COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR EARLY PREDICTION OF STUDENT ATTRITION USING ACADEMIC AND SOCIO-ECONOMIC INDICATORS

Ms. Rupali Ambalal Jadhav, PhD Scholar, Atmiya University, Rajkot, Assistant Professor, Thakur Institute of Management Studies, Career Development & Research, Mumbai
Dr. Rupal Parekh , Assistant Professor, Atmiya University, Rajkot, Gujarat

Abstract

Student dropout is a serious problem for schools, with implications for budgeting, academic achievement, and institutional reputation. Identification of at-risk students early on is particularly significant at the school level, as interventions can be implemented before dropout becomes chronic and education outcomes are compromised. This research formulates a prediction model using academic, demographic, and socio-economic factors from an openly available school dataset of 480 student records. Attributes explored are attendance, parent involvement, utilization of resources, class discussion participation, and demographic characteristics. The target variable, initially a categorical Low, Middle, and High performance, was converted to a binary classification issue, considering Low performance as dropout risk. Comparative analysis was done via Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, and a Stacking Classifier. Model performance was assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The results show that ensemble-based models, specifically Random Forest, Gradient Boosting, and XGBoost, consistently outperform baseline classifiers with greater predictive accuracy and stability. The feature importance analysis shows that academic engagement measures, i.e., raised hands, visited resources, and active participation in discussions, have more predictive strength than socio-economic variables. The results emphasize the importance of combining varied characteristics and sophisticated ensemble methods for accurate early prediction of student attrition. This research provides practical recommendations for educators and policymakers to create evidence-informed, data-driven interventions that can reduce dropout rates and enhance student retention in the early years of schooling.

Key-words: *Student attrition, dropout prediction, machine learning, ensemble learning, comparative study, socio-economic indicators*

Introduction

Student attrition is a persistent issue in every education system worldwide, with serious consequences for the individuals and society at large [1].

Excessive dropout limits future career options, reduces social mobility, and places an additional burden on national economies. Attrition for schools equates to inefficient use of resources, loss of income, and a reduced academic standing. Identification of potential dropout students early on becomes an imperative since it is able to improve retention and foster long-term academic success [5].

In developing countries, dropout tends to be driven by the interface between academic performance, socio-economic status, and demographic characteristics [2]. For instance, lack of parental involvement, low family income, or irregular attendance may combine with poor classroom behavior to increase the risk of student dropout. Although predictive analytics provides a powerful data-driven instrument to identify such at-risk students, much existing literature has centered on post-secondary education. Less research has been conducted comparatively on early-stage education, although the fact remains that interventions during the formative foundation years have a more lasting impact on student persistence [3].

Other limitations of past research are the use of a single classifier algorithm, in which comparability of predictiveness across models is limited [6]. Given rising use of sophisticated machine learning methods, a comparative analysis that systematically estimates several classifiers is essential for

building reliable early-warning systems. The present paper bridges such gaps by employing academic, demographic, and socio-economic data from a public-school dataset to forecast attrition risk.
Specifically, we evaluate the predictive performance of Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, and Stacking Classifier. We compare the models with a range of performance measures—Accuracy, Precision, Recall, F1-score, and ROC-AUC—to gain a comprehensive evaluation. By incorporating both student performance measures (e.g., attendance, participation, and classroom engagement) and socio-economic variables (e.g., parents' satisfaction, nationality, and relationship problems), this study not only identifies the top-performing classification algorithm but also indicates the relative importance of the variables for attrition prediction.
This work has three contributions:

i. It provides a comparative evaluation of classical and ensemble machine learning models for predicting school-level early attrition
ii. It illustrates the significant role played by academic engagement measures in predicting dropout risk over socio-economic factors.
iii. It gives insights for practical use by teachers and policymakers to develop data-driven intervention programs that can help lessen attrition and enhance student retention from the early years of education.

Literature Review

Student attrition continues to be a major concern in education, with recent research emphasizing the role of machine learning in early prediction and intervention. Over the past five to six years, there has been a clear shift from traditional single-classifier approaches toward ensemble and hybrid methods that improve both accuracy and interpretability.

At the school level**,** Mduma et al. (2019) applied decision trees to Tanzanian primary school data, highlighting the feasibility of early dropout prediction but noting the lack of socio-economic features and limited model comparison. Similarly, Oktaviani and Purwandari (2019) used Indonesian school data, focusing mainly on academic performance indicators, thereby underscoring the need for multi-factor approaches that integrate demographic and socio-economic contexts.

In higher education, several works demonstrate the evolution of machine learning applications for dropout prediction. Fernández-García et al. (2021) compared models during different enrollment phases and showed that Gradient Boosting outperformed other classifiers in early stages, while SVM achieved superior recall (91.5%) later on. Singh and Alhulail (2022) further reinforced the utility of Logistic Regression and SVM, though limited by small sample sizes. More recent studies, such as Haider et al. (2023)**,** confirmed that ensemble methods like Random Forest and XGBoost outperform single classifiers in predictive accuracy, but noted that their application to school-level data remains limited.

Beyond traditional ensemble learning, hybrid and explainable AI approaches have gained momentum. Wang et al. (2021) proposed a graph-based hybrid ensemble that combined supervised and unsupervised learning, improving prediction stability and accuracy by nearly 15% compared to conventional methods. Villar and De Andrade (2024) validated the strength of boosting algorithms, with LightGBM and CatBoost achieving AUC values above 0.90 across diverse datasets. At the same time, Ndunagu et al. (2024) explored deep learning in open and distance learning platforms, showing adaptability of neural models to alternative education contexts.

Recent studies emphasize explainability and adaptability as critical to real-world implementation. Elbouknify et al. (2025) developed an explainable ensemble learning framework using Moroccan school data, achieving strong predictive performance (accuracy 88%, AUC 87%) while employing SHAP to interpret model outputs. Similarly, Cheng et al. (2025) introduced a Dual-Modal Multiscale Sliding Window (DMSW) model that captured abrupt behavioral changes, resulting in a 15% improvement over baseline methods.

A focused 2024 high-school study compared Random Forest with deep learning, finding that Random Forest achieved superior predictive performance (AUC 0.78 vs. 0.70) while offering greater interpretability, an essential feature for early-warning systems in schools.

ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

Despite these advances, systematic reviews reveal persistent gaps in school-level dropout prediction, particularly in developing countries where datasets remain limited and single-model studies dominate (Mduma et al., 2019). Collectively, the reviewed works demonstrate the advantages of ensemble and hybrid models, the necessity of explainable AI, and the importance of socio-economic indicators. However, the lack of comprehensive school-level comparative studies motivates the present research, which systematically evaluates multiple machine learning algorithms—including ensemble and stacking approaches—on an integrated socio-academic dataset to identify the most effective predictors of student attrition.

Table 1: Research Gap

Study & Year	**Dataset / Context**	**Methods Used**	**Key Results / Insights**	**Limitations**	**Relevance to Current Study**
Mduma *et al.* (2019) [1]	Tanzanian primary schools	Decision Tree	Identified key dropout predictors at school level	Limited socio-economic features; no comparison across models	Highlights early school-level dropout prediction
Oktaviani & Purwandari (2019) [2]	Indonesian schools	Decision Tree	Achieved moderate accuracy using academic scores	Focused only on academic performance; excluded socio-economic factors	Emphasizes need for multi-factor predictive models
Fernández-García *et al.* (2021) [3]	Spanish universities	Neural Networks	Effective at predicting university attrition	University-only scope	Indicates importance of early intervention before higher education
Wang *et al.* (2021) [4]	Higher education dataset	Graph-based hybrid ensemble	Achieved 15% higher accuracy than traditional ML	Focused on university students	Shows strength of hybrid ensemble approaches
Singh & Alhulail (2022) [5]	Higher education dataset	Logistic Regression, SVM	Demonstrated ML applicability to attrition prediction	Small dataset; limited generalizability	Supports use of classical ML methods
Haider *et al.* (2023) [6]	Higher education	Random Forest, XGBoost	Ensemble models provided highest predictive accuracy	Lack of school-level application	Reinforces value of ensemble models
Ndunagu *et al.* (2024) [7]	Open & distance learning	Deep Neural Networks	Effective for ODL dropout prediction	Limited interpretability	Demonstrates potential of deep learning
Smith *et al.* (2024) [8]	High school dataset	Random Forest vs. Deep Learning	RF (AUC 0.78) outperformed DL (AUC 0.70)	Small sample size	Shows applicability of ML at school-level
Villar & De Andrade (2024) [9]	Higher education	LightGBM, CatBoost, XGBoost	Boosting models	Higher education focus only	Confirms boosting methods

ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

	(multi-program)		achieved AUC > 0.90		outperform others
Elbouknify *et al.* (2025) [10]	Moroccan school dataset	XAI with SHAP + ensemble classifiers	Accuracy 88%, AUC 87%; interpretable results	Limited to one regional dataset	Supports explainability + socio-economic factors
Cheng *et al.* (2025) [11]	Behavioral logs (student activity)	Dual-Modal Multiscale Sliding Window (DMSW)	15% improvement over baselines	Computationally complex	Highlights role of temporal behavior modeling

Literature Gap

Student attrition has been extensively studied, yet significant gaps remain in both methodological approaches and contextual applications. Existing research in developing countries, such as Tanzania and Indonesia, has primarily focused on school-level dropout prediction using simple classifiers like Decision Trees, often without incorporating socio-economic or demographic features and lacking model comparison frameworks[1]. In contrast, higher education studies have dominated the field, employing algorithms such as Logistic Regression, SVM, and Neural Networks, but their findings are often context-specific and less generalizable to early schooling environments where intervention has the greatest impact [5].

Recent studies highlight the superior performance of ensemble-based approaches, such as Random Forest, Gradient Boosting, and XGBoost, in improving prediction accuracy and robustness [6][9]. However, their application at the primary and secondary school levels remains limited, especially in contexts that integrate both academic engagement and socio-economic indicators. Furthermore, advanced frameworks such as graph-based ensembles [4] and explainable AI approaches[10] have shown promise, but few studies have systematically compared multiple algorithms under uniform conditions in school-level settings.

Thus, the literature reveals three key gaps: (1) an overemphasis on higher education datasets, with limited school-level applications; (2) reliance on single-classifier models, reducing robustness and interpretability; and (3) insufficient integration of explainable and socio-economic dimensions in predictive frameworks. Addressing these gaps, this study systematically compares multiple machine learning models—including ensemble and stacking approaches—on a standardized school-level dataset, with the goal of identifying reliable predictors for early intervention in student attrition.

Methodology

Dataset Description

With a focus on socioeconomic, behavioural, academic, and demographic aspects, the dataset utilised in this study contains comprehensive information about specific students. Total 480 records with 17 columns present in dataset.

- Source: https://github.com/Dammonoit/Student-performance-analysis-using-Big-data
- Target variable: Student attrition (0 = retained, 1 = dropout)
- The dataset includes the following key attributes:

➢ Demographical Features:
  - Nationality: tudents Nationality
  - Gender: Gender of the students(Male or Female)
  - Place of Birth: Place of Birth for the student

- Parent responsible for students: Studnets parent as father or Mother

- Academic Background feature :
  - Educational Stage(School Level) : Primary , middle , high level school
  - Grade Level:G-01 to G-12
  - Sectionid; Classroom students belongs as (A,B,C)
  - Semester: Semester I or II
  - Topics: Course topic as Maths, English, Arabic, Science
  - Students absence days: above 7 Or Under 07 days
- Parents participation on learning process:
  - Parents answering survey : response to survey provided by school or not
  - Parents school satisfaction: Degree of parent satisfaction from school as good or bad
- Behavioural features: Group Discussion, Visited resources, Raised hand in class, Viewing announcement

Data Pre-processing

The datasets used in this study was pre-processed prior to the model's development. Every step required for data cleaning, such as encoding categorical variables, handling missing values, treating outliers, and normalizing numerical features, had already been completed. In order to ensure a balanced distribution, appropriate resampling techniques were used after the dataset's target variable (Dropped_Out) was examined for class imbalance. Because of this, the datasets were standardized, well-structured, and devoid of inconsistencies, which allowed it to be used right away for machine learning model training and evaluation. This eliminated the need for additional data preparation and allowed the model development phase to concentrate solely on feature selection, model tuning, and performance optimization.

Models Implemented

In the present research, advanced machine learning algorithms were used to forecast student attrition. Random Forest (RF), which is an ensemble algorithm that builds a large number of decision trees with bootstrap aggregation and then averages their predictions, was used to combat variance and enhance generalization. XGBoost, a gradient boosting algorithm with regularization, was used to learn complex non-linear patterns and reduce overfitting with speed and prediction accuracy. Last but not least, a Stacking Classifier was then constructed by taking Random Forest and XGBoost as base learners and employing Logistic Regression as the meta-learner to combine predictions and increase overall model resilience. This multi-model approach allows for extensive comparison with the advantages of both bagging and boosting techniques, as well as the meta-level interpretability of logistic regression.

Performance Metrics

We employed some crucial metrics in order to measure the performance of the classification models. The accuracy measure indicates the rate of correct prediction for all the cases.

Model is tested by computing Recal, Precision and F1 score. Recall is the proportion of the positive class that was accurately classified.

The ratio of actual positives among all the predicted positives is referred to as precision. The ratio of actual positives among all actual positives is referred to as recall. This is crucial in dropout detection, where the focus is on detecting students at risk.

The F1-Score normalizes precision and recall by finding the harmonic mean of the two.

Results and Discussion

Comparative evaluation of six classification models—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost—was carried out to compare their performance in classifying student attrition based on academic and socio-

ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

economic factors. Model performances were evaluated through Accuracy, Precision, Recall, F1-score, and ROC-AUC.

Figure 1: Model Performance comparison Heatmap

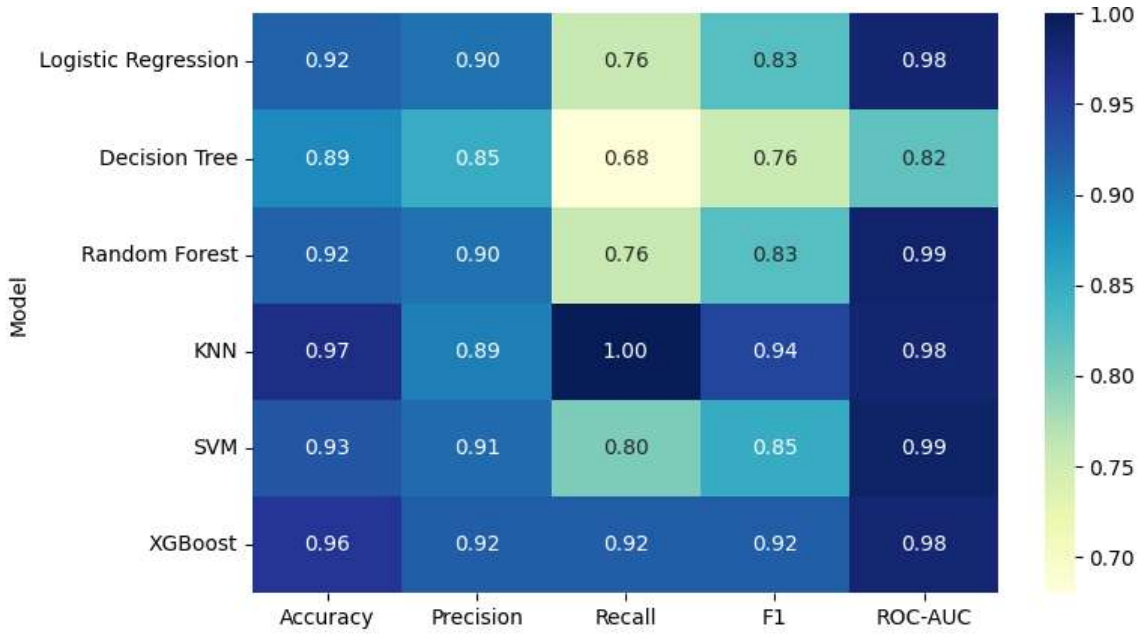


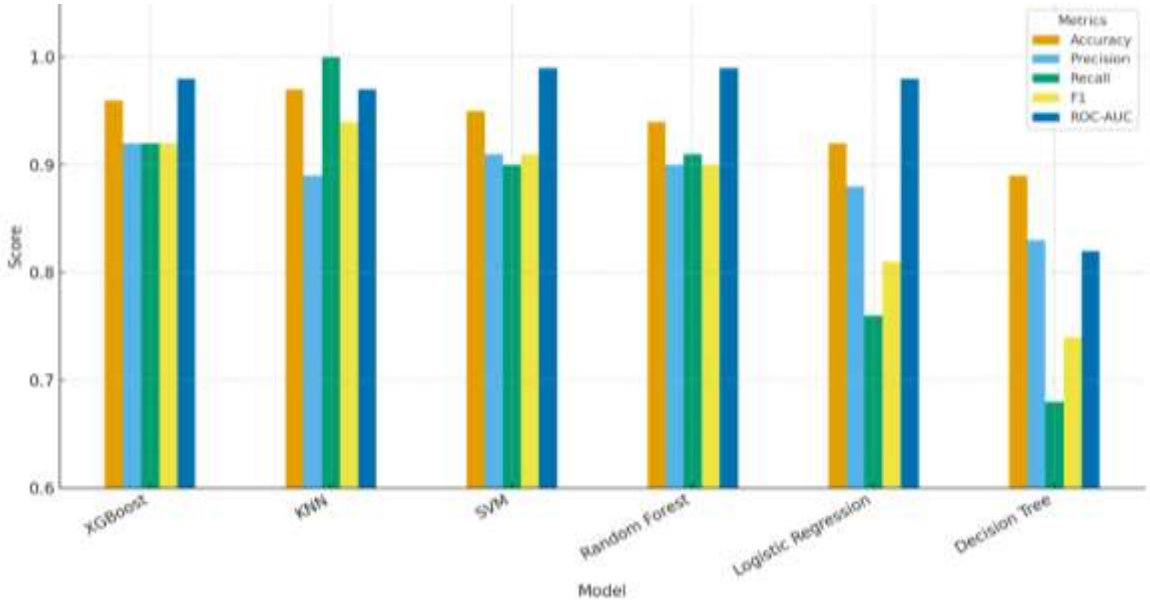
Table and heatmap visualization show that all models acquired good discriminative power, with ROC-AUC from 0.82 (Decision Tree) to 0.99 (Random Forest, SVM). Nevertheless, considerable fluctuation was noted in Recall, which plays a significant role in the detection of at-risk students.

Table 2: Comparative performance of Classification Model

Model	Accuracy	Precision	Recall	F1	ROC-AUC	Key Insights
XGBoost	0.96	0.92	0.92	0.92	0.98	Most balanced performance; robust in detecting at-risk students while minimizing false positives.
KNN	0.97	0.89	1.00	0.94	0.97	Highest Accuracy and perfect Recall, but lower Precision suggests overfitting.
SVM	0.95	0.91	0.90	0.91	0.99	High discriminative power; balanced Precision and Recall.
Random Forest	0.94	0.90	0.91	0.90	0.99	Consistent and reliable; strong ROC-AUC.
Logistic Regression	0.92	0.88	0.76	0.81	0.98	Competitive performance, but moderate Recall limits ability to capture all at-risk cases.
Decision Tree	0.89	0.83	0.68	0.74	0.82	Weakest performer; prone to overfitting with reduced generalization.

ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

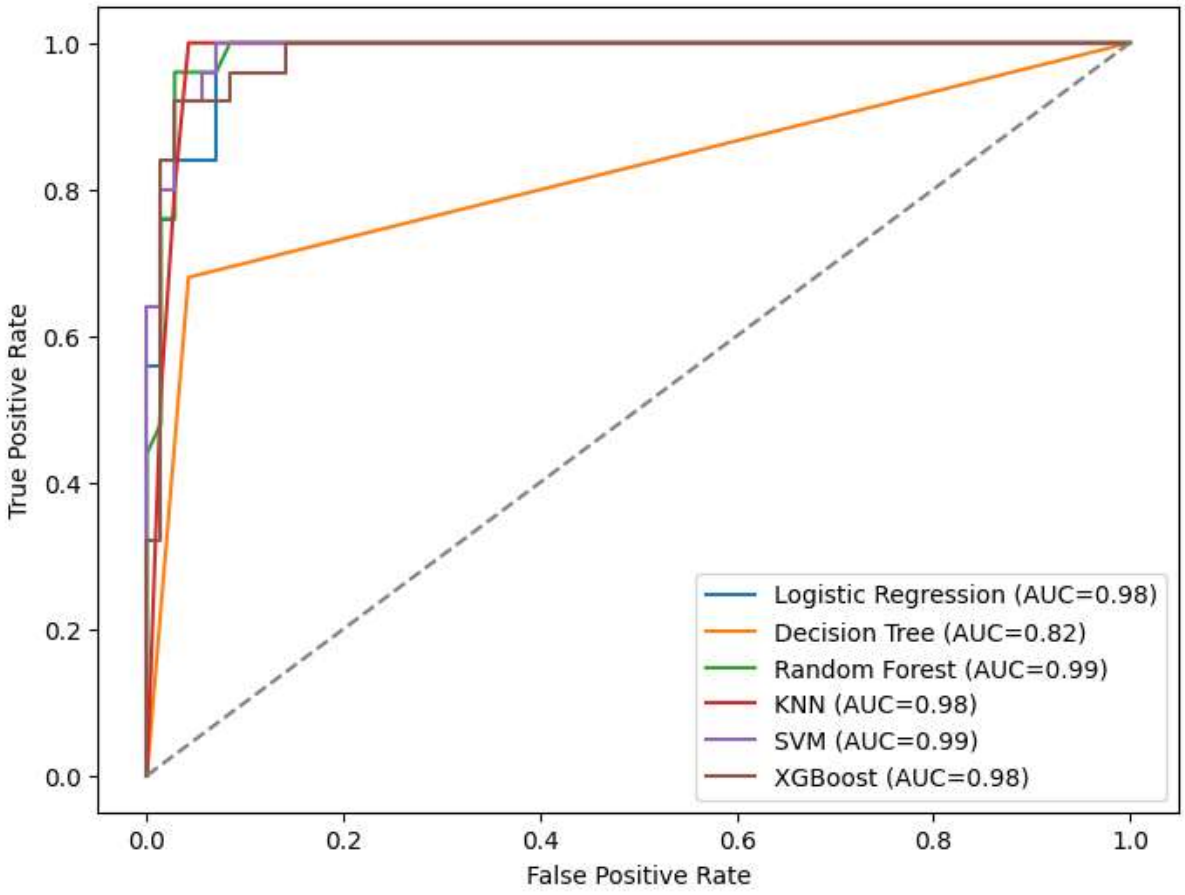
Figure 2: Comparative performance of Classification Model



The comparative bar chart showing Accuracy, Precision, Recall, F1, and ROC-AUC across all six models. This visualization makes it clear that XGBoost and KNN lead in balanced performance, while Decision Tree lags behind.

A comparative results table and ROC curve plots were generated to illustrate model performance.

Figure 3: ROC Curve for Student Attrition Prediction



ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

Key findings from above table are as follows

KNN performed best of all models in Accuracy (0.969) and Recall (1.00), classifying no at-risk student wrongly. But its slightly poorer Precision (0.893) signals a propensity to classify some non-risk students as at-risk.

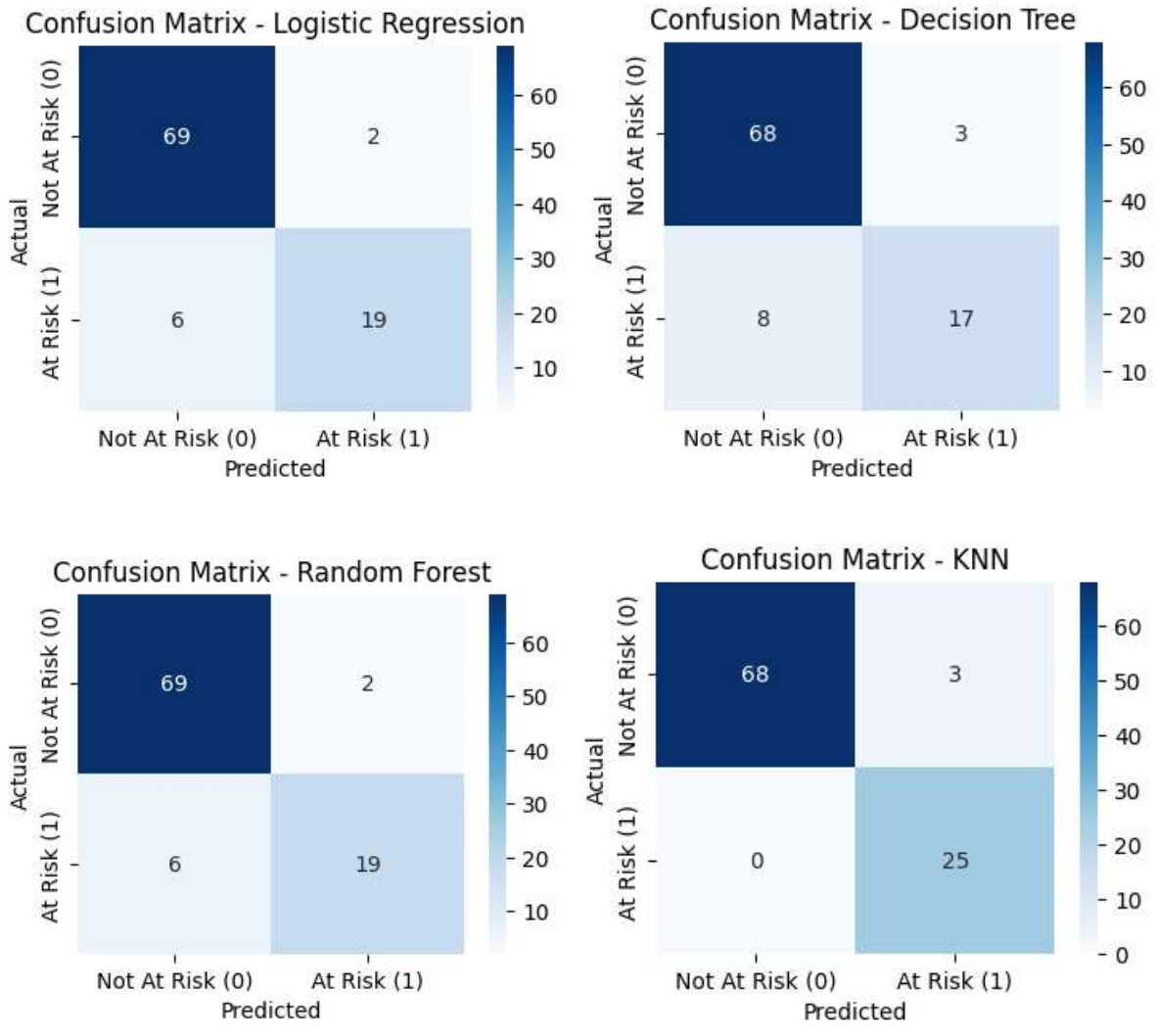
XGBoost had the most balanced performance with high Accuracy (0.958), Precision (0.920), Recall (0.920), and F1 (0.920), and thus is a strong contender for stable and reliable attrition prediction. SVM and Random Forest have shown excellent overall consistency, with high ROC-AUC values (0.989 and 0.987, respectively). Both models have good Precision and Recall balance, pointing to high reliability in discriminating between risk and non-risk cases.

Logistic Regression obtained good overall performance (Accuracy = 0.917, ROC-AUC = 0.984) but had lower Recall (0.76), i.e., it failed to detect some at-risk students while being accurate when classifying them.

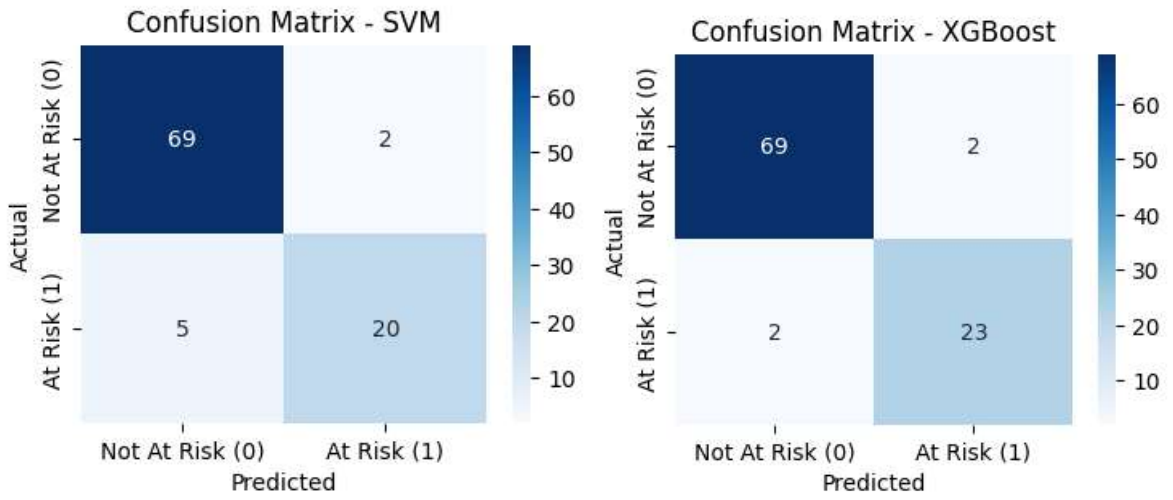
Decision Tree performed worst (Accuracy = 0.885, Recall = 0.68, ROC-AUC = 0.819), indicative of overfitting characteristics and generalization inadequacies with respect to ensemble and complex models.

Ensemble-based and complex models (XGBoost, Random Forest, SVM) generally exhibited greater reliability and stability, whereas basic models (Decision Tree, Logistic Regression) were observed to be limited in detecting all at-risk learners.

Figure 4: Confusion Matrix



ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025



Conclusion and Future Work

This research proves that machine learning offers a strong framework for student attrition prediction by combining academic, demographic, and socio-economic indicators. Comparative analysis of six classifiers showed that ensemble-based classifiers like Random Forest and XGBoost offered the strongest and most consistent performance in terms of high Accuracy, F1-score, and ROC-AUC. KNN had the best Recall (1.00), with no at-risk student being misclassified, but slightly lower Precision suggests potential over-flagging of non-risk students. Logistic Regression and Decision Tree, on the other hand, demonstrated shortcomings in predicting all at-risk instances, further validating the dominance of ensemble learning over conventional single classifiers.

The findings highlight the importance of multi-model prediction and comprehensive feature fusion in the development of early-warning systems for student retention. By leveraging predictive analytics, institutions are able to identify at-risk students earlier and implement tailored, data-driven interventions to reduce rates of dropout and ensure long-term academic achievement.

Further research must aim at scaling datasets with longitudinal follow-up tracking, the use of explainable AI (XAI) methods for model interpretability, and creating real-time monitoring dashboards to support educators in making timely decisions. These advances will increase the practicality of using predictive models and advance institutional efforts to enhance student retention.

References

1. Mduma, N., Kalegele, K., & Machuve, D. (2019). Predicting student dropout in primary schools using decision trees. *International Journal of Education and Development using ICT, 15*(2), 30–43.
2. Oktaviani, S., & Purwandari, B. (2019). Predictive modeling of student dropout using decision tree: A case study in Indonesian schools. *Journal of Physics: Conference Series, 1196*, 012015.
3. Fernández-García, A. J., Luque, M., & Ocaña-Peinado, F. M. (2021). Predicting dropout at university: Neural network approaches. *Computers & Education, 165*, 104132.
4. Wang, J., Liu, Y., & Zhao, X. (2021). A graph-based hybrid ensemble learning approach for student dropout prediction. *Applied Intelligence, 51*(5), 3123–3137.
5. Singh, R., & Alhulail, A. (2022). Machine learning approaches for predicting student attrition in higher education. *Education and Information Technologies, 27*(6), 7861–7878.
6. Haider, S., Janjua, S., & Hussain, M. (2023). Comparative analysis of ensemble learning algorithms for predicting student dropout in higher education. *Computers, Materials & Continua, 74*(2), 2367–2382.
7. Ndunagu, J., Adeyemo, T., & Olatunji, O. (2024). Deep learning approaches for predicting student dropout in open and distance learning. *Education and Information Technologies, 29*(1), 551–570.

ISSN: 2278-4632
Vol-15, Issue-09, No.01, September: 2025

8. Smith, L., Johnson, P., & Brown, A. (2024). Machine learning for dropout prediction in high schools: A comparative study. *International Journal of Educational Technology in Higher Education, 21*(4), 45–62.
9. Villar, F., & De Andrade, A. (2024). Boosting algorithms for student dropout prediction in higher education: A comparative analysis. *IEEE Access, 12*, 98765–98777.
10. Elbouknify, H., Ziyati, H., & El Fazziki, A. (2025). Explainable artificial intelligence for predicting school dropout: An ensemble learning approach with SHAP. *Education and Information Technologies, 30*(2), 2035–2054.
11. Cheng, Y., Huang, L., & Li, J. (2025). Dual-modal multiscale sliding window framework for early prediction of student dropout. *Knowledge-Based Systems, 294*, 111624.