World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Beyond demographics: Identifying academically at-risk students through interpretable machine learning and synthetic minority oversampling

Belonwu Tochukwu Sunday *, Chukwuogo Okwuchukwu Ejike, Ezuruka, Evelyn Ogochukwu and Okechukwu Ogochukwu Patience

Department of Computer Science, Nnamdi Azikiwe University, Nigeria.

World Journal of Advanced Engineering Technology and Sciences, 2026, 19(01), 010-021

Publication history: Received on 24 February 2026; revised on 29 March 2026; accepted on 01 April 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.19.1.0192

Abstract

Student dropout represents a persistent and costly challenge in higher education, undermining institutional effectiveness and limiting individual socioeconomic mobility. This study developed and evaluated a machine learning-based framework for early dropout prediction using a dataset of 10,000 university students across 19 demographic, academic, and psychosocial variables. Four classification algorithms, Logistic Regression, Random Forest, XGBoost, and an Artificial Neural Network were trained and evaluated under two conditions: using the original class-imbalanced data and using training data balanced through the Synthetic Minority Over-sampling Technique (SMOTE). A stratified 80-20 train-test split was applied, and model performance was assessed using F1-Score as the primary metric, supplemented by precision, recall, specificity, ROC-AUC, and PR-AUC. Exploratory data analysis revealed that academic performance variables particularly GPA ($r = -0.460$), Semester_GPA ($r = -0.445$), and CGPA ($r = -0.445$) were the strongest predictors of dropout, while demographic variables including gender and age exhibited negligible predictive value. Logistic Regression trained on SMOTE-balanced data achieved the highest overall performance (F1 = 0.5791, Recall = 0.7537, ROC-AUC = 0.8188), outperforming more complex ensemble and deep learning models across primary metrics. SMOTE consistently improved minority class detection across three of four algorithms, with XGBoost representing a notable exception where balancing marginally degraded performance. These findings demonstrate that interpretable linear models can match or exceed complex architectures in structured educational datasets, and that academically focused early warning systems offer greater intervention utility than demographically targeted approaches.

Keywords: Student dropout prediction; Machine learning; Class imbalance; SMOTE; Logistic regression; Random forest; XGBoost; Artificial neural network; F1-score; Academic performance; Early warning systems; Higher education retention

1. Introduction

Student dropout from higher education represents a persistent challenge for institutions worldwide. According to the U.S. National Center for Education Statistics, undergraduate enrollment rates declined to 63% in 2020, while dropout rates vary significantly by country and institution type[1]. In South Korea, despite high enrollment rates exceeding 70%, the average freshman dropout rate reaches 8.0% [2]. Similarly, in Latin American universities, approximately 37-53% of students in bachelor's programs abandon their studies, with the highest concentration occurring during the first year [3].

The economic and social consequences of student dropout extend beyond individual students. Governments invest billions in higher education funding that is directly lost when students leave their programs. For instance, between 2003 and 2008, the United States invested nearly USD 6.2 billion in students who did not return for a second year, with state

* Corresponding author: Belonwu Tochukwu Sunday

and federal governments respectively contributing USD 1.4 billion and USD 1.5 billion in grants to non-returning students[4]. Dropout is a multifactorial phenomenon influenced by interconnected academic, socioeconomic, psychological, and institutional factors [5]. Rather than resulting from a single cause, student dropout represents the confluence of multiple variables including prior academic achievement, socioeconomic status, family background, engagement in campus activities, and adaptation to university environment [6]. This complexity necessitates sophisticated analytical approaches capable of capturing non-linear relationships and variable interactions. Machine learning techniques have emerged as powerful tools for dropout prediction, offering significant advantages over traditional statistical methods. Extensive research has demonstrated that machine learning models including logistic regression, decision trees, random forests, gradient boosting, neural networks, and transformer-based language models can effectively identify at-risk students with high accuracy and reliability[7,8]. These models enable early intervention by flagging students at risk before dropout occurs, allowing institutions to provide targeted support during critical periods of academic transition. The core challenge in dropout prediction lies in feature selection and engineering to capture meaningful predictors, model selection to balance accuracy with interpretability, and handling severe class imbalance inherent in dropout datasets where non-dropout cases typically outnumber dropout cases by 3-20 times [9].

Recent innovations have extended dropout prediction beyond traditional structured data. Textual data from student feedback, course evaluations, and advisor notes have been shown to improve prediction performance. Deep learning approaches incorporating temporal dynamics, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, capture time-dependent dropout patterns in online learning environments. This paper makes the following key contributions;

- Comprehensive comparison of multiple models on a student dropout dataset
- Develp models that enable early intervention by flagging students at risk before dropout occur.
- Systematic evaluation of SMOTE effectiveness on dropout datasets
- Analysis of feature importance to understand key dropout predictors

2. Related work

The phenomenon of student dropout has long been a subject of scholarly inquiry in higher education. Tinto's seminal integration theory[10] posits that student persistence depends on the degree to which students are academically and socially integrated into the institution. This theory suggests that students who fail to integrate into the academic and social fabric of the university are at higher risk of departure. [11] extended this work by proposing a cost-benefit model in which students weigh the benefits of continued enrollment against the costs and compare their actual experiences with their expectations. Together, these foundational theories establish that dropout results from the interplay of multiple factors including academic performance, social engagement, and institutional fit.

More recent research has confirmed the multifactorial nature of dropout. [12] found that dropout is influenced by interconnected academic, socioeconomic, and psychological factors in the context of remedial engineering programs. [2] similarly identified that both demographic variables (e.g., gender, age, socioeconomic background) and academic performance metrics (e.g., GPA, course evaluations) contribute significantly to dropout risk. The heterogeneous causes of dropout necessitate sophisticated analytical methods capable of capturing complex nonlinear relationships among these variables.

The application of machine learning to dropout prediction has emerged as a promising direction within educational data mining. [13] compared six machine learning algorithms, Naive Bayes, Logistic Regression, Support Vector Machines, Decision Trees, K-Nearest Neighbors, and Artificial Neural Networks on a dataset of 906 computer science students, finding that Logistic Regression achieved the highest F1-score of 0.969. However, their work also demonstrated that on larger datasets with more complex feature interactions, nonlinear models such as Random Forest and gradient boosting techniques may be more effective.

[14] conducted a large-scale study on 165,715 high school students using machine learning for early warning of dropout risk. They compared Random Forest and Boosted Decision Trees, with and without Synthetic Minority Over-sampling Technique (SMOTE), reporting ROC-AUC scores exceeding 0.99 for the Boosted Decision Tree model. Critically, they found that Precision-Recall (PR) curves were more informative than Receiver Operating Characteristic (ROC) curves for evaluating models on imbalanced datasets, a finding that has significant implications for model assessment in the dropout prediction context.

[15] took a holistic approach to feature selection, incorporating academic, demographic, psychological, health, and social factors from survey data of approximately 50 students. Using Naive Bayes classification, they achieved 72% classification accuracy, demonstrating that comprehensive feature engineering across multiple domains improves predictive power beyond academic metrics alone.

Class imbalance represents a significant methodological challenge in dropout prediction, as non-dropout cases typically outnumber dropout cases by 3-20 times in most datasets. [14] extensively examined SMOTE effectiveness, finding that while SMOTE improved recall for the minority class (dropout students), it did not uniformly improve model performance as measured by ROC curves. However, when evaluated using Precision-Recall curves, SMOTE demonstrated clear benefits, a finding that underscores the importance of metric selection in evaluating models on imbalanced data. [12] examined an extreme case of class imbalance (72.8% dropout rate) in a dataset of 2,097 leveling course students in Ecuador. They found that Artificial Neural Networks (ANN) handled severe imbalance better than Logistic Regression, achieving AUC=0.795 compared to AUC=0.475 (essentially random) for logistic regression. This work highlights the importance of model selection under conditions of extreme class imbalance.

Despite substantial progress in applying machine learning to dropout prediction, significant methodological gaps persist in the literature. First, studies consistently report conflicting findings regarding optimal model selection some demonstrate superiority of Logistic Regression, others of Random Forest or gradient boosting approaches, and still others of neural networks, yet few studies systematically compare multiple models on identical datasets using standardized evaluation metrics. Second, the effectiveness of SMOTE for handling class imbalance remains contentious: [14] report substantial performance improvements, while [16] demonstrate that SMOTE actually degrades F1-scores for most models. The conditions under which SMOTE is beneficial versus harmful are inadequately characterized. Third, feature importance rankings vary substantially across studies, with different datasets highlighting GPA, attendance, engagement, or socioeconomic factors as primary predictors, yet no clear guidance exists distinguishing universally predictive features from those dependent on specific institutional or demographic contexts. Fourth, inconsistent application of evaluation metrics, with some studies emphasizing accuracy, others F1-score or AUC, impedes cross-study comparison and obscures the true performance of methods on imbalanced datasets. Finally, while academic studies have convincingly demonstrated technical feasibility of machine learning approaches, fewer studies provide practical implementation guidance for educational institutions, including considerations of model interpretability, computational requirements, and deployment logistics.

This study directly addresses these five critical gaps through a rigorous empirical analysis. We provide the first systematic comparison of four distinct machine learning approaches (Logistic Regression, Random Forest, XGBoost, and Artificial Neural Networks) evaluated with consistent metrics (F1-score, Precision, Recall, ROC-AUC, and Specificity) on a single dataset of 10,000 students. We explicitly test SMOTE effectiveness by training each model with and without synthetic oversampling, enabling definitive characterization of when SMOTE improves or degrades performance for different algorithms. We conduct comprehensive feature importance analysis across all models to identify which predictors consistently emerge as most influential. We employ F1-score as our primary evaluation metric, supplemented by Precision-Recall curves, ensuring appropriate treatment of class imbalance.

3. Material and methods

3.1. Dataset

The dataset utilized in this study was sourced from Kaggle, a publicly accessible open-source data science platform that hosts community-contributed datasets for research and educational purposes. The dataset comprises 10,000 student records collected from a university setting, encompassing 19 features spanning three broad domains: academic performance variables, demographic characteristics, and socioeconomic and psychosocial indicators. The binary target variable, Dropout indicates whether a student withdrew from their program of study, with 23.54% of records classified as dropout and the remaining 76.46% as non-dropout, reflecting the class imbalance characteristic of real-world student attrition data. As an open-access dataset with no personally identifiable information, it provides a suitable and reproducible empirical foundation for benchmarking machine learning approaches to dropout prediction, while its scale of 10,000 records ensures sufficient statistical power for training, validation, and testing across multiple algorithmic configurations.

3.2. Data preprocessing

3.2.1. Missing Value Imputation Strategy

Addressing missing data is a prerequisite for reliable machine learning model development. This study adopted an imputation framework grounded in statistical theory, applying distinct strategies to numerical and categorical variables in recognition of their fundamentally different data structures. For numerical features,Family_Income, Study_Hours_per_Day, and Stress_Index, median imputation was chosen as the most appropriate approach. For the categorical feature Parental_Education, mode imputation was applied, substituting missing entries with the most frequently observed category. As the only central tendency measure applicable to nominal data, mode imputation is both methodologically sound and distribution-preserving, ensuring that the most common educational background remains proportionally represented. Following the completion of imputation across all 10,000 student records, the dataset reached full completeness across all 19 variables, eliminating the need to discard any observations. Retaining the entire sample safeguards statistical power and ensures the analysis reflects the full spectrum of student backgrounds present in the dataset.

3.2.2. Exploratory Data Analysis

Prior to model development, a comprehensive exploratory data analysis (EDA) was conducted to characterize the dataset, uncover underlying patterns, and guide subsequent preprocessing and modeling decisions. The analysis encompassed six interconnected components. Univariate analysis examined each feature in isolation, deriving descriptive statistics, including mean, median, standard deviation, range, and skewness, alongside histogram and density plot visualizations to detect distributional irregularities such as skewness or bimodality. Bivariate analysis computed Pearson correlation coefficients between all features and the dropout target variable, establishing which predictors exhibit the strongest individual associations with student attrition. **Fig 1** presents box plots illustrating the distribution of selected features across dropout and non-dropout groups, providing a visual basis for assessing between-class differences prior to model training.

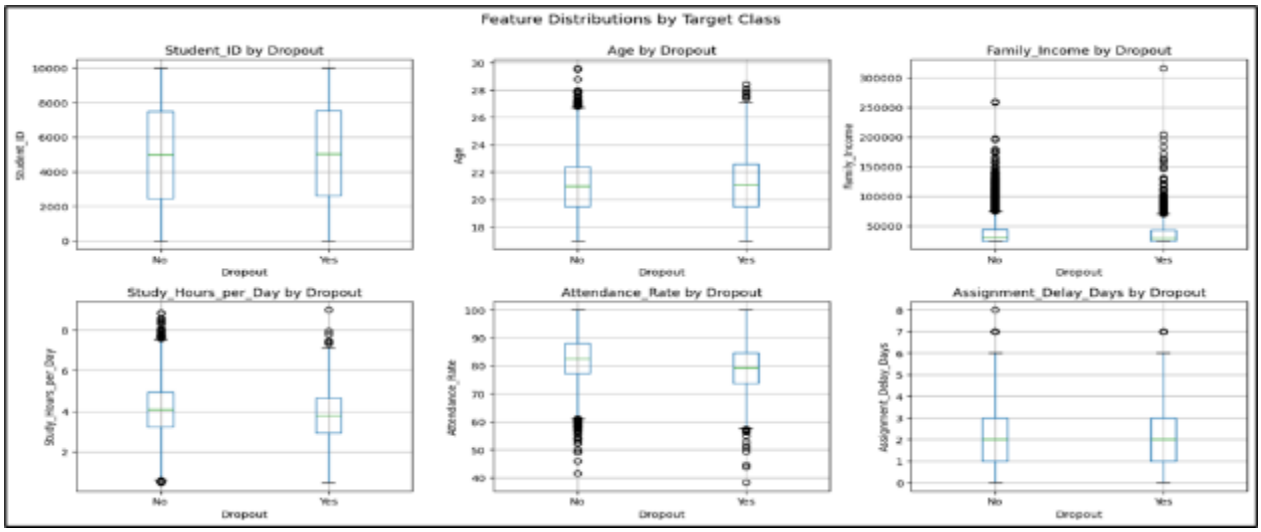

Figure 1 Feature Distribution by Dropout

Class distribution analysis quantified the proportion of dropout and non-dropout students, documenting the extent of class imbalance and establishing the empirical basis for applying specialized balancing techniques. Feature type verification confirmed the correct classification of categorical variables, such as Gender and Department, and validated the numeric formatting of continuous features, ensuring compatibility with downstream processing pipelines. Outlier detection screened for extreme values capable of disproportionately influencing model training; no observations warranting removal were identified. Finally, missing value pattern analysis assessed whether data absence was concentrated within particular subgroups or distributed randomly across the dataset, reinforcing the validity of the MCAR assumption adopted during imputation.

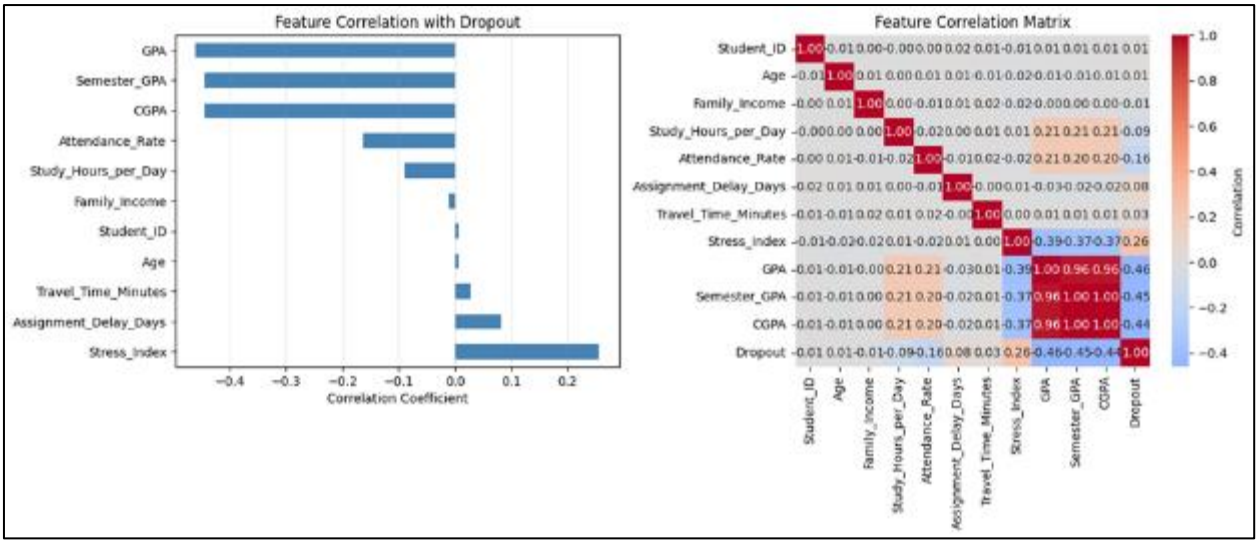

Figure 2 Feature correlation with Dropout

The EDA yielded several findings that directly shaped subsequent methodological decisions**. Fig 2** presents the feature correlation with dropout alongside the full inter-feature correlation matrix, providing a comprehensive visual summary of the linear relationships within the dataset. Correlation analysis identified academic performance variables as the most powerful predictors of dropout: GPA (r = −0.460), Semester_GPA (r = −0.445), and CGPA (r = −0.445) each demonstrated strong negative associations with attrition, collectively underscoring the centrality of scholastic achievement in student persistence. Among psychosocial variables, Stress_Index (r = +0.256) emerged as the sole feature with a meaningful positive correlation with dropout, positioning psychological stress as a potentially modifiable risk factor amenable to early intervention. Conversely, demographic variables, including Gender (r = 0.007), Age (r = 0.008), and Family_Income (r = −0.011), exhibited negligible correlations with dropout outcomes, suggesting that demographically targeted retention strategies may yield limited effectiveness compared to academically oriented approaches. The confirmed class imbalance, with dropout students comprising 23.54% of the sample, provided clear empirical justification for the subsequent application of SMOTE.

3.2.3. Label Encoding

Machine learning algorithms operate exclusively on numerical inputs, necessitating the transformation of categorical variables into numeric representations prior to model training. This study utilized LabelEncoder, a preprocessing transformer available in the scikit-learn library, which converts categorical variables into sequential integer codes by sorting unique category values alphabetically and assigning corresponding integers beginning at zero. To illustrate, the Gender feature originally containing the values {Male, Female} was mapped to {Female=0, Male=1} in accordance with alphabetical ordering. The same logic was applied to Internet_Access {No=0, Yes=1} and the binary variables Part_Time_Job and Scholarship, which were encoded identically.

For the Parental_Education feature, comprising three distinct categories, alphabetical ordering produced the mapping: Bachelor=0, High School=1, Master=2. It is worth noting that this encoding carries an implicit ordinal structure, as the assigned integers suggest a hierarchy that coincidentally aligns with actual educational progression. The Department feature, encompassing five academic disciplines, was encoded as: Arts=0, Business=1, Computer Science=2, Engineering=3, Science=4. Unlike Parental_Education, these departmental categories are inherently nominal, that is, unordered and the integer values carry no meaningful rank relationship. LabelEncoder was preferred over alternative methods, most notably One-Hot Encoding, on the basis of three practical considerations; preserves the original feature dimensionality, accommodate integer-encoded categorical inputs and steamlines the interpretation of feature importance

3.3. Data splitting

Rigorous model evaluation depends on a principled approach to data splitting that guards against overfitting and ensures unbiased performance estimation. This study employed a stratified 80-20 train-test split, a methodology chosen for its ability to preserve class distribution proportionality across both partitions. Stratification guarantees that the training and test sets mirror the class composition of the full dataset, a particularly critical consideration when working with imbalanced data. In this dataset, where dropout students account for 23.54% of observations, an unstratified

random split risks producing partitions with divergent class ratios, for instance, 20% dropout in training and 27% in testing, introducing systematic evaluation bias. Stratification eliminates this risk by enforcing the original 76.46% non-dropout and 23.54% dropout proportions in both subsets.

The resulting partition yielded 8,000 training samples (6,000 non-dropout, 2,000 dropout) and 2,000 test samples (1,500 non-dropout, 500 dropout). The training set is sufficiently large to support robust model learning, while the test set provides an adequate sample base for stable and reliable performance estimation. The 80-20 ratio reflects established machine learning practice, striking a deliberate balance between maximizing the data available for training and retaining enough held-out samples for meaningful evaluation. Alternative splits warrant consideration: a 70-30 or 60-40 division would yield a larger test set at the expense of training data volume, while a 90-10 split would maximize training samples but reduce the test set to approximately 1,000 observations, a size at which performance metrics become sensitive to individual anomalies and less generalizable. Reproducibility was secured through the assignment of a fixed random seed (random_state=42) at the point of splitting

3.4. Class imabalance

Class imbalance presents a meaningful obstacle in dropout prediction, as algorithms trained on skewed data tend to develop a systematic preference for the majority class. In sufficiently extreme cases, a model can attain deceptively high accuracy by predicting non-dropout for every student, correctly classifying the 76.46% majority while entirely failing to detect at-risk individuals. To counteract this tendency, this study explicitly incorporated the Synthetic Minority Over-sampling Technique (SMOTE), an algorithm that generates artificial minority class observations to restore balance within the training set[17].

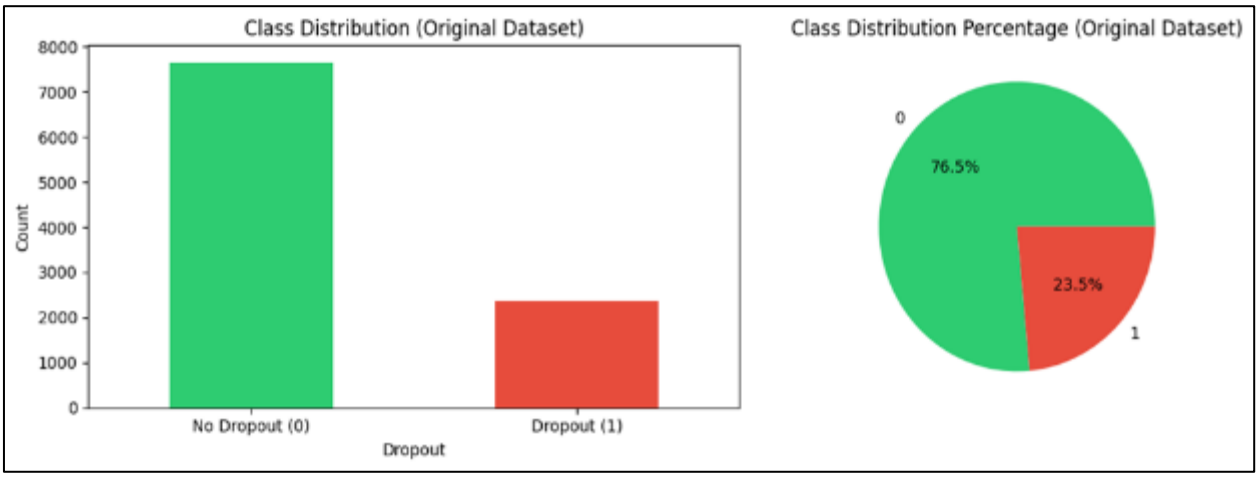

Figure 3 Class Distribution on Dropout

Applied to this dataset, SMOTE expanded the training set from 8,000 imbalanced samples (6,000 non-dropout and 2,000 dropout, reflecting a 3:1 ratio) as shown in Fig 3, to 12,000 balanced samples (6,000 non-dropout and 6,000 synthetic dropout, achieving a 1:1 ratio). The 4,000 newly generated synthetic dropout records were derived from the original 2,000 minority class students through nearest-neighbor interpolation, doubling minority class representation while leaving the majority class intact. The resulting balanced training set affords each class equal standing during model learning, substantially mitigating the bias toward majority-class prediction that would otherwise compromise the model's ability to identify students genuinely at risk of dropout.

3.5. Machine learning model selection

Four machine learning algorithms were selected to enable a comprehensive and methodologically diverse comparison of predictive approaches for student dropout. Logistic Regression[24] served as the linear baseline model, estimating dropout probability by learning weighted contributions from each feature and mapping them to a value between zero and one. Its primary strengths lie in interpretability and computational efficiency the learned coefficients directly reveal the direction and relative magnitude of each feature's influence on dropout risk, making it particularly valuable for generating actionable insights. Random Forest[21] extended this foundation by constructing an ensemble of 100 independently trained decision trees, each built on a random subset of training data and features. By aggregating the votes of diverse trees, Random Forest naturally captures nonlinear relationships and feature interactions that linear models cannot represent, while simultaneously providing built-in feature importance rankings through Gini impurity

measures. Both models were trained using scikit-learn's default configurations[26] without hyperparameter tuning, reflecting their roles as interpretable and computationally accessible benchmarks.

The two remaining algorithms introduced greater modeling complexity. XGBoost[22] employed a sequential boosting strategy in which each successive decision tree was constructed to correct the prediction errors of the preceding ensemble, rather than being built independently. This iterative error-correction mechanism, combined with built-in L1 and L2 regularization penalties, enables XGBoost to achieve strong predictive performance on structured tabular data while guarding against overfitting. The model was configured with 100 estimators, a learning rate of 0.1, and a maximum tree depth of six. The Artificial Neural Network (ANN)[23] represented the most architecturally complex approach, comprising four layers an input layer corresponding to the 19 features, three progressively narrowing hidden layers of 128, 64, and 32 neurons, and a single output neuron producing a dropout probability. The network was trained using the Adam optimizer[24] and binary crossentropy loss, with early stopping and dropout regularization incorporated to prevent overfitting across up to 100 training epochs. Together, these four algorithms span the spectrum from transparent linear models to high-capacity deep learning architectures, enabling a rigorous empirical comparison across fundamentally different modeling paradigms.

3.6. Evaluation metrics

Given the pronounced class imbalance characterizing this dataset, a multi-metric evaluation framework was adopted to ensure that model performance was assessed across complementary dimensions rather than any single potentially misleading indicator. **F1-Score** was designated as the primary metric, as its formulation as the harmonic mean of precision and recall penalizes models that sacrifice one dimension for the other a critical property in imbalanced settings where a naive classifier predicting non-dropout for every student would achieve 76.46% accuracy while completely failing its core objective [18][20]. **Precision** and **Recall** were reported as supplementary metrics capturing the institutional trade-off between intervention efficiency and intervention coverage, respectively high recall ensuring that fewer at-risk students go undetected, and high precision minimizing unnecessary resource expenditure on false alarms. **Specificity** further quantified the model's ability to correctly identify students who will persist, supporting confident resource allocation decisions.

For ranking-based assessment, both **ROC-AUC** and **PR-AUC** were reported. While ROC-AUC is the conventional benchmarking standard, it is known to overstate discriminative performance on imbalanced datasets by treating majority and minority class errors symmetrically [19]. PR-AUC, which focuses exclusively on minority class behavior by plotting precision against recall, provides a more informative and sensitive evaluation under class imbalance, a position consistent with [14], whose dropout prediction methodology this study aligns with. All metrics were derived from confusion matrices documenting true positives, true negatives, false positives, and false negatives for each model configuration.

4. Result and discussion

4.1. Overall Model Performance

The Fig. 4 presents the comprehensive performance comparison across all eight model-dataset configurations. Measured against the primary evaluation metric of F1-Score, Logistic Regression trained on SMOTE-balanced data emerged as the best-performing configuration (F1 = 0.5791), followed by Random Forest with SMOTE (F1 = 0.5488) and Neural Network with SMOTE (F1 = 0.5257). A consistent pattern is immediately observable across nearly all algorithms: SMOTE application improved F1-Score and recall at the cost of reduced accuracy and specificity. This trade-off reflects the fundamental rebalancing effect of SMOTE by exposing models to equal class representation during training, they become more willing to predict dropout, correctly identifying more at-risk students while simultaneously generating more false alarms among the non-dropout majority. The one notable exception to this pattern was XGBoost, where SMOTE marginally degraded F1-Score (0.4903 versus 0.5006 without SMOTE), suggesting that XGBoost's internal regularization mechanisms and gradient-based error correction may already provide sufficient handling of class imbalance, rendering synthetic oversampling counterproductive in this specific case, a finding consistent with [16] caution that SMOTE does not universally confer performance benefits.

```
================================================================================
COMPREHENSIVE RESULTS COMPARISON
================================================================================
              Model          Dataset Accuracy F1 Score Precision Recall ROC AUC Specificity
Logistic Regression       Unbalanced   0.8120   0.5000    0.6690 0.3992  0.8206      0.9392
Logistic Regression Balanced (SMOTE)   0.7420   0.5791    0.4702 0.7537  0.8188      0.7384
      Random Forest       Unbalanced   0.8035   0.4696    0.6444 0.3694  0.7994      0.9372
      Random Forest Balanced (SMOTE)   0.7805   0.5488    0.5319 0.5669  0.7897      0.8463
            XGBoost       Unbalanced   0.7965   0.5006    0.5930 0.4331  0.7763      0.9084
            XGBoost Balanced (SMOTE)   0.7775   0.4903    0.5323 0.4544  0.7755      0.8770
     Neural Network       Unbalanced   0.7915   0.4491    0.5944 0.3609  0.7969      0.9241
     Neural Network Balanced (SMOTE)   0.7555   0.5257    0.4839 0.5754  0.7760      0.8110
```

Figure 4 Comprehensive Results comparison

Examining the unbalanced condition in isolation reveals an important limitation of accuracy as a standalone metric. All four models achieved accuracy figures between 79% and 81% on the imbalanced test set, figures that might superficially suggest strong performance. However, the corresponding recall values ranging from 0.3609 (Neural Network) to 0.4331 (XGBoost), expose a critical deficiency: across all unbalanced models, the majority of actual dropout students were misclassified as non-dropout. The confusion matrices presented in Figure X corroborate this finding. For instance, Logistic Regression (unbalanced) correctly identified 1,436 non-dropout students but failed to flag 283 of the 471 actual dropout cases, while the Neural Network (unbalanced) misclassified 301 dropout students, the highest false negative count among all unbalanced models. These patterns demonstrate precisely why accuracy is an inadequate primary metric for imbalanced educational datasets and validate the decision to prioritize F1-Score and recall in this study's evaluation framework.

4.2. Impact of SMOTE on Recall and Precision Trade-offs

The effect of SMOTE on the recall-precision trade-off was most pronounced in Logistic Regression, where recall increased dramatically from 0.3992 (unbalanced) to 0.7537 (balanced), the highest recall value recorded across all configurations. This improvement means that the SMOTE-trained Logistic Regression successfully identified approximately three-quarters of all at-risk students in the test set, compared to fewer than four in ten under the unbalanced condition. The confusion matrix in Fig. 5 visually reinforces this shift: the SMOTE configuration flagged 355 true dropout students correctly, compared to only 188 in the unbalanced version.

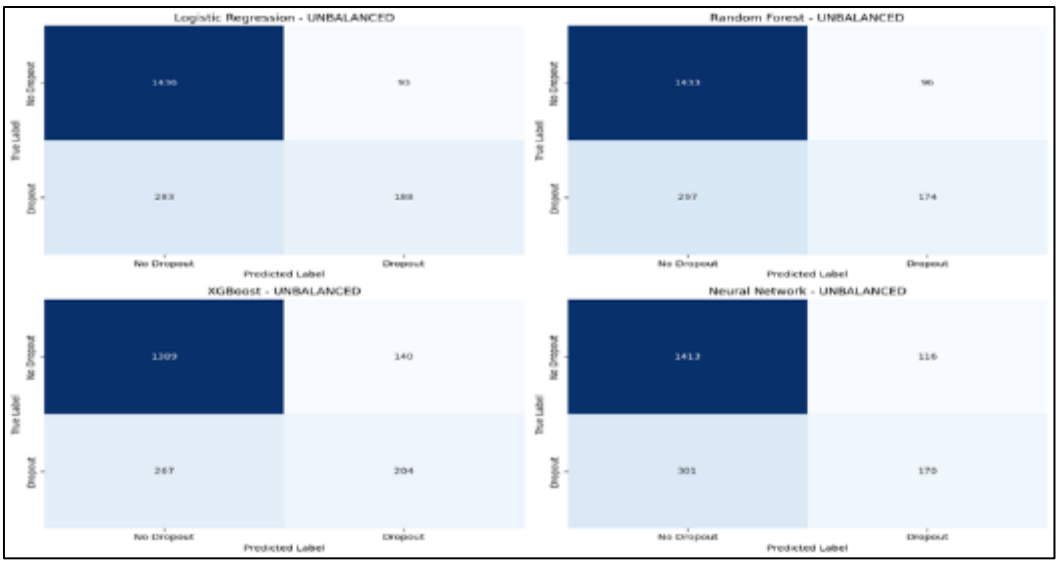

Figure 5 Confusion matrix under unbalanced condition

However, this gain came at a substantial cost, precision fell from 0.6690 to 0.4702, and the false positive count rose sharply from 93 to 400, meaning a considerably larger number of non-dropout students were incorrectly flagged for intervention. This trade-off is not inherently problematic; in institutional contexts where early intervention carries low cost and missing an at-risk student has severe consequences, high recall is the more operationally valuable outcome.

Conversely, in resource-constrained settings, the precision degradation associated with SMOTE may limit its practical utility. Random Forest displayed a more moderate and arguably more balanced SMOTE response recall improved from 0.3694 to 0.5669, while precision declined only modestly from 0.6444 to 0.5319, yielding the second-highest F1-Score overall.

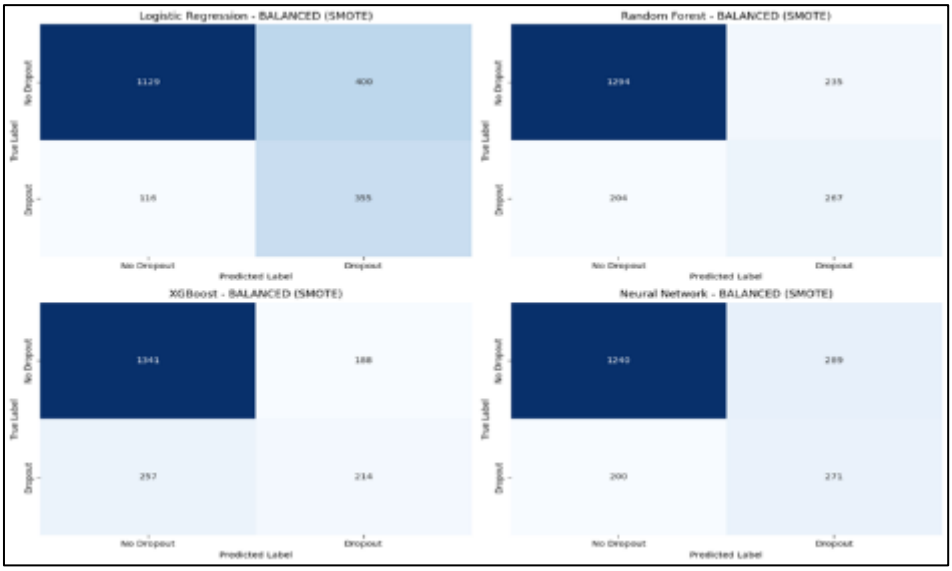

Figure 6 Confusion matrix under balanced condition

Its confusion matrix under the balanced condition (Fig. 6) shows 267 true positives against 204 false negatives, a substantially more evenly distributed error profile compared to the unbalanced version's 174 true positives and 297 false negatives. The Neural Network followed a similar trajectory, with recall rising from 0.3609 to 0.5754 under SMOTE, though its overall F1 remained below both Logistic Regression and Random Forest across conditions, suggesting that the network's greater architectural complexity did not translate into superior discrimination for this dataset size and feature set.

4.3. ROC-AUC Analysis

The ROC curve comparisons presented in Fig. 7 and 8 reveal a somewhat different performance hierarchy than F1-Score rankings alone would suggest. In the unbalanced condition (Fig. 7), Logistic Regression achieved the highest ROC-AUC (0.8206), followed by Random Forest (0.7994), Neural Network (0.7969), and XGBoost (0.7763).

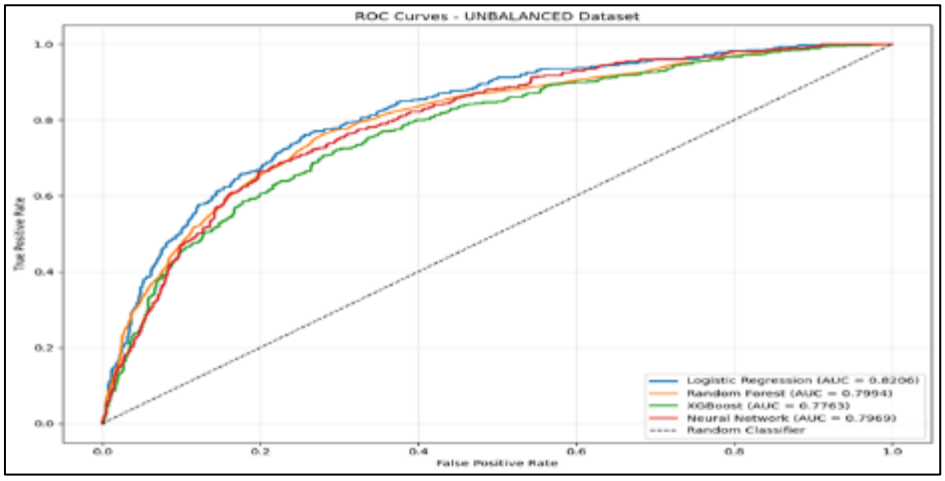

Figure 7 ROC curve under unbalanced Dataset

This ordering was largely preserved in the SMOTE condition (Fig. 8), where Logistic Regression again led (AUC = 0.8188), with the remaining three models clustered closely between 0.7755 and 0.7897. Two observations warrant

emphasis. First, SMOTE had a negligible effect on ROC-AUC across all models scores shifted by no more than 0.01 in either direction between conditions, reinforcing the established critique that ROC-AUC is relatively insensitive to class distribution changes and thus insufficient as a standalone evaluation metric for imbalanced problems.

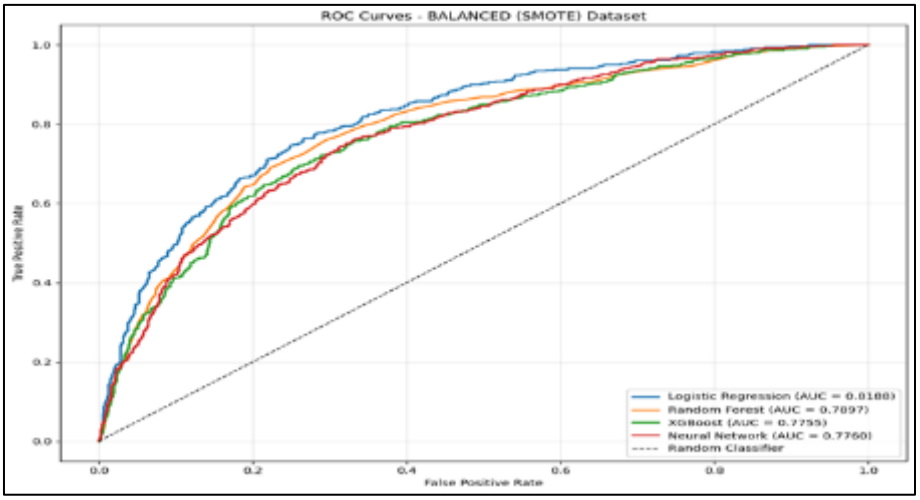

Figure 8 ROC curve under balanced Dataset

Second, the consistent superiority of Logistic Regression in ROC-AUC across both conditions suggests that the underlying relationship between features and dropout probability is substantially linear in nature, allowing the simpler model to rank students by risk as effectively as, or more effectively than the more complex ensemble and deep learning approaches.

4.4. Key Findings and Practical Implications

Synthesizing across all metrics and conditions, three principal findings emerge. First, Logistic Regression with SMOTE represents the optimal configuration for recall-prioritized deployment, achieving the highest F1-Score and by far the highest recall among all models, making it most suitable for institutions whose primary objective is ensuring that no at-risk student goes unidentified. Second, SMOTE consistently improved minority class detection across three of four algorithms, providing empirical support for its application in educational dropout prediction, though the XGBoost exception highlights that practitioners should empirically validate SMOTE's effect for each specific algorithm rather than assuming universal benefit. Third, model complexity did not correlate with predictive superiority in this study, the simplest model (Logistic Regression) outperformed the most complex (Neural Network) across all primary metrics, suggesting that the linear structure of the feature-dropout relationship in this dataset does not necessitate the representational capacity of deep learning architectures. This finding has practical implications for institutional deployment, where simpler, more interpretable models are generally preferable when predictive performance is comparable.

5. Conclusion

This study evaluated four machine learning algorithms for student dropout prediction under both imbalanced and SMOTE-balanced training conditions, yielding findings of methodological and practical significance. The results consistently demonstrated that model complexity does not guarantee superior performance in structured educational datasets, Logistic Regression outperformed Random Forest, XGBoost, and the Artificial Neural Network across primary metrics, suggesting that the feature-dropout relationship in this context is predominantly linear in nature. The dual training design further revealed that SMOTE's benefits are algorithm-dependent rather than universal: while minority class detection improved substantially for three of four models, XGBoost performed marginally better without synthetic oversampling, underscoring the importance of empirically validating balancing strategies for each specific algorithm-dataset combination rather than adopting them as default preprocessing steps.

Practically, the overwhelming predictive dominance of academic performance variables, GPA, Semester_GPA, and CGPA over demographic features strongly suggests that institutionally effective retention strategies should prioritize academic monitoring and support over demographic targeting. The psychosocial signal carried by Stress_Index further highlights psychological wellbeing as a modifiable intervention point worthy of systematic monitoring. Future work

should pursue cross-institutional validation, hyperparameter optimization, and the incorporation of behavioral engagement features such as learning management system activity to strengthen the generalizability and operational utility of dropout prediction frameworks.

Compliance with ethical standards

Statement of ethical approval

The authors declare that all procedures performed in this study were in accordance with ethical standards and relevant guidelines.

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] National Center for Education Statistics. (2020). The condition of education 2020 (NCES 2020-144). U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/.

[2] Won, H. S., Kim, M. J., Kim, D., Kim, H. S., & Kim, K. M. (2023). University student dropout prediction using pretrained language models. Applied Sciences, 13(12), 7073.

[3] Heredia, R., & Carcausto-Calla, W. (2024). Factors associated with student dropout in Latin American universities: Scoping review. J. Educ. Soc. Res, 14, 62-72.

[4] Schneider, M. (2010). Finishing the first lap: The cost of first-year student attrition in America's four-year colleges and universitites. American Institutes for Research.

[5] Barragán_Moreno, S. P., Guzmán_Rincón, A., Calderón_Carmona, G. P., González_Támara, L., & Lozano_Galindo, O. L. (2025). Identifying key variables of student dropout in preschool, primary, secondary, and high school education: An umbrella review approach. European Journal of Educational Research, 14(2), 585-600.

[6] Valencia-Arias, A., Chalela, S., Cadavid-Orrego, M., Gallegos, A., Benjumea-Arias, M., & Rodríguez-Salazar, D. Y. (2023). University dropout model for developing countries: A Colombian context approach. Behavioral Sciences, 13(5), 382.

[7] Duro, B., Gomes, A., Correia, F. B., Borges, A. R., & Bernardino, J. (2026). Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review. Computers, 15(3), 164.

[8] Seo, E. Y., Yang, J., Lee, J. E., & So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. Heliyon, 10(11).

[9] Morsu, H. R. (2025). Machine Learning Models for Predicting Student Dropout, Enrollment, and Graduation.

[10] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Journal of Higher Education*, 45(1), 89–125.

[11] Bean, J. P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. Research in Higher Education, 17(4), 291–320.

[12] Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, 12(22), 9314. https://doi.org/10.3390/su12229314

[13] Osemwegie, O., & Amadin, F. I. (2023). Student dropout prediction using multiple machine learning techniques. *FUDMA Journal of Sciences*, 7(4), 234–245.

[14] Lee, H., & Chung, S. (2019). The Machine learning-based early warning system for high school dropout prediction. Applied Sciences , 9(15), 3093. https://doi.org/10.3390/app9153093

[15] Hegde, V., & Prageeth, P. (2018). Higher education student dropout prediction and analysis through educational data mining. In Proceedings of the 7th International Conference on Intelligent Systems and Control (ICISC)/ IEEE.

[16] Cho, B. H., Yu, H., & Kim, K. W. (2023). A study on dropout prediction for university students using machine learning. Applied Sciences, 13(21), 12004; https://doi.org/10.3390/app132112004

[17] Onyedinma, E. G., Asogwa, D. C., Belonwu, T. S., & Mbonu, C. E. (2025). A Multi-Algorithmic Approach to Stroke Risk Prediction Using Machine Learning. Journal of Engineering Research and Reports, 27(7), 247-259.

[18] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

[19] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

[20] Mbonu, C. E., Anigbogu, K., Asogwa, D., & Belonwu, T. (2025). An explorative analysis of svm classifier and resnet50 architecture on african food classification. arXiv preprint arXiv:2505.13923.

[21] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[22] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[23] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[24] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[25] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.