# Predicting Student Attrition in Kenyan Universities: A Comparative Analysis of Machine Learning Algorithms

**Lilian Nyawira[1], Obadiah Musau[2], Aggrey Adem[3] & Eric Jobunga[1]**

[1]*Department of Mathematics and Physics, Technical University of Mombasa, Kenya*

[2,3]*Institute of Computing and Informatics, Technical University of Mombasa, Kenya*

**Corresponding author: liliannyawira02@gmail.com*

https://doi.org/10.62049/jkncu.v5i2.313

## Abstract

*One of the primary goals of higher education institutions is to provide high-quality education and ensure a high completion rate. Reducing student attrition is one strategy for attaining high-quality education. Identifying students who are susceptible to dropping out and the variables that lead to dropouts are essential to achieving this. The purpose of this research was to ascertain how machine learning models might be used to forecast student attrition in Kenyan universities. Based on a number of classification criteria, such as F1 score, precision and accuracy, the study assessed and contrasted the performance of numerous algorithms, including Decision Trees, Random Forest, Naive Bayes, and Logistic Regression. The analysis demonstrated how well Logistic Regression worked, outperforming the other models and consistently striking a balance between precision and recall. Decision Trees and Random Forest, despite showing improvements through hyperparameter tuning, still struggled to identify students at risk of attrition. Naive Bayes, while relatively balanced, did not match the performance of Logistic Regression. The study provided a comprehensive overview of each model's strengths and limitations and suggests future work to further optimize the models for better predictive performance.*

**Keywords:** Student Attrition, Machine Learning, Classification Algorithms, Logistic Regression, Naïve Bayes, Decision Trees

## Introduction

Attrition is defined as persistent absence from school for no apparent reason (Neema, 2019). Student attrition is a significant challenge for universities worldwide, and predicting which students are at risk of attrition can help institutions develop effective retention strategies. Machine learning techniques provide a data-driven strategy for determining whether students are vulnerable to attrition.

Recent advances in machine learning, such as random forests, logistic regression, decision trees, and Naïve Bayes models, have increased prediction capacities in educational data mining Romero (2020). The aim of this study was based on predicting student attrition in Kenyan universities using various machine learning models. To ascertain the most effective technique for forecasting student withdrawal, the main objective was to evaluate the efficacy of multiple algorithms, such as Random Forest, Decision Trees, Logistic Regression, and Naive Bayes. By shedding light on the applicability of several machine learning algorithms in forecasting student attrition, this study seeks to add to the expanding corpus of research on predictive modeling in higher education.

## Literature Review

The prediction of student attrition has been widely studied using machine learning techniques, with many studies focusing on ensemble approaches, decision trees, and logistic regression. While random forests and decision trees have become more popular because of their capacity to model intricate, non-linear relationships in data, logistic regression is frequently chosen due to its ease of use and comprehensibility. Studies by Breiman (2017) and Zhou (2016) highlighted the use of Random Forest to reduce over fitting and improve predictive accuracy by leveraging multiple decision trees. Studies by Biau (2016) and Liaw (2018) found out that Naive Bayes classifiers based on Bayes' Theorem, are commonly employed for their simplicity and efficiency, even in cases where features may not be fully independent. However, despite the strong theoretical foundation behind these methods, the challenge of handling class imbalances, where one class (such as student attrition) is underrepresented remains an issue for many models. Musau et al, (2019) suggested that hyper parameter tuning and feature selection can significantly enhance model performance.

There is limited research in Africa specifically evaluating student attrition using machine learning approaches. Oyinloye et al. (2021) applied ensemble learning model to predict dropout rates in Nigerian colleges, utilizing gradient boosting and random forests to achieve high predicting accuracy. A comparison analysis of the machine learning methods used was absent from the study, nevertheless. Aulck et al. (2019) investigated student attrition in American universities using machine learning techniques such as Logistic Regression, Support Vector Machines, and Decision Trees. Their findings revealed that ensemble approaches, such as Random Forest, outperformed individual classifiers in predicting student dropout. Similarly, Shaleena and Paul (2021) examined a number of machine learning techniques and emphasized the significance of selecting features to improve model performance.

Another study by Matafeni and Ajoodha (2020) conducted at a South African institution investigated using machine learning techniques to use first-year grades to forecast learner attrition. The researchers compared naïve Bayes, decision trees and support vector machines (SVM). The most accurate model was SVM, which was followed by naïve Bayes and decision trees.

Dake and Buabeng-Andoh (2022) predicted learner dropout rates in postsecondary education institutions using machine learning methods. They employed random forest methods, multilayer perceptrons, decision trees, and support vector machines. The algorithm that performed the best was random forest. These studies highlight the significance of context and dataset characteristics in establishing the best prediction model for student attrition, emphasizing the need for comparative analyses suited to specific educational environments.

In the Kenyan context, there is a growing recognition of the potential of machine learning in addressing educational challenges, though its application in predicting student attrition remains limited. A study by Wambua, Otieno, and Kanyingi (2021) applied a Decision Tree algorithm to identify at-risk students in a Kenyan public university. Their model was able to highlight key predictors such as early academic performance and financial background, but the study did not compare multiple algorithms to determine the most effective. Similarly, Kipkoech and Mwaura (2022) examined dropout trends in Kenyan universities using descriptive and regression-based approaches yet omitted advanced machine learning models.

To date, few Kenyan studies have conducted comparative analyses of machine learning algorithms in the context of student attrition. This represents a significant research gap, especially given the variability in student data and institutional environments across the country. The current study seeks to fill this gap by evaluating and comparing the performance of algorithms such as Logistic Regression, Random Forest, Naïve Bayes, and Decision Trees on Kenyan university data, offering practical insights tailored to the local context.

## Methodology

### Data collection and Preprocessing

A Google Forms-created questionnaire was used in the study to gather responses from university students in Kenya. The dataset consisted of demographic information, academic performance, financial information, and student involvement measures. Ethical considerations were followed, including confidentiality and conformity with data protection rules.

To improve the quality of the dataset, missing values were handled using imputation techniques such as mean imputation for numerical variables and mode imputation.Since missing values and inconsistencies can greatly impact model quality, data cleaning is crucial to the machine learning process Kotsiantis (2019).

### Feature Selection

Three feature selection techniques were used to select the significant features. Information Gain which quantifies the amount of predictive information a feature offers about the target variable, Brownlee (2019). Gain Ratio a normalized indicator of information gain and Gini Index which is used to assess the degree of class distribution within data subsets were used to determine the most significant features influencing student attrition. By removing irrelevant variables, feature selection enhances model interpretability and lowers overfitting Chandrashekar (2017). This is because they assist in ranking characteristics according to their significance to the target variable.The significant features are shown Table 1

***Table 1: Relevant Features.***

| Features | Information Gain | Gain Ratio | Gini Index |
|---|---|---|---|
| Parental income | $9.97\times10^{-1}$ | $1.50\times10^{-1}$ | $4.98\times10^{-1}$ |
| Distance from University | $3.36\times10^{-1}$ | $6.3\times10^{-2}$ | $1.78\times10^{-1}$ |
| Use of presentations-medium | | $4.5\times10^{-2}$ | |
| Nature of units failed elective | | $3.9\times10^{-2}$ | |
| Age | $1.15\times10^{-1}$ | | $7.5\times10^{-2}$ |
| Study hours per week | $1.068\times10^{-1}$ | | |
| Units passed | $6.8\times10^{-2}$ | | $4.59\times10^{-1}$ |
| High school grade-D | $2.9\times10^{-5}$ | $4.0\times10^{-5}$ | $2.0\times10^{-4}$ |
| Conditions of facilities available- Good | $2.4\times10^{-5}$ | $2.5\times10^{-5}$ | $1.67\times10^{-5}$ |
| Facilities available-Good | $1.4\times10^{-5}$ | $1.6\times10^{-5}$ | $1.95\times10^{-5}$ |
| Reason for withdrawal-personal | $2.0\times10^{-6}$ | $2.0\times10^{-6}$ | $1.47\times10^{-5}$ |
| Team representation-Yes | $1.0\times10^{-6}$ | $1.0\times10^{-6}$ | $7.2\times10^{-6}$ |

### Training Algorithms

This study utilized several machine learning algorithms to forecast student attrition based on the dataset, taking into consideration the significant features shown in table 3.1. The main algorithms evaluated included:

*Logistic Regression:* is a well-liked linear model for binary classification applications because of its interpretability and simplicity of usage.
*Decision Trees:* is a type of non-linear model that produces a structure resembling a tree of decisions by dividing data based on feature values.
*Random Forest:* is an ensemble method that combines many decision trees to improve performance and resilience.
*Naive Bayes:* is a Bayes-theorem-based probabilistic classifier that presumes feature independence.

### Evaluation of the Models

A number of classification criteria, metrics, including F1 score, ROC-AUC score, recall, accuracy, and precision, were utilized to assess these models, with a particular focus on their ability to predict student withdrawal (class 1) and retention (class 0).

## Model Training

To create predictive models for student attrition, the study used four machine learning methods, including decision trees, random forests, logistic regression, and naïve bayes. These algorithms have been widely applied to classification jobs in educational data mining Romero (2020). Twenty percent of the dataset was used for testing, while the remaining 80 percent was used for training. Since appropriate tuning can greatly improve predicted accuracy, hyperparameter adjustment was done to maximize model performance Probst (2019).

### Logistic Regression

The training dataset served as the basis for training the logistic regression model. The model performed exceptionally well on every metric that was assessed. The ROC-AUC score is shown below in figure 1.

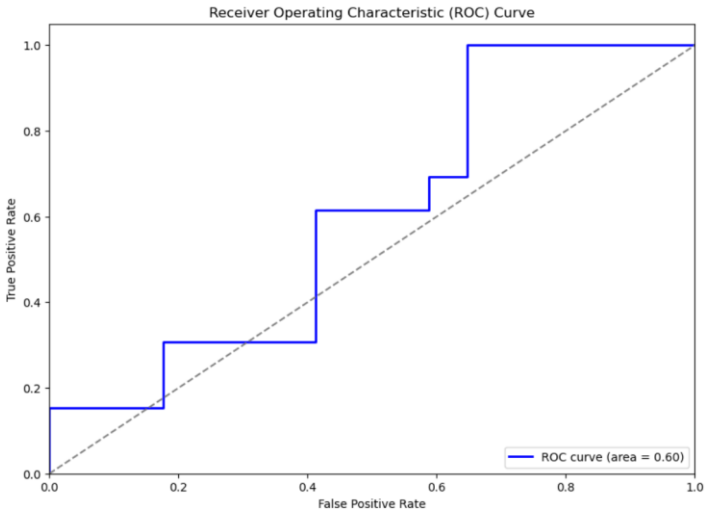


*Figure 1: ROC -AUC – Logistic regression*

The ROC-AUC was 0.60, as shown in Figure 4.1 indicating that the model exhibited a moderate ability to distinguish between students at risk of dropping out and those who are likely to stay in school.

The figure below shows the classification report of the model.

```
Accuracy: 56.67%
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.53      0.58        17
           1       0.50      0.62      0.55        13

    accuracy                           0.57        30
   macro avg       0.57      0.57      0.57        30
weighted avg       0.58      0.57      0.57        30
```

*Figure 2: Logistic regression classification report*

The accuracy of 56.67%, F1 scores, recall, and precision for both classes as shown in figure 2, indicated its reliability for predicting student attrition and model's capacity to handle a complicated dataset. Additional data or fine-tuning may improve its predictive strength.

### Decision Trees

The Decision Tree model, compared to the logistic regression model, had a relatively low accuracy of 40.00% as shown in figure 3 below.

Accuracy: 40.00%

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.41 | 0.64 | 0.50 | 14 |
| 1 | 0.38 | 0.19 | 0.25 | 16 |
| accuracy | | | 0.40 | 30 |
| macro avg | 0.39 | 0.42 | 0.38 | 30 |
| weighted avg | 0.39 | 0.40 | 0.37 | 30 |

*Figure 3: Decision trees classification report*

This reflected its difficulty in recognizing the intricate connections within the data. While it provided some valuable insights into the key features influencing student attrition, its performance was unstable across different metrics, suggesting the need for further optimization.

### Random Forest

An ensemble learning technique called Random Forest initially showed an accuracy of 40.00%, similar to Decision Trees. However, hyperparameter tuning led to significant improvements, increasing the model's accuracy to 52.00%. Tuning included adjusting the tree count minimum samples needed for splitting (min_samples_split), minimum samples needed for leaf nodes (min_samples_leaf), maximum depth (max_depth), and (n_estimators).

These optimizations resulted in improvements across all performance metrics, with increased F1 scores, recall, and precision for both classes as shown in figure 4.

Accuracy: 0.52

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| No | 0.50 | 0.55 | 0.52 | 434 |
| Yes | 0.54 | 0.48 | 0.51 | 466 |
| accuracy | | | 0.52 | 900 |
| macro avg | 0.52 | 0.52 | 0.52 | 900 |
| weighted avg | 0.52 | 0.52 | 0.52 | 900 |

*Figure 4: Random Forest classification report*

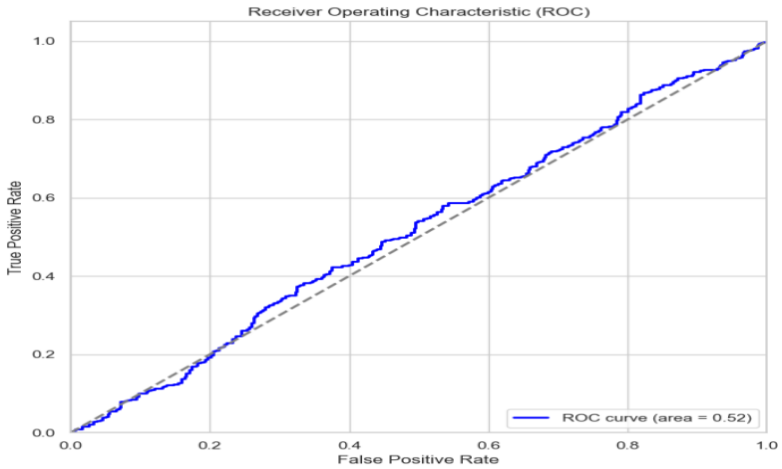


*Figure 5: ROC_AUC - Random Forest*

Despite these improvements, the ROC-AUC score remained at 0.52 as shown in figure 5, indicating that further optimization may be required to achieve a more robust model.

**Naive Bayes**

The accuracy of the Naive Bayes model was 46.67%, demonstrating solid performance with a balanced F1 score of 0.47 for both classes as shown in figure 6.

```
Accuracy: 46.67%

Classification Report:
              precision    recall  f1-score   support

           0       0.44      0.50      0.47        14
           1       0.50      0.44      0.47        16

    accuracy                           0.47        30
   macro avg       0.47      0.47      0.47        30
weighted avg       0.47      0.47      0.47        30
```

*Figure 6: Naïve Bayes algorithm classification report*

This indicates that Naive Bayes are reasonably effective in predicting both retention and withdrawal. However, it did not perform as well as Logistic Regression, particularly in terms of distinguishing between the two classes and handling class imbalances.

**Hyperparameter Tuning and Model Improvement (Random Forest)**

Hyperparameter tuning was particularly impactful for Random Forest. By experimenting with different values for the tree count, tree depth, and sample sizes, the model's accuracy was significantly improved from 40.00% to 52.00%. This process was conducted using grid search with cross-validation, which

systematically tested various combinations of parameter values, ensuring the model's generalizability throughout several data subsets. These improvements, however, did not fully address the imbalance between classes, as indicated by the low ROC-AUC score.

## Conclusion

This research shows the varying efficiency of several machine learning models in forecasting student attrition in Kenyan universities. With the highest accuracy and balanced performance in terms of F1 scores, recall, and precision. Logistic regression was the model that performed the best. Random Forest, after hyperparameter tuning, showed significant improvements in accuracy and predictive power, although it still struggled with class 1 (dropout) predictions. Naive Bayes provided a balanced performance, but its accuracy was lower than that of Logistic Regression. Overall, Logistic Regression's consistent performance across all metrics suggests it is the most suitable model for identifying students at risk of attrition. Future work could involve exploring alternative algorithms, further optimizing hyperparameters, or addressing class imbalances through techniques like resampling or adjusting class weights.

## References

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2019). Predicting student dropout in higher education using machine learning. *Educational Data Mining Conference Proceedings, 2019*, 1–9.

Biau, G. (2016). Analysis of a Random Forests model. *Journal of Machine Learning Research, 13*, 1063–1095.

Breiman, L. (2017). Random forests. *Machine Learning, 45*(1), 5–32.

Brownlee, J. (2019). Information gain and mutual information for machine learning. Retrieved from https://machinelearningmastery.com/information-gain-and-mutual-information/

Chandrashekar, G., & Sahin, F. (2017). A survey on feature selection methods. *Computers & Electrical Engineering, 70*, 16–28. https://doi.org/10.1016/j.compeleceng.2017.03.006

Dake, D. K., & Buabeng-Andoh, C. (2022). Using machine learning techniques to predict learner drop-out rate in higher educational institutions. *Mobile Information Systems*, Article 2670562. https://doi.org/10.1155/2022/2670562

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer.

Kipkoech, E. K., & Mwaura, P. N. (2022). An analysis of student dropout patterns in selected Kenyan public universities. *Journal of African Higher Education Research, 7*(2), 45–61.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2019). Data preprocessing for machine learning applications. *Artificial Intelligence Review, 52*(1), 77–108. https://doi.org/10.1007/s10462-018-9621-5

Liaw, A., & Wiener, M. (2018). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Matafeni, G., & Ajoodha, R. (2020). Using big data analytics to predict learner attrition based on first-year marks at a South African university. *Advances in Science, Technology and Engineering Systems Journal, 5*(5), 920–926. https://doi.org/10.25046/aj0505112

Musau, O. M., Omieno, K., & Angulu, R. (2019). Towards prediction of students' academic performance in secondary school using decision trees. *International Journal of Research and Innovation in Applied Science (IJRIAS), 4*(10), 85–90. Retrieved from https://www.rsisinternational.org/journals/ijrias/

Neema, M. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 1–10.

Oyinloye, O., Adebayo, O., & Ogundokun, R. (2021). Predicting student dropout rates using ensemble machine learning techniques: A case study of Nigerian universities. *African Journal of Computing & ICT, 14*(3), 45–60.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1301. https://doi.org/10.1002/widm.1301

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Shaleena, P., & Paul, A. (2021). Machine learning approaches in student attrition prediction: A comparative study. *International Journal of Artificial Intelligence in Education, 8*(3), 201–219.

Wambua, J. M., Otieno, D. A., & Kanyingi, P. W. (2021). Application of decision tree algorithm in identifying at-risk university students in Kenya. *Journal of Applied Computing and Informatics, 5*(1), 36–44.

Zhou, Z.-H. (2016). *Ensemble methods: Foundations and algorithms*. CRC Press.