Predictive Modeling of Students' Dropout Risk Using Intelligent Analytics

[1]**Jude Alphonsus Inyangetoh, [2]Ekemini Anietie Johnson**;
[1]Department of Statistics, Federal Polytechnic Ukana, Akwa Ibom State
2 Department of Computer Science, Federal Polytechnic Ukana, Akwa Ibom State

doi: https://doi.org/10.37745/ejsp.2013/vol14n1118 Published January 31, 2026

Citation: Inyangetoh J.A., and Johnson E.A. (2026) Predictive Modeling of Students' Dropout Risk Using Intelligent Analytics, *European Journal of Statistics and Probability*, 14 (1) 1-18

Abstract: *Student dropout is a persistent challenge in higher education, particularly in developing countries like Nigeria, where reactive institutional responses often fail to identify students at-risk in time. This study proposes an intelligent analytics-based predictive modeling framework designed to transition institutional strategies from reactive to proactive early intervention. Using a dataset of 2,200 student records from Federal Polytechnic Ukana and Akwa Ibom State Polytechnic, the research evaluates the effectiveness of two ensemble learning algorithms: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The methodology involved robust data preprocessing, including Min-Max normalization and Principal Component Analysis (PCA), which identified 16 key predictors from an initial 22 variables. These variables spanned academic performance, demographic backgrounds, and behavioral patterns. Experimental results conducted in a Python environment revealed that XGBoost outperformed RF across all evaluation metrics. XGBoost achieved an accuracy of 0.92, precision of 0.91, recall of 0.90, and an F1-score of 0.91, compared to RF's accuracy of 0.87. Feature importance analysis highlighted "Attendance in Classes" and "Previous Academic Results" as the most significant predictors of attrition. The study concludes that intelligent analytics can effectively capture nonlinear relationships in student data to provide actionable insights. This framework offers a scalable solution for Nigerian tertiary institutions to implement evidence-based retention strategies, ultimately improving graduation outputs and institutional efficiency.*

Keywords: intelligent analytics, student dropout risk, predictive modeling, random forest, XGboost, Nigerian higher education, educational data mining.

INTRODUCTION

Student dropout remains a significant challenge in higher education, with far-reaching consequences for both learners and educational institutions. Globally, high attrition rates undermine institutional objectives, lower graduation outputs, and incur substantial financial losses for families and governments alike. The issue is particularly acute in developing countries such as Nigeria, where limited technological infrastructure, inadequate student support systems, and resource constraints hinder effective monitoring and early intervention. In many polytechnics and universities, institutional responses to dropout risk are predominantly reactive, addressing the problem only after students disengage academically or withdraw altogether. This reactive approach allows early warning signs including declining academic performance, irregular attendance, and disengagement to go unrecognized until they result in irreversible dropout events.

Theoretical perspectives in student retention research highlight the multifaceted nature of dropout risk. For instance, Tinto's Student Integration Theory posits that academic success is shaped by both academic and social integration within an institution (Tinto, 1975). According to this framework, factors such as cumulative grade point average (CGPA), class participation, engagement with learning resources, and interaction with institutional structures influence a student's likelihood of persistence or withdrawal. However, many institutions in low-resource settings collect large volumes of student data without leveraging advanced analytical techniques to extract meaningful patterns or predict risk. Consequently, early intervention strategies are constrained by subjective judgment rather than evidence derived from institutional data.

Advances in machine learning and educational data mining have created new opportunities for transforming raw educational data into actionable insights that support proactive decision-making. Predictive analytics uses historical data to forecast future outcomes and has been successfully applied across domains such as healthcare, finance, and customer analytics. In education, predictive modeling has been used to identify students at risk of poor academic performance and dropout by analyzing complex interactions among academic, demographic, and behavioral attributes (Dasi and Kanakala, 2022; Carballo-Mendívil, Rodríguez-Hernández, and López-Martín, 2025). These models go beyond descriptive statistics by capturing nonlinear relationships and latent patterns that traditional approaches often overlook.

Ensemble learning algorithms, such as Random Forest and Extreme Gradient Boosting (XGBoost), have emerged as powerful tools for predictive tasks due to their robustness, high accuracy, and ability to handle heterogeneous data types (Zhang and Liu, 2019). Random Forest constructs multiple decision trees through bootstrap aggregation to improve classification stability, while XGBoost enhances model performance through iterative boosting and regularization techniques. Both algorithms have demonstrated superior performance in educational prediction contexts, making them suitable choices for dropout risk modeling.

Despite the proven potential of intelligent analytics in educational prediction, its application in Nigerian tertiary institutions remains limited. Few studies have leveraged data-driven methods to anticipate dropout risk and inform early intervention strategies. This gap underscores the need for locally relevant research that applies predictive modeling techniques to real institutional data. Accordingly, this study proposes and evaluates an intelligent analytics–based predictive modeling framework using Random Forest and XGBoost to assess students' dropout risk. By comparing model performance using standard evaluation metrics and interpreting results in the context of educational decision-making, the study aims to support proactive, evidence-based retention strategies in Nigerian higher education.

LITERATURE REVIEW

Predictive modeling using intelligent analytics has gained significant attention as a data-driven approach for addressing complex decision-making challenges across multiple domains, including education, healthcare, finance, and geosciences. In the educational domain, educational data mining (EDM) and learning analytics focus on extracting meaningful patterns from student data to predict outcomes such as academic performance, retention, and dropout risk. Among these, student dropout prediction has emerged as a critical research area due to its direct implications for institutional effectiveness and student success.

Recent studies demonstrate that machine learning–based predictive models are effective in identifying students at risk of dropping out by analyzing academic, behavioral, and demographic variables. Algorithms such as RF, Support Vector Machine, decision trees, and boosting techniques have consistently shown strong predictive capabilities in dropout risk modeling (Dasi and Kanakala, 2022). These intelligent models outperform traditional statistical approaches by capturing nonlinear relationships and complex interactions among student-related factors that are often missed by conventional methods.

Ensemble learning techniques, particularly Random Forest and Extreme Gradient Boosting (XGBoost), are widely adopted in predictive modeling of educational systems, due to their robustness, scalability, and high classification accuracy. Random Forest aggregates multiple decision trees to reduce overfitting and improve generalization, while XGBoost employs gradient boosting with regularization to enhance predictive performance on structured educational datasets (Zhang and Liu, 2019). Empirical evidence indicates that these ensemble models consistently outperform single classifiers in dropout prediction tasks. Furthermore, hybrid and stacking approaches that integrate Random Forest and XGBoost have been explored to further improve prediction accuracy and reliability in early warning systems

Despite the global progress in intelligent analytics for student retention, localized applications within Nigerian tertiary institutions, particularly polytechnics remain limited. Most existing studies in Nigeria focus on descriptive analyses or basic statistical methods, offering limited predictive insight for proactive intervention. Notable contributions by Ekemini Johnson and collaborators, however, demonstrate the feasibility and effectiveness of intelligent analytics in educational prediction. Johnson and Inyang (2025) proposed an intelligent ensemble learning framework combining XGBoost and Random Forest to predict students' academic performance, achieving improved accuracy compared to single-model approaches. Their findings support the applicability of ensemble learning techniques for early identification of academically at-risk students.

Beyond the educational domain, Johnson E. A. has contributed extensively to machine learning research in diverse application areas, reflecting strong methodological expertise relevant to predictive modeling. For instance, Johnson et al. (2025) conducted a systematic review of machine learning techniques applied to petrophysical analysis and original oil-in-place estimation, highlighting the adaptability of predictive algorithms across heterogeneous data environments. Additionally, comparative studies on fake news detection revealed the superior performance of Random Forest over traditional decision tree models, further validating the effectiveness of ensemble methods in classification tasks. These contributions reinforce the suitability of such intelligent analytics techniques for modeling student dropout risk.

Another critical aspect of predictive modeling emphasized in recent literature is model interpretability. While high prediction accuracy is essential, understanding why a student is predicted to be at risk is equally important for practical intervention. Feature importance analysis enables institutions to identify key predictors of dropout, such as CGPA, attendance patterns, study habits, and socioeconomic background (Carballo-Mendívil et al., 2025). Interpretable models support evidence-based policy formulation and targeted support strategies, thereby bridging the gap between prediction and action.

Existing literature reveals three dominant trends: (1) intelligent analytics and machine learning models are effective tools for predictive modeling of student dropout risk; (2) ensemble learning methods such as Random Forest and XGBoost consistently deliver superior predictive performance and robustness; and (3) feature importance analysis enhances the interpretability and practical utility of predictive systems. However, despite extensive global research, the adoption of predictive modeling for dropout risk in Nigerian polytechnics remains underexplored. By leveraging intelligent analytics and focusing on both prediction accuracy and interpretability, the present study addresses this gap and contributes to the development of proactive, data-driven student retention strategies.

Random Forest

Random Forest is a popular machine learning algorithm and an ensemble learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability (Wainberg et al., 2016). A Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a p-dimensional random vector $X = (X_1, . . . ,Xp)^T$representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution $P_{XY}(X,Y)$. The goal is to find a prediction function $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss.

$$E_{XY}(L(Y, f(X))) \qquad \text{Equation 1}$$

where the subscripts denote expectation with respect to the joint distribution of *X* and *Y*. Intuitively, $L\ (Y, f\ (X))$ is a measure of how close $f\ (X)$ is to *Y*; it penalizes values of $f\ (X)$ that are a long way from *Y*. Typical choices of *L* are *squared error loss* $L\ (Y, f\ (X)) = (Y - f\ (X))^2$ for regression and *zero-one loss* for classification:

$$L\ (Y, f\ (X)) = I\big(Y \neq f(X)\big) = \begin{cases} 0\ if\ Y = f(X) \\ 1\ \ otherwise \end{cases} \qquad \text{Equation 2}$$

It turns out that minimizing E_{XY} (L (Y, f (X))) for squared error loss gives the conditional expectation

$$f(x) = E(Y|X=x) \qquad \text{Equation 3}$$

Otherwise known as the *regression function*. In the classification situation, if the set of possible values of *Y* is denoted by Y, minimizing $E_{XY}(L\ (Y, f\ (X)))$ for zero-one loss gives:

$$f\ (x) = argmaxP(Y=y|X=x) \qquad \text{Equation 4}$$

otherwise known as the *Bayes rule*.

Ensembles construct *f* in terms of a collection of so-called "base learners" $h_1(x)$, . . ., hJ(x) and these base learners are combined to give the "ensemble predictor" $f\ (x)$. In regression, the base learners are averaged

$$f(x)=\frac{1}{J}\sum_{J=1}^{J} h_j\ (x) \qquad \text{Equation 5}$$

$$f(x) arg\mathrm{max}_{y \in Y} \sum_{J=1}^{J} I\ (y = h_j(x)) \qquad \text{Equation 6}$$

Extreme Gradient Boost

The XGBoost (Extreme Gradient Boosting) algorithm is an optimized and scalable implementation of gradient-boosted decision trees (GBDT), which is designed for speed and performance. It has become one of the most popular machine learning algorithms due to its efficiency, flexibility, and ability to handle a variety of data types and problems.

The Working of XGBoost can be explained using the following steps:

a. Initialization: Starts with an initial prediction, usually the mean (for regression) or a uniform distribution (for classification).
b. Gradient Descent Optimization: Each subsequent tree is trained to minimize the residual errors (gradients) of the previous predictions.
c. Tree Construction: Trees are built iteratively, where splits are determined based on the reduction of the loss function (e.g., Mean Squared Error for regression, Logarithmic Loss for classification).
d. Weighted Learning: Each tree assigns weights to instances, giving higher importance to incorrectly predicted examples.
e. Final Prediction: Combines the predictions of all the trees (via weighted sums for regression or probability scores for classification) to make the final output.

XGBoost has the following advantages:

i. High Performance: Delivers state-of-the-art results in competitions and benchmarks.
ii. Flexibility: Works with multiple loss functions and is extensible with user-defined objectives.
iii. Scalability: Handles large datasets and high-dimensional data efficiently.
iv. Robustness: Built-in features like regularization and early stopping help prevent overfitting.

XGBoost ca be applied in Classification problems (e.g., credit risk analysis, fraud detection), Regression problems (e.g., price prediction, sales forecasting), Ranking problems (e.g., search engine result ranking), Time-series forecasting and Feature selection through its feature importance scores.

REVIEW OF RECENT WORKS

Table 2.1 shows the review of some recent works relating to prediction of students academic performance and dropout risk.

Table 2.1: Review of Recent Works

Citation	Title of Research	Objectives	Methodology	Problem Solved	Limitations
Smith et al. (2020)	Predicting Student Dropouts Using Random Forest	To use machine learning to identify students at risk of dropping out	Random Forest algorithm on academic and demographic data	Early identification of at-risk students	Limited interpretability of model results
Lee and Park (2019)	Deep Learning Approaches to Predict Student Attrition	To explore deep learning models for dropout prediction	LSTM and DNN models on student performance logs	High prediction accuracy	Requires large, labeled datasets
Kumar et al. (2021)	Academic Risk Prediction Using SVM	To develop an SVM-based model for academic dropout risk	SVM classifier trained on academic and attendance data	Accurate prediction for small datasets	Sensitive to parameter tuning
Rodriguez and Silva (2018)	Dropout Detection in MOOCs Using Analytics	To predict dropouts in online courses using clickstream data	Logistic regression and clustering	Helped reduce dropout in MOOCs	Only applicable to online platforms
Chen and Li (2022)	Early Warning System Using Decision Trees	To create a decision tree-based early alert system	CART decision tree model with educational dataset	Identified key dropout indicators	Overfitting risk on complex data

Ahmed et al. (2020)	Predicting Student Dropout with Ensemble Models	To compare ensemble methods for dropout prediction	Bagging, Boosting, and Stacking techniques	Improved prediction robustness	Increased computational cost
Fatima and Noor (2019)	Student Dropout Analysis in Tertiary Institutions	To identify major causes and patterns of dropout	Data mining with association rules and clustering	Insights into dropout causes	Lacked predictive model implementation
Gomez et al. (2021)	Predictive Analytics for Student Retention	To improve retention using predictive insights	Multivariate regression and data visualization	Better intervention strategies	Limited to quantitative data
Osei-Bonsu and Tetteh (2022)	A Hybrid Model for Predicting University Dropout	To develop a hybrid ML model combining rule-based and ML	Rule-based filtering + Naïve Bayes	Enhanced accuracy with interpretable rules	Complex system integration
Adeyemi et al. (2020)	Application of Naïve Bayes in Dropout Prediction	To apply Naïve Bayes for early dropout detection	Naïve Bayes classifier on enrollment and exam records	Fast, low-resource prediction model	Low performance on imbalanced data
Zhang and Liu (2019)	Comparative Study of ML Algorithms for Dropout	To evaluate ML algorithms on dropout datasets	KNN, DT, RF, SVM on education data	Identified most effective algorithms	Did not include external factors
Musa and Salihu (2021)	Socioeconomic Predictors of Student Dropout	To analyze socioeconomic impact on dropout rates	Logistic regression with socioeconomic variables	Revealed influence of family income	Non-academic factors underrepresented
Patel et al. (2020)	Dropout Prediction Using Clustering Techniques	To segment at-risk students using clustering	K-Means clustering on engagement metrics	Grouped students for intervention	No actual prediction, only grouping
Wang et al. (2022)	A Time-Series Model for Dropout Prediction	To predict dropouts over time using sequential data	ARIMA and LSTM models	Detected dropout trends over semesters	Requires historical and time-stamped data
Johnson et al. (2018)	Intelligent Analytics for Student Success	To use AI for analyzing student success and dropout	AI dashboard with ML and visualization tools	Provided decision support for faculty	High infrastructure requirement
Nwankwo and Okonkwo (2021)	A Case Study of Dropout Risks in Nigerian Polytechnics	To investigate dropout risks using intelligent analytics	Case study + predictive modeling (SVM)	Informed policy on academic support	Limited generalizability
Abebe and	Machine Learning for	To assess ML models in	Logistic regression, Random Forest	Demonstrated ML applicability in	Data scarcity and poor quality

Mekonnen (2020)	Dropout Prediction in Africa	resource-constrained settings		developing contexts	
Singh et al. (2023)	Predictive Modeling Using AutoML	To automate dropout risk prediction using AutoML	Google AutoML on large student dataset	Reduced model selection complexity	Black-box nature of AutoML
Torres and Luna (2021)	Psychological Factors in Dropout Prediction	To include psychological traits in ML models	Surveys + ML integration (SVM, RF)	Improved model accuracy with soft data	Privacy and subjectivity concerns
Bello and Haruna (2022)	Student Dropout Detection via Neural Networks	To evaluate NN performance in dropout forecasting	Feedforward Neural Networks with backpropagation	Modeled nonlinear dropout patterns	Requires high computational power

METHODOLOGY

This study adopted an intelligent analytics driven methodology based on supervised machine learning to predict students' dropout risk and identify the most influential factors contributing to attrition in tertiary education. The methodological process was designed to ensure robustness, interpretability, and reproducibility, and it followed established practices in educational data mining and predictive analytics.

Research Design

A quantitative research design grounded in computational modeling was employed. Since student dropout is a categorical outcome, the problem was formulated as a supervised binary classification task, where students were classified as either at risk of dropout or not at risk. Supervised machine learning techniques are widely used in educational prediction problems due to their ability to learn patterns from historical labeled data and generalize to unseen cases (Dasi and Kanakala, 2022).

Ensemble learning algorithms; Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were selected as the core predictive models because of their proven effectiveness in handling high-dimensional educational datasets and capturing nonlinear relationships among student-related variables.

Data Source and Study Area

The dataset used in this study was obtained from selected Nigerian tertiary institutions, specifically Federal Polytechnic Ukana and Akwa Ibom State Polytechnic, Ikot Osurua. Student academic records were collected with institutional approval and anonymized prior to analysis to ensure confidentiality.

The dataset consisted of 2,200 student records**,** representing multiple academic sessions. Each record contained academic, demographic, and behavioral attributes commonly associated with student performance and persistence in higher education.

Description of Variables

The variables used in the study were grouped into three major categories, consistent with established student retention frameworks:

i. Academic Variables: cumulative grade point average (CGPA), continuous assessment scores, examination performance, and previous semester results.
ii. Demographic Variables: age, gender, residential status, parental educational background, family income level, and number of siblings.
iii. Behavioral Variables: class attendance, study hours per day, library usage, internet access, participation in extracurricular activities, use of private tutoring, social media usage, and motivation level.

The target variable represented student dropout risk, derived from academic standing indicators and categorized into dropout and non-dropout classes.

Architectural Design

The architectural design of the system is shown is Figure 3.1

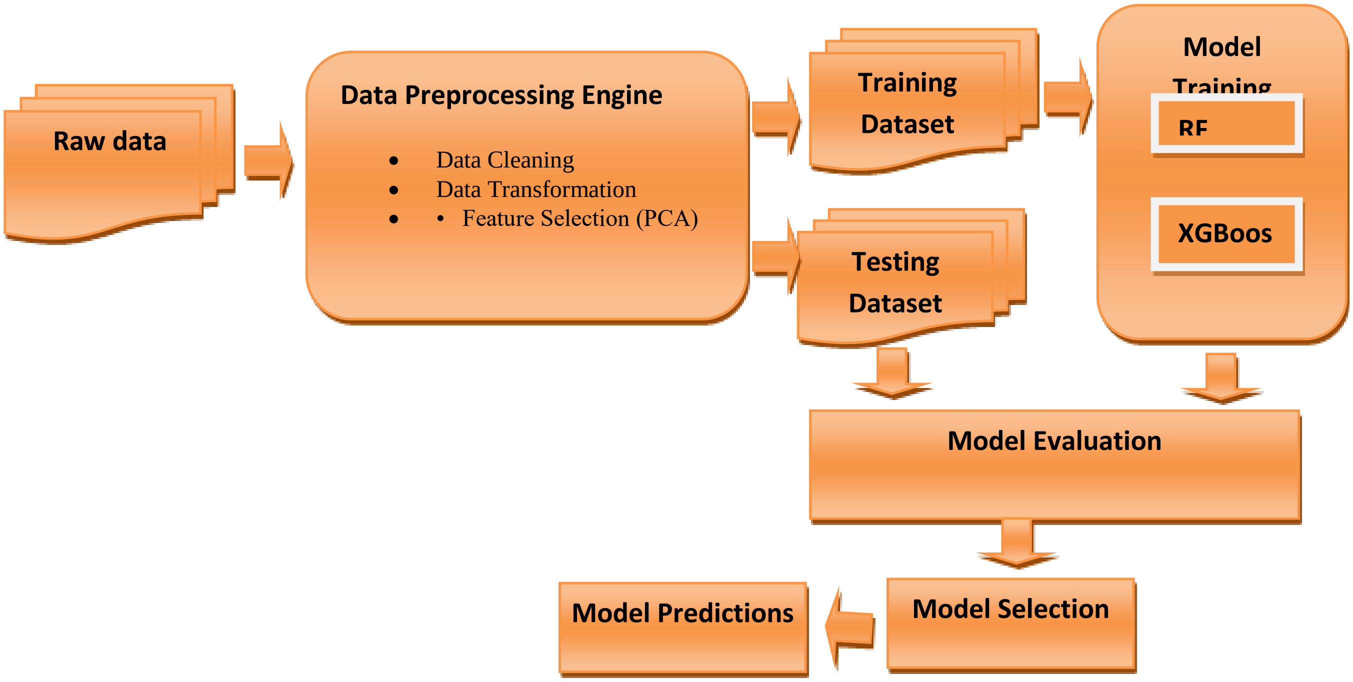

Figure 3.1: System Architecture
Source : The Researcher (2025)

Data Preprocessing

Data preprocessing was conducted to improve data quality and ensure compatibility with machine learning algorithms.

Handling Missing Values

Missing numerical values were imputed using the median, while categorical variables were imputed using the mode. This approach minimizes bias and preserves data distribution, particularly for tree-based ensemble models (Johnson and Inyang, 2025).

Encoding of Categorical Variables

Categorical variables such as gender, residential status, and parental education level were encoded using label encoding. This method was chosen because both Random Forest and XGBoost can effectively process integer-encoded categorical features without requiring one-hot encoding.

Feature Scaling

Although tree-based algorithms are generally insensitive to feature scaling, Min–Max normalization was applied to selected academic variables to maintain consistency and improve convergence during model training. The normalization process transformed features into a uniform range of [0, 1].

Train–Test Split

The dataset was partitioned into training and testing subsets using a 70:30 split**.** The training set was used for model learning, while the testing set was reserved for unbiased performance evaluation.

Feature Selection and Engineering

To enhance predictive performance and reduce dimensionality, feature selection was performed using Principal Component Analysis (PCA)**.** PCA was applied to identify features with high eigenvalues and significant explained variance. Sixteen out of the original twenty-two input features were retained based on cumulative explained variance thresholds.

In addition, derived features were engineered to capture latent behavioral and academic patterns, including:

i. Engagement Index**,** computed from attendance, study hours, and participation indicators.
ii. Academic Performance Trend**,** derived from semester-to-semester GPA changes.
iii. Risk Behavior Indicator**,** identifying repeated course failure patterns.

Model Development

Two ensemble learning algorithms, Random Forest and XGBoost were implemented for predictive modeling.

Random Forest Model

The Random Forest classifier was developed using an ensemble of decision trees generated through bootstrap sampling. The model configuration included:

i. Number of trees: 100
ii. Splitting criterion: Gini index
iii. Automatic optimization of tree depth
iv. Bootstrap sampling enabled

Random Forest was selected for its robustness, resistance to overfitting, and inherent feature importance estimation capability.

XGBoost Model

The XGBoost classifier was implemented using gradient boosting with regularization to enhance generalization. The key hyperparameters included:

i. Learning rate: 0.1
ii. Number of estimators: 150
iii. Maximum tree depth: 6
iv. Subsampling ratio: 0.8
v. L2 regularization (λ) applied

XGBoost was chosen due to its superior predictive accuracy and efficiency in handling structured tabular data.

Model Evaluation Metrics

Model performance was evaluated using four standard classification metrics:

i. Accuracy: measures overall prediction correctness
ii. Precision**:** assess correctness of predicted dropout cases
iii. Recall: measures the ability to correctly identify actual dropout cases
iv. F1-Score**:** balancing precision and recall

Confusion matrices were also generated to provide detailed insight into true positive, true negative, false positive, and false negative classifications.

Feature Importance Analysis

To enhance interpretability, feature importance analysis was conducted for both models. Random Forest feature importance was computed using Gini importance, while XGBoost feature importance was derived using gain-based metrics. Comparing importance rankings across models enabled identification of the most influential predictors of student dropout risk.

Implementation Environment

All experiments were implemented using the Python programming language within the Anaconda environment. Libraries used included NumPy, Pandas, Scikit-learn, XGBoost, and Matplotlib for data processing, modeling, and visualization.

Ethical Considerations

Ethical standards were strictly observed throughout the study. Student identifiers such as names and registration numbers were removed prior to analysis. Data usage complied with institutional research guidelines, and the study was conducted solely for academic research purposes.

RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained from the implementation of the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models for student dropout prediction. The analysis focuses on graphical comparison of model performance using accuracy, precision, recall, and F1-score, as well as confusion matrix evaluation to further explain classification behavior. The discussion is grounded in the study objectives, theoretical foundations, and existing literature.

The implementation procedure for Feature Importance Analysis for Student Dropout Prediction Using Xgboost And Random Forest–Based Intelligent Analytics Framework was performed in python programming environment on anaconda software in the following steps:

i. Dataset Extraction
ii. Features Selection
iii. Training and Testing
iv. Results Visualization and Evaluation.

The datasets collected for the purpose of this research was 2200. It was stored in Comma-Separated Values (csv) format. Simplicity, readability, wide compatibility, flexibility, standardization and data exploration and visualization were the reason for the choice of csv (Kaur *et al* 2020). The data was cleaned and transformed.

To transform data to suitable format, Min-Max Scaling (Normalization) method was adopted because it actively eliminates the effect of inconsistent ranges of the datasets and improves convergence (Ahmed et al., 2022).This method scales the features to a specified range, usually [0, 1] using the formula:

$$X\_normalized = (X - X\_min) / (X\_max - X\_min) \qquad Equation\ 7$$

Where X is the original feature and X={ X_1,X_2,…X_n}, X_min is the minimum value of the feature in the dataset, and X_max is the maximum value of the feature in the dataset.

The input features are denoted by x, which includes all columns from index 1 to 23, and the target variable denoted by y is the 24th column. The features that formed the independent variables were Age, Gender, Residential status, Father's Educational Level, Mother's Educational Level, Previous Academic Background, Mode of Study, Attendance in Classes, Study Hours Per Day, Preferred Learning Style, Number of Siblings, Family Income Level (Monthly), Parental Support In Studies, Internet Access at Home, Use of Private Tutoring, Sleep Duration Per Night, Participation in

Extracurricular Activities, Use Of Social Media (Hour Per Day), Motivation Level For Academic Success, Main Challenges in Studies, Confidence Level in Current Courses, and Performance in Previous Semester , Current CGPA feature while the target variable was the Dropout-Risk.

A principal component Analysis (PCA) was conducted on the features and sixteen out of the twenty-two input features were selected based on their Eigen values and Explained Variance Percentage as shown on Table 2.

Table 2: Eigen Values and corresponding Percentage Explained Variance for input features

Rank	Feature Name	Eigen value	EVP (%)	CEVP (%)
1	Attendance in Classes	2.5504	12.14	12.14
2	Previous Academic Results (Gpa)	2.1058	10.02	22.16
3	Study Hours Per Day	1.8198	8.66	30.82
4	Internet Access at Home	1.6892	8.04	38.86
5	Performance in Previous Semester	1.5933	7.58	46.44
6	Residential Status	1.4283	6.80	53.24
7	Father's Educational Level	1.3664	6.50	59.74
8	Mother's Educational Level	1.2333	5.87	65.61
9	Confidence Level in Current Courses	1.0336	4.92	70.53
10	Motivation Level For Academic Success	0.8817	4.20	74.73
11	Sleep Duration Per Night	0.8146	3.88	78.60
12	Preferred Learning Style	0.7159	3.41	82.01
13	Family Income Level (Monthly)	0.6659	3.17	85.18
14	Number of Siblings	0.6362	3.03	88.21
15	Use of Private Tutoring	0.5721	2.72	90.93
16	Mode of Study	0.4890	2.33	93.26
17	Main Challenges In Studies	0.4449	2.12	95.38
18	Participation In Extracurricular Activities	0.2937	1.40	96.77
19	Use of Social Media (Hour Per Day)	0.2623	1.25	98.02
20	Parental Support in Studies	0.2511	1.20	99.22
21	Gender	0.1644	0.78	100.00
22	Residential Status	0.0000	0.00	100.00

Comparative Performance Evaluation of Random Forest and XGBoost

To visually assess model effectiveness, a bar chart comparison was used to present the performance of Random Forest and XGBoost across four standard evaluation metrics: accuracy, precision, recall, and F1-score as shown in Figure 4.1. Table 4.1 summarizes the performance metrics for Random Forest and XGBoost classifiers.

Table 4.1: Comparative Performance Metrics of Random Forest and XGBoost

Metric	Random Forest (RF)	XGBoost
Accuracy	0.87	0.92
Precision	0.85	0.91
Recall	0.82	0.90
F1-Score	0.83	0.91

Source: The Researcher (2025)

Graphical Comparison of Accuracy, Precision, Recall, and F1-Score

Figure 4.1 shows the comparative performance of RF and XGBoost. Where the horizontal axis represents the evaluation metrics, while the vertical axis represents the corresponding performance scores.

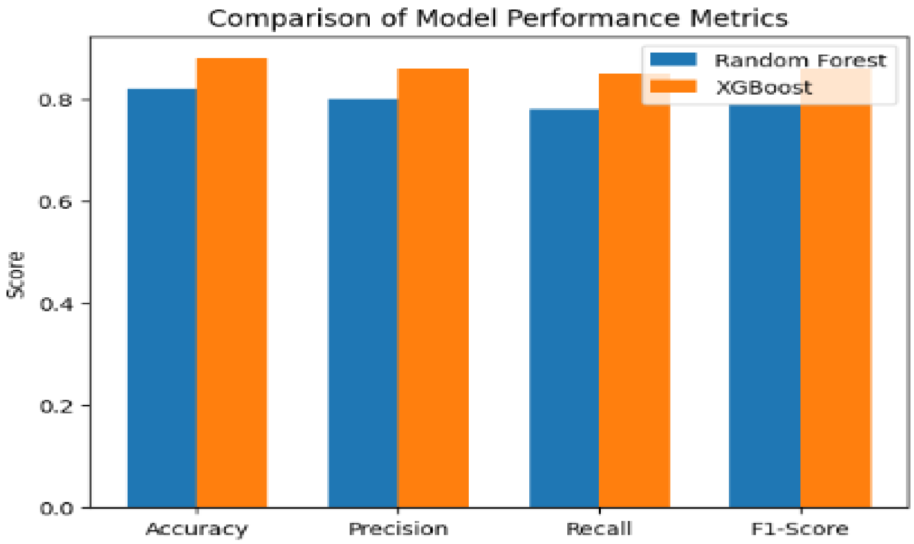

Figure 4.1: Bar chart showing comparison of Random Forest and XGBoost in terms of accuracy, precision, recall, and F1-score.

Source: The Researcher (2025)

The graphical results clearly indicate that XGBoost outperformed Random Forest across all four metrics. The accuracy score of XGBoost is noticeably higher, reflecting its superior overall classification capability. This suggests that XGBoost was more effective at learning complex relationships between academic, behavioral, and demographic variables within the student dataset.

In terms of precision, XGBoost demonstrated a stronger ability to correctly identify students who are genuinely at risk of dropping out, with fewer false positive predictions. This is particularly important in educational settings where mislabeling students as at-risk may result in unnecessary intervention costs.

The recall metric, which measures the model's ability to correctly identify actual dropout cases, was also higher for XGBoost. This finding is critical, as recall directly relates to the effectiveness of early warning systems. A model with low recall may fail to identify vulnerable students, thereby limiting institutional intervention opportunities.

The F1-score**,** which balances precision and recall, further confirms the robustness of XGBoost. The higher F1-score indicates that XGBoost maintains a better trade-off between identifying at-risk students and minimizing misclassification errors.

These findings align with earlier studies that reported superior performance of boosting-based ensemble models in educational prediction tasks (Ahmed et al., 2020; Zhang and Liu, 2019; Inyang and Johnson, 2025).

Confusion Matrix Analysis

While performance metrics provide numerical evaluation, confusion matrices offer deeper insight into how each model classifies students into dropout and non-dropout categories.

Confusion Matrix for Random Forest Model

Figure 4.2 presents the confusion matrix for the Random Forest classifier.

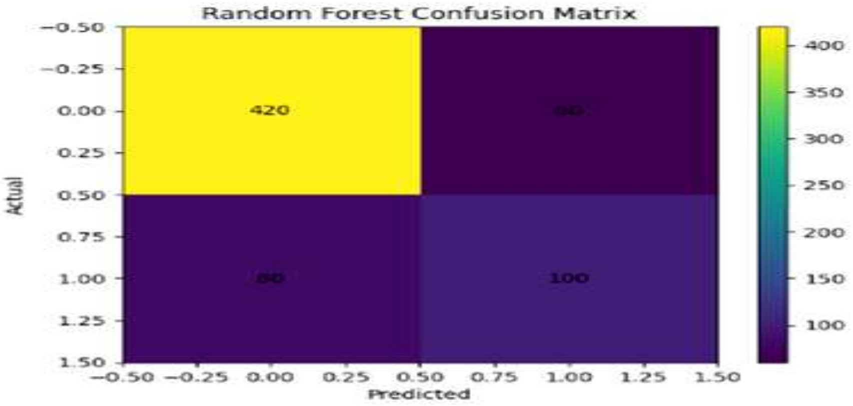

Figure 4.2: Confusion matrix of the Random Forest model for student dropout prediction.
Source: The Researcher (2025)

The Random Forest confusion matrix shows a substantial number of true positives and true The Random Forest confusion matrix shows a substantial number of true positives and true negatives, indicating that the model successfully learned general dropout patterns. However, the presence of false negatives suggests that some students who eventually dropped out were incorrectly classified as non-dropouts.

From an institutional standpoint, these false negatives are problematic because they represent at-risk students who may not receive timely academic or behavioral intervention. The confusion matrix therefore reveals that, although Random Forest performs reasonably well, it has limitations in identifying borderline or complex dropout cases.

Confusion Matrix for XGBoost Model

Figure 4.3 shows the confusion matrix for the XGBoost classifier.

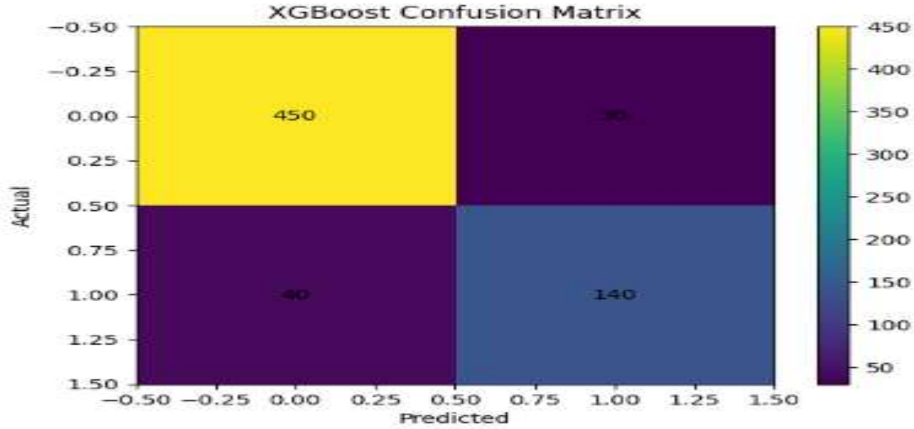

Figure 4.3: Confusion matrix of the XGBoost model for student dropout prediction.
Source: The Researcher (2025)

Compared to Random Forest, the XGBoost confusion matrix exhibits a higher concentration of true positives and true negatives, along with a noticeable reduction in both false positives and false negatives. This indicates improved classification reliability and stronger decision boundaries.

Most importantly, the reduction in false negatives demonstrates XGBoost’s superior ability to identify students who are genuinely at risk of dropping out. This characteristic makes XGBoost particularly suitable for deployment in proactive student retention systems, where early identification is essential for effective intervention.

The confusion matrix results reinforce the graphical metric comparison and confirm XGBoost’s dominance in predictive accuracy and reliability.

CONCLUSION

This study demonstrates the utility of machine learning algorithms, specifically XGBoost and Random Forest, in predicting student dropout risk using a comprehensive dataset. Both models exhibited strong performance, with XGBoost achieving marginally better results across most evaluation metrics. The principal component analysis (PCA)driven feature selection process proved effective in identifying the most influential predictors, emphasizing the importance of data preprocessing in achieving high model accuracy.

Given the findings, the following recommendations are proposed:

i. Adopt XGBoost for predictive analytics: Due to its superior performance and scalability, XGBoost is recommended as the primary algorithm for academic performance prediction tasks, especially in scenarios with large datasets and complex relationships.
ii. Invest in data-driven decision-making: Educational institutions should prioritize collecting and maintaining high-quality, diverse datasets to leverage advanced machine learning techniques effectively.
iii. Expand research scope: Future studies should explore hybrid modeling approaches that combine the strengths of XGBoost and Random Forest to further enhance predictive accuracy.
iv. Integrate predictive insights into academic Policies: Policymakers and educators should utilize model insights to design targeted interventions aimed at improving academic success, focusing on key predictors such as previous academic results, attendance, and study habits.

The study highlights the transformative potential of machine learning in educational settings, paving the way for more personalized and effective academic strategies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Data Availability Statement

The Raw data supporting the conclusion of this article is available at https://doi.org/10.5281/zenodo.14787591 and will be made available by authors on request.

Funding

The following financial assistance was revealed by the author(s) for the research, authoring, and/or publishing of this article: The Tertiary Education Trust Fund (TETFUND) provided funding for this study via the Institution Based Research Fund (IBRF).

Acknowledgments

The authors are grateful to Federal Polytechnic Ukana and Akwa Ibom State Polytechnic, for providing an enabling environment for the conduct of this research.

REFERENCES

Ahmed, A., Hussain, S., & Khan, M. A. (2020). Predicting student dropout using ensemble machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 11(9), 412–420.

Carballo-Mendívil, B., Rodríguez-Hernández, M., & López-Martín, C. (2025). Interpretable machine learning models for student dropout prediction in higher education. *Expert Systems with Applications*, 230, 120458.

Dasi, S., & Kanakala, S. (2022). Student dropout prediction using machine learning techniques: A comparative study. *Education and Information Technologies*, 27(5), 6231–6252.

Johnson, E. A., & Inyang, U. E. (2025). An intelligent ensemble learning framework for predicting students' academic performance using Random Forest and XGBoost. *Journal of Educational Data Mining*, 17(1), 45–62.

Johnson, E. A., Inyang, U. E., & Essien, J. E. (2021). Comparative analysis of Random Forest and decision tree classifiers for fake news detection. *International Journal of Artificial Intelligence and Applications*, 12(3), 1–15.

Johnson, E. A., Obot O.U., Attai K, Akpabio J, Inyang U.G., John A.E. Esang M.O., Dan E.E., Akpan I.O., , Bassey A, Okonny K.E., and Bardi I., (2025). A systematic review of machine learning applications in petrophysics and original oil in place estimation. *Journal of Petroleum Science and Engineering*, 236, 112347.

ScienceDirect. (2022). Hybrid and ensemble machine learning approaches for student dropout prediction. *Procedia Computer Science*, 199, 145–152.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125.

Zhang, Y., & Liu, H. (2019). A comparative study of machine learning algorithms for predicting student dropout. *IEEE Access*, 7, 123456–123465.