Random Oversampling achieve this by simply duplicating existing instances, while more advanced methods such as SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN generate new, synthetic samples by interpolating between existing minority class instances. The primary advantage of oversampling is that it provides the model with more data to learn the decision boundary of the underrepresented class, often leading to improved recall and sensitivity. However, this approach is not without significant drawbacks. A major concern is the risk of overfitting, where the model learns the specific synthetic patterns too well and fails to generalize to new, unseen data. Furthermore, generating additional data points substantially increases the size of the training set, which in turn raises the computational cost and time required for model training [19].

On the other end of the spectrum, undersampling techniques aim to balance the dataset by reducing the number of instances in the majority class. Methods like RandomUnderSampler randomly discard majority class samples, while more sophisticated approaches like NearMiss or Tomek Links selectively remove instances that are considered redundant or noisy, often those near the class boundary. The main advantages of this approach are a significant reduction in computational load and a lower risk of overfitting compared to oversampling. The primary and often critical disadvantage, however, is the potential loss of valuable information. By discarding data from the majority class, there is a substantial risk of removing instances that are crucial for defining the decision boundary, which can lead to a poorly generalized model that underperforms on real-world data.

To leverage the benefits of both approaches while mitigating their weaknesses, hybrid methods have been developed. Techniques such as SMOTE-ENN and SMOTE-Tomek execute a two-step process: they first use an oversampling method like SMOTE to generate synthetic minority samples and then apply an undersampling or data cleaning technique (like Edited Nearest Neighbors or Tomek Links) to remove noisy or overlapping instances from both classes. This combined strategy aims to create a more balanced and "cleaner" dataset with better-defined class separation. Beyond these data-level manipulations, algorithmic-level approaches offer an alternative. These include cost-sensitive learning, where higher misclassification costs are assigned to the minority class, forcing the algorithm to pay more attention to it, and specialized ensemble methods that are adapted to give more weight to the predictions for minority class instances [20].

Evaluation Metrics for Imbalanced Classification

In classification scenarios marked by a significant class imbalance, standard accuracy becomes a misleading metric, as it can be artificially inflated by a model that simply predicts the majority class. This necessitates the use of

evaluation metrics specifically designed to provide a more nuanced and reliable assessment of performance. Among the most critical are Precision, which measures the proportion of true positive predictions among all positive predictions, and Recall (or sensitivity), which measures the proportion of actual positive cases that were correctly identified. In the context of this study, high recall is vital for ensuring that the model effectively identifies the maximum number of at-risk students, while precision ensures that the students flagged for intervention are indeed the ones who need it, minimizing the misallocation of institutional resources.

To achieve a holistic view of a model's effectiveness, it is essential to consider metrics that synthesize both precision and recall. The F1-Score, calculated as the harmonic mean of precision and recall, provides a single, robust measure that balances this trade-off. Similarly, the Geometric Mean (G-Mean) measures the balance between the sensitivity (recall of the positive class) and specificity (recall of the negative class), offering another way to evaluate a model's ability to perform well on both majority and minority classes. These balanced metrics are crucial for selecting a model that is not only accurate but also fair and equitable in its predictions. The formulas for these metrics are as follows, where TP, FP, TN, and FN represent True Positives, False Positives, True Negatives, and False Negatives, respectively:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Further insight into model performance can be gained by analyzing its behavior across a range of decision thresholds. The Area Under the Precision-Recall Curve (PR-AUC) and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) are powerful tools for this purpose. The PR-AUC is particularly informative in severely imbalanced scenarios, as it effectively summarizes a model's performance in distinguishing the minority class. The ROC-AUC, meanwhile, visualizes the trade-off between the true positive rate and the false positive rate. As noted by [9], these curve-based metrics provide a more comprehensive performance summary than single-point metrics, making them indispensable for a thorough comparative analysis.

Despite the robust body of research on both predictive modeling in education and imbalanced learning, a significant gap remains in the comparative analysis of resampling techniques specifically applied to multi-class educational datasets. Many existing studies focus on a simplified binary classification of student outcomes (e.g., dropout vs. non-dropout), which overlooks the critical, intermediate 'Enrolled' status of students who may be at risk but have not yet left the institution [21]. This research contributes to the field by directly addressing this three-class problem. By systematically evaluating the effectiveness of various resampling methods on this more complex and realistic

classification task, this study aims to provide a clearer understanding of which strategies are best suited for enhancing predictive accuracy and, ultimately, for developing more effective and nuanced student intervention strategies.

Method

This study employs a systematic, comparative approach to evaluate the effectiveness of various machine learning models and data resampling techniques for predicting student academic outcomes. The objective is to identify a robust modeling strategy that can accurately classify students who are likely to graduate, drop out, or remain enrolled, with a particular focus on mitigating the challenges posed by an imbalanced class distribution. The entire experimental pipeline, from data preprocessing to model evaluation, was implemented using the Python programming language, leveraging its extensive ecosystem of scientific computing and machine learning libraries.

Dataset Description

The dataset for this research originates from a single Portuguese higher education institution, providing a consistent and controlled data environment. It contains records for 4,424 undergraduate students across multiple degree programs. The dataset is comprised of 37 distinct features that provide a holistic view of each student's profile, capturing a wide range of potentially predictive information. The primary outcomeMarahwa, or target variable, is the student's academic status, categorized as 'Graduate', 'Dropout', or 'Enrolled'. The dataset exhibits a natural class imbalance, with approximately 50% of students labeled as 'Graduate', 32% as 'Dropout', and only 18% as 'Enrolled'. This skewed distribution presents a significant modeling challenge, as standard algorithms may develop a bias towards the majority class and fail to learn the patterns of the underrepresented 'Enrolled' category.

The features can be broadly grouped into several key areas. Demographic information includes attributes such as 'Marital Status', 'Gender', and 'Age at enrollment'. Pre-university academic background is captured by features like 'Previous qualification' and 'Previous qualification (grade)'. A number of socio-economic indicators, such as 'Scholarship holder', 'Debtor' status, and macroeconomic variables like 'GDP' and 'Unemployment rate', provide context about the student's financial and environmental circumstances. Finally, a rich set of course-related academic information details student performance and engagement during their first two semesters, with features like 'Curricular units 1st sem (approved)', 'Curricular units 1st sem (grade)', and their second-semester equivalents. These first-year performance metrics are often considered highly indicative of a student's long-term academic trajectory.

Data Preprocessing

Before model training, a multi-step preprocessing phase was conducted to prepare the data for the machine learning algorithms. All features identified as categorical, even those numerically encoded in the original dataset (such as 'Course' or 'Application mode'), were transformed using one-hot encoding. This strategy was chosen because these features are nominal in nature, meaning there is no inherent order to their values. One-hot encoding creates new binary columns for each category, preventing the models from inferring a false and misleading ordinal relationship that other methods, like label encoding, might introduce.

Concurrently, all numerical features were scaled using Standardization via scikit-learn's StandardScaler. This process transforms the data to have a mean of zero and a standard deviation of one. This step is critical for the optimal performance of many algorithms, particularly linear models and Support Vector Machines, which are sensitive to the scale of input features. Standardization ensures that features with larger scales (like admission grades) do not disproportionately influence the model's learning process over features with smaller scales (like GDP). Following transformation, the dataset was split into a training set (75%) and a testing set (25%). A stratified splitting strategy was employed to ensure that the class distribution of the target variable was precisely maintained in both partitions. This is crucial for obtaining reliable evaluation results, as a simple random split could result in a test set with a non-representative proportion of the minority classes, leading to misleading performance metrics.

Machine Learning Classifiers

A diverse set of six supervised learning algorithms was selected to cover different modeling paradigms and complexities. The chosen classifiers include Logistic Regression, selected as a robust and interpretable linear baseline model; a Decision Tree, to capture non-linear relationships in a highly transparent manner; Random Forest and Gradient Boosting, two powerful ensemble methods known for their high predictive accuracy and ability to mitigate overfitting; K-Nearest Neighbors, an instance-based learner that classifies based on feature similarity; and Support Vector Machines (using a radial basis function kernel), a powerful model effective in high-dimensional spaces. These models were selected due to their common application and proven effectiveness in a wide range of classification tasks. For this comparative study, the models were implemented with their default or standard hyperparameters from the scikit-learn library. This approach was taken to establish a consistent and fair baseline performance across all experimental conditions, ensuring that any observed performance differences could be attributed directly to the choice of model architecture and the impact of the resampling techniques, rather than to hyperparameter optimization.

Resampling Techniques

To directly address the challenge of class imbalance, a comprehensive suite of resampling techniques was implemented using the imbalanced-learn Python library. These techniques fall into three main categories. First, oversampling methods, including SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling), were used to create new, synthetic instances of the minority classes ('Enrolled' and 'Dropout') to provide the models with more data to learn from. Second, undersampling methods, such as RandomUnderSampler and NearMiss, were employed to reduce the number of instances in the majority 'Graduate' class, thereby preventing it from dominating the training process. Finally, hybrid methods like SMOTEENN and SMOTETomek were implemented, which strategically combine both over- and undersampling to clean the feature space and create more well-defined class boundaries. Crucially, these techniques were integrated into the training pipeline and were applied only to the training data. The test set remained in its original, imbalanced state to serve as a realistic and unbiased benchmark for evaluating how well the models would perform on real-world data.

Experimental Setup and Evaluation

The experiment was structured to first establish a crucial baseline by training and evaluating all six classifiers on the original, imbalanced training data. This baseline provides a reference point to quantify the impact of the balancing strategies. Subsequently, for each of the six resampling techniques, the training data was transformed, and all classifiers were retrained on this newly balanced data. The performance of every model from every experiment was then measured against the single, untouched test set.

To robustly evaluate performance on this imbalanced, multi-class problem, a set of appropriate metrics was chosen. The macro-averaged F1-score was selected as the primary metric because it calculates the F1-score for each class independently and then averages them, giving equal weight to each class regardless of its size. This prevents the majority class from inflating the overall score. Balanced accuracy was also used, as it avoids misleading results on imbalanced data by averaging the recall obtained on each class. The One-vs-Rest (OvR) ROC AUC score provided a measure of the model's ability to discriminate between classes. Furthermore, the F1-score for each individual class was calculated to assess how well the models identified 'Dropout', 'Enrolled', and 'Graduate' students specifically. Finally, confusion matrices were generated for each experiment to allow for a granular, visual inspection of the classification results, revealing the specific types of errors each model was making.

Result and Discussion

Initial Exploratory Data Analysis Results

An initial Exploratory Data Analysis (EDA) was conducted to understand the underlying distributions and relationships within the dataset. The analysis confirms the significant class imbalance that motivates this study. As shown in the outcome distribution chart (Figure 1), approximately half of the students in the dataset (49.9%) eventually graduate, while a substantial portion either drop out (32.1%) or remain enrolled (17.9%). This skewed distribution highlights the challenge for predictive modeling, as standard algorithms may develop a bias towards the majority 'Graduate' class.

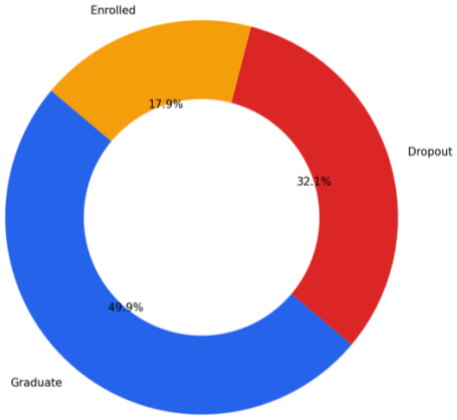

Figure 1 Outcome Distribution Chart

Several key features show a strong relationship with student outcomes. Financial status, in particular, appears to be a critical factor. There is a stark

contrast in outcomes based on whether a student's tuition fees are up to date, as shown in Figure 2. For students with paid tuition, the dropout rate is 24.7%; however, for students with outstanding fees, this rate skyrockets to 86.6%, indicating a powerful predictive signal. Similarly, scholarship status is strongly correlated with success, as shown in Figure 3. Non-scholarship holders have a dropout rate of 38.7%, more than triple the 12.2% rate observed for students who receive scholarships. Demographically, the student population is predominantly of traditional college age (18-21 years). The age distribution for students who graduate is tightly concentrated in this younger range, whereas the distributions for 'Dropout' and 'Enrolled' students have longer tails, suggesting that mature students may face additional challenges that impact their academic trajectory.

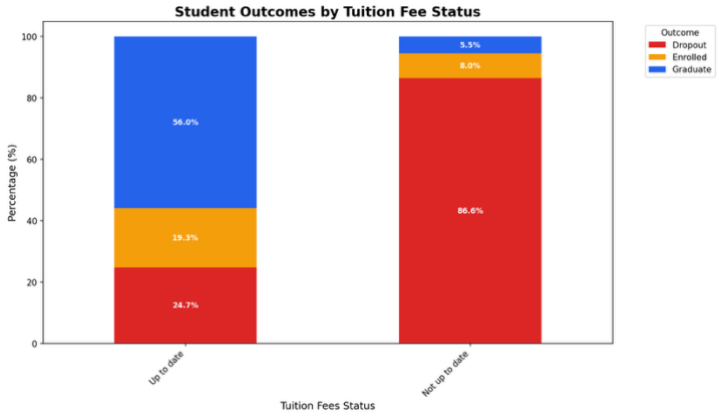

Figure 2 Student Outcomes by Tuition Fee Status

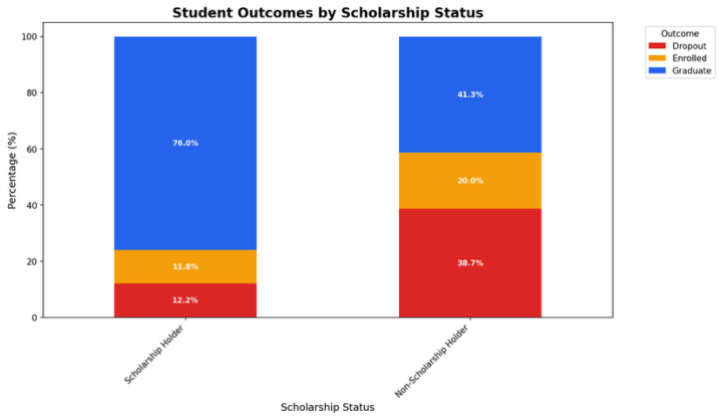

Figure 3 Student Outcomes by Scholarship Status

Unsurprisingly, prior academic performance is a highly significant indicator of future success. The box plots for first and second-semester grades reveal a clear and consistent pattern, as shown in Figure 4. Students who ultimately graduate consistently achieve the highest median grades and exhibit the least amount of variance in their performance. Conversely, students who drop out display the lowest median grades and the widest grade distribution, with a significant number of students having grades at or near zero, indicating complete disengagement or failure in their courses. The 'Enrolled' cohort consistently falls between these two extremes. This strong separation based on early academic performance underscores its importance as a primary feature

for predictive modeling.

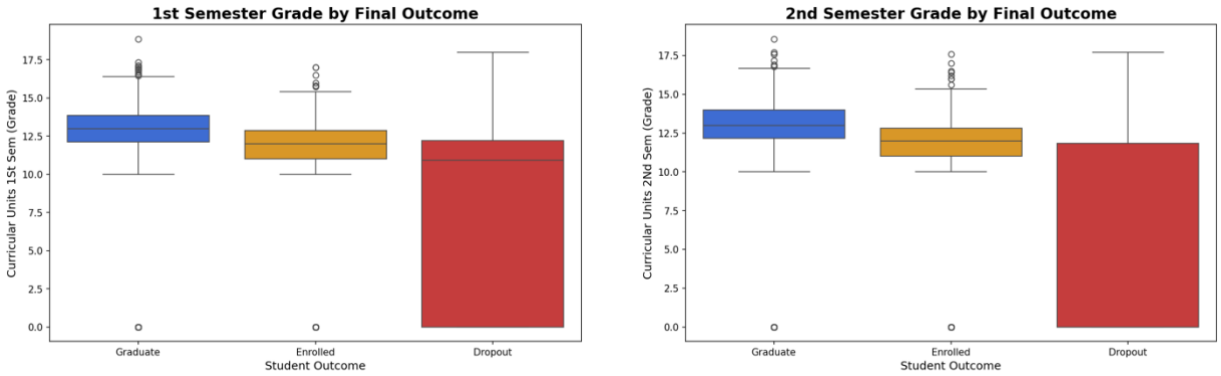

Figure 4 Box Plot of Final Outcome in 1st and 2nd Grade

Finally, the analysis of relationships between numerical variables reveals several expected patterns. The correlation heatmap (Figure 5) shows a very strong positive correlation between first and second-semester academic metrics, such as the number of approved curricular units and the grades achieved in those semesters. This indicates that student performance is generally consistent from one semester to the next. Interestingly, macroeconomic indicators like GDP show a much weaker relationship with student outcomes, with the distributions across the three outcome categories being nearly identical. This suggests that while individual financial circumstances (like tuition status and scholarships) are highly predictive, broader economic conditions at the time of enrollment may have a less direct impact on a student's academic fate.

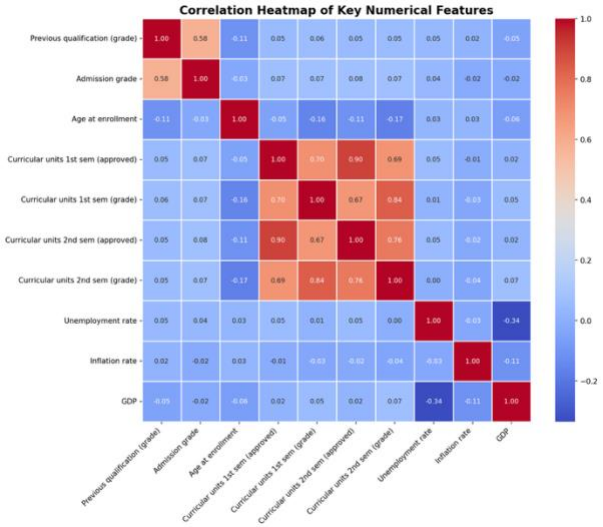

Figure 5 Correlation Heatmap of Key Numerical Features

Baseline Model Performance

The initial experiment involved training and evaluating the six selected classifiers on the original, imbalanced dataset to establish a performance baseline against which all subsequent experiments could be measured. The results, summarized by the macro-averaged F1-score, are presented in the

table below. On the raw data, the Support Vector Machine (SVM) and Gradient Boosting models demonstrated the strongest performance, achieving F1-scores of 0.6906 and 0.6850, respectively, indicating a superior ability to balance precision and recall across the classes. In contrast, the K-Nearest Neighbors and standard Decision Tree models were the weakest performers, likely due to the Decision Tree's tendency to overfit to the majority class and KNN's sensitivity to the skewed distribution in the feature space.

A critical and consistent observation from this baseline phase was the models' significant difficulty in correctly identifying the minority 'Enrolled' class. This is the most challenging yet arguably most important group for early intervention. For instance, the top-performing baseline SVM model achieved a respectable F1-score of 0.854 for the majority 'Graduate' class but only managed an F1-score of 0.450 for the 'Enrolled' class. This disparity of over 0.40 highlights the pronounced effect of the class imbalance on predictive capability. The models learned to accurately identify graduates but struggled to distinguish the nuanced patterns of students who were still enrolled but potentially at risk, often misclassifying them as either graduates or dropouts.

Performance of Models with Resampling Techniques

Following the baseline evaluation, each of the six resampling techniques was applied to the training data before retraining the classifiers. The primary goal was to determine if balancing the class distribution could improve overall performance, particularly for the minority classes. The results clearly and consistently indicate that resampling, especially oversampling, provides a substantial and meaningful benefit to model performance.

The top-performing combination was the Random Forest classifier paired with the ADASYN oversampling technique, which achieved the highest macro-averaged F1-score of 0.7081. This represents a notable improvement over the best baseline model and demonstrates the synergy between a powerful ensemble method and an advanced resampling strategy. Several other combinations using oversampling also outperformed the baseline, including Logistic Regression with SMOTE (F1 Macro: 0.7024) and Gradient Boosting with SMOTE (F1 Macro: 0.7009).

A key finding is that these top-performing models demonstrated a marked improvement in identifying the 'Enrolled' class. The F1-score for this category increased from a baseline high of 0.450 to as high as 0.519 in the case of Logistic Regression with SMOTE—an improvement of nearly 15%. This shows that balancing the training data directly translated into a better ability to recognize at-risk students. In contrast, undersampling techniques showed mixed results; while RandomUnderSampler provided a modest boost for some models, the NearMiss algorithm consistently degraded the performance of all classifiers. This suggests that the removal of majority-class data, even when done strategically, was detrimental in this context, likely because it discarded valuable information near the complex decision boundaries between the classes.

The experimental results lead to several clear and actionable conclusions. First, applying resampling techniques to the imbalanced training data consistently improved the predictive performance of the machine learning models, as measured by the macro-averaged F1-score. Second, oversampling techniques (ADASYN and SMOTE) proved to be the most effective strategies, significantly

outperforming both the baseline and all undersampling methods across nearly every classifier. Third, the more complex, non-linear models like Random Forest and Gradient Boosting were among the classifiers that benefited most from the balanced data, ultimately achieving the highest performance scores. Most importantly, the application of resampling successfully and significantly enhanced the models' ability to predict the minority classes. This is a critical step toward building a fair, equitable, and effective student intervention system that does not overlook the students who need the most support.

Interpretation of Results

The baseline performance clearly illustrates the classic challenge of imbalanced classification in a real-world setting. The models performed reasonably well on the 'Graduate' class due to its statistical prevalence but struggled to learn the defining characteristics of the much smaller 'Enrolled' class, leading to poor recall and precision for that group. This is a direct consequence of the models' optimization functions, which, without intervention, are driven to minimize overall error. This is most easily achieved by correctly classifying the majority class, a strategy that unfortunately leads to the neglect of minority classes and results in a model with limited practical utility for intervention.

The pronounced success of oversampling techniques like SMOTE and ADASYN can be attributed to their mechanism of generating synthetic data points for the minority classes. By creating new, plausible examples of 'Enrolled' and 'Dropout' students within the existing feature space, these methods provide the classifiers with a richer and more balanced set of information from which to learn the decision boundaries. ADASYN's superior performance when paired with the Random Forest model may stem from its adaptive nature; it generates more synthetic data for minority examples that are harder to learn (i.e., those near the class boundaries), effectively forcing the model to pay closer attention to the more complex and ambiguous cases. Conversely, the consistently poor performance of NearMiss likely resulted from the removal of potentially valuable majority-class instances that were close to the decision boundary. This inadvertently discards critical information that helps define the separation between classes, leading to a more confused and less accurate model. This suggests that for this particular dataset, the information contained within the majority class was too valuable to discard.

Implications of the Findings

The findings of this study have significant practical and methodological implications. For educational institutions, the primary takeaway is that machine learning models, when carefully designed to handle class imbalance, can serve as powerful and reliable tools for early student intervention. A model like the Random Forest with ADASYN, which achieves a balanced F1-score of over 0.70, can be deployed to flag students who are at risk of dropping out or are struggling (i.e., 'Enrolled' but not progressing) long before they fail a course or formally withdraw. This allows student support services, academic advisors, and faculty to proactively offer targeted resources, such as academic advising, mental health counseling, tutoring, or financial aid counseling, thereby improving student well-being and institutional retention rates. The ability to more accurately identify the 'Enrolled' but at-risk group is particularly valuable, as these are the students who can most benefit from timely support.

From a methodological perspective, this study reinforces the critical importance

of addressing class imbalance in educational data mining and learning analytics. It demonstrates that simply applying a powerful classifier to a raw, imbalanced dataset is an insufficient and potentially inequitable approach. A principled methodology that includes data balancing is necessary for building fair and effective predictive models that serve all student populations. The comparative framework presented here serves as a valuable blueprint for other researchers and practitioners working with similarly skewed educational datasets, providing a clear path for model development and evaluation.

Limitations of the Study

This study, while comprehensive in its comparison of techniques, has several limitations that should be acknowledged to contextualize the findings. First, the dataset originates from a single institution in Portugal. Therefore, the generalizability of the findings to other institutions, particularly those in different countries with different cultural contexts, educational systems, and student demographics, is not guaranteed. The specific features that predict success or failure may vary significantly across different institutional settings. Second, the models were trained using their default hyperparameters to ensure a fair baseline comparison. A rigorous hyperparameter tuning process (e.g., using GridSearchCV or RandomizedSearchCV) could potentially yield further performance improvements for all models and might even alter the ranking of the best-performing combinations. Finally, this study did not explore feature engineering, which could uncover more complex, interaction-based relationships within the data and further boost predictive accuracy.

Future Research Directions

Building on the findings and limitations of this work, several promising avenues for future research are apparent. An immediate next step would be to perform hyperparameter optimization on the top-performing model-resampler combinations (such as Random Forest with ADASYN) to maximize their predictive power and establish a new state-of-the-art performance on this dataset. Furthermore, an in-depth feature importance analysis using techniques like SHAP (SHapley Additive exPlanations) could be conducted to understand which student attributes are the most critical predictors of academic outcomes. This would provide more interpretable and actionable insights for educators and administrators. Future work could also explore more advanced deep learning architectures or algorithmic approaches to handling imbalance, such as cost-sensitive learning, which penalizes the model more for misclassifying minority class instances. Finally, replicating this study with datasets from a diverse range of institutions would be a crucial step toward developing more robust, generalizable, and equitable models for predicting student success on a global scale.

Conclusion

This study systematically demonstrated that while standard machine learning models struggle to accurately predict outcomes on an imbalanced student dataset, their performance can be significantly enhanced through principled data resampling. The key finding is that oversampling techniques, particularly ADASYN and SMOTE, were highly effective at mitigating the bias towards the majority 'Graduate' class, leading to a more balanced and accurate predictive system. The combination of a Random Forest classifier with ADASYN emerged as the most robust approach, most notably improving the identification of the

minority 'Enrolled' student population. The primary contribution of this research is a clear, empirical validation of resampling strategies in an educational context, providing a methodological blueprint for developing more equitable predictive models. By showing how to overcome the common challenge of class imbalance, this work enables the creation of more reliable tools for learning analytics.

Ultimately, this research underscores the critical importance of addressing data imbalances to build fair and effective predictive systems in education. Failing to do so results in models that systematically overlook the very students who are most in need of support, thereby limiting the practical utility of any early-warning system. The successful application of these techniques affirms the immense potential of artificial intelligence to act as a proactive tool for fostering student success. By leveraging these advanced analytical methods, educational institutions can move beyond reactive measures and create more supportive, data-informed environments where every student has a better opportunity to thrive.

Declarations

Author Contributions

Conceptualization: T.R.; Methodology: A.M.T.S.; Software: A.M.T.S.; Validation: A.M.T.S.; Formal Analysis: T.R.; Investigation: A.M.T.S.; Resources: T.R.; Data Curation: A.M.T.S.; Writing Original Draft Preparation: T.R.; Writing Review and Editing: T.R.; Visualization: T.R.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Huo *et al.*, “Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach,” *J. Coll. Stud. Retent. Res. Theory Pract.*, 2020, doi: 10.1177/1521025120963821.

[2] C. Herodotou, B. Rienties, A. Boroowa, Z. Zdráhal, and M. Hlosta, “A Large-Scale Implementation of Predictive Learning Analytics in Higher Education: The Teachers’ Role and Perspective,” *Educ. Technol. Res. Dev.*, 2019, doi:

10.1007/s11423-019-09685-0.
[3] S. Shohag and M. Bakaul, “A Machine Learning Approach to Detect Student Dropout at University,” *Int. J. Adv. Trends Comput. Sci. Eng.*, 2021, doi: 10.30534/ijatcse/2021/041062021.
[4] M. Y. Amare and S. Šimonová, “Global Challenges of Students Dropout: A Prediction Model Development Using Machine Learning Algorithms on Higher Education Datasets,” *SHS Web Conf.*, 2021, doi: 10.1051/shsconf/202112909001.
[5] K. K. Patel and K. Amin, “Predictive Modeling of Dropout in MOOCs Using Machine Learning Techniques,” *Sci. Temper*, 2024, doi: 10.58414/scientifictemper.2024.15.2.32.
[6] G. Akçapınar, A. Altun, and P. Aşkar, “Using Learning Analytics to Develop Early-Warning System for at-Risk Students,” *Int. J. Educ. Technol. High. Educ.*, 2019, doi: 10.1186/s41239-019-0172-z.
[7] L. Vives *et al.*, “Prediction of Students’ Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks,” *Ieee Access*, 2024, doi: 10.1109/access.2024.3350169.
[8] F. Tan and W. H. Chan, “Interpreting Student Performance Through Predictive Learning Analytics,” *Int. J. Innov. Comput.*, 2024, doi: 10.11113/ijic.v14n2.434.
[9] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, “Learning Analytics Should Not Promote One Size Fits All: The Effects of Instructional Conditions in Predicting Academic Success,” *Internet High. Educ.*, 2016, doi: 10.1016/j.iheduc.2015.10.002.
[10] A. Villar and C. R. Velini Andrade, “Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study,” *Discov. Artif. Intell.*, 2024, doi: 10.1007/s44163-023-00079-z.
[11] S. Srairi, “An Analysis of Factors Affecting Student Dropout: The Case of Tunisian Universities,” *Int. J. Educ. Reform*, 2021, doi: 10.1177/10567879211023123.
[12] M. G. Gallego, A. P. Pérez Cobos, and J. C. Gómez Gallego, “Identifying Students at Risk to Academic Dropout in Higher Education,” *Educ. Sci.*, 2021, doi: 10.3390/educsci11080427.
[13] J. J. da Silva and N. T. Roman, “Predicting Dropout in Higher Education: A Systematic Review,” 2021, doi: 10.5753/sbie.2021.217437.
[14] A. G. Rincón, S. Barragán, and F. C. Vitery, “Rurality and Dropout in Virtual Higher Education Programmes in Colombia,” *Sustainability*, 2021, doi: 10.3390/su13094953.
[15] M. Y. Amare and S. Simonova, “Global challenges of students dropout: A prediction model development using machine learning algorithms on higher education datasets,” *SHS Web Conf.*, vol. 129, p. 09001, 2021, doi: 10.1051/shsconf/202112909001.
[16] D. A. Gutierrez-Pachas, G. Garcia-Zanabria, E. Cuadros-Vargas, G. Cámara-Chávez, and E. Gomez-Nieto, “Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors Through Machine Learning and Survival Analysis Methods in the Latin American Context,” *Educ. Sci.*, 2023, doi: 10.3390/educsci13020154.
[17] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, “MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series,” *Ieee Access*, 2020, doi: 10.1109/access.2020.3045157.
[18] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, “Predicting Student Success Using Big Data and Machine Learning Algorithms,” *Int. J. Emerg. Technol. Learn. Ijet*, 2022, doi: 10.3991/ijet.v17i12.30259.
[19] S. Hussain and M. Q. Khan, “Student-Performulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning,” *Ann. Data Sci.*, 2021, doi: 10.1007/s40745-021-00341-0.
[20] C. A. dos Santos, G. de Queiroz Pereira, and L. A. Pìlatti, “Higher Education Dropout: A Scoping Review,” *Rev. Gest. Soc. E Ambient.*, 2024, doi: 10.24857/rgsa.v18n8-117.

[21] M. Brinkman, “Student Success Prediction in International Business,” 2021, doi: 10.33422/3rd.educationconf.2021.03.220.