

Methodology Article

Machine Learning Model for Prediction of Student Attrition in E-learning Environment: Research Methodology

Musyimi Samuel Muthama^{*} , Fullgence Mwakondo , Kennedy Khadullo, Kevin Tole 

Institute of Computing and Informatics, Technical University of Mombasa, Mombasa, Kenya

Abstract

This paper presents a novel machine learning methodology for predicting student attrition in eLearning environments. Recognizing the limitations of traditional approaches, research exploring the power of ensemble machine learning, combining the strengths of Naïve Bayes, Gradient Boosting, and Random Forest algorithms. To further enhance predictive accuracy based on machine learning, integrating Simulated Annealing for parameter optimization and validation, allowing for fine-tuning of each individual model within the ensemble. An investigation of why students in Kenya public universities dropout from particular course, early identification and mitigation procedures of students attrition. The ensemble weights are iteratively adjusted and optimized to create a robust predictive machine learning model. This paper allows the machine learning model to learn complex patterns within the data that contribute to student's attrition identification. Using a mixed method research design for optimal predictive machine learning in student attrition identification offers a robust approach to understanding and addressing the multifaceted issue of student dropout. Both quantitative and qualitative methods, researchers can develop more accurate, interpretable, and actionable models, ultimately leading to more effective interventions and improved student retention rates. Research validate the proposed framework using real-world eLearning datasets, comparing its performance against standalone models. The results demonstrate the effectiveness of combining ensemble learning with optimization techniques, highlighting the potential for improved precision in identifying at-risk students. This methodology contributes to the field of educational data mining by pioneering the use of Simulated Annealing for attrition prediction, offering a scalable solution for institutions to proactively support student retention and improve eLearning outcomes.

Keywords

Students Attrition Prediction, eLearning, Optimization, Machine Learning, Simulated Annealing

1. Introduction to Research Methodology

Student attrition poses a significant challenge in higher education, particularly within developing regions where digital learning platforms are being rapidly adopted. In Kenya, attrition rates have reached approximately 35% in traditional classroom settings and around 25% in eLearning and blended

environments. Such statistics underscore a critical institutional and national issue: the premature withdrawal of students disrupts academic progression, reduces institutional efficiency, and undermines broader educational development goals. With the increasing integration of Learning Manage-

^{*}Corresponding author: musyimisamuelmuthama@tum.ac.ke (Musyimi Samuel Muthama)

Received: 10 May 2025; **Accepted:** 29 May 2025; **Published:** 23 July 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

ment Systems (LMS) and digitized administrative data, higher education institutions are now in a position to implement intelligent, data-driven approaches to understanding and predicting student disengagement [1, 2].

Recent advances in machine learning (ML) have shown promise in addressing educational problems, particularly in the domain of student retention. A growing body of literature explores how ML algorithms can analyze large volumes of student data to predict attrition. For instance, [3] employed decision trees and logistic regression to forecast student dropout, while [4] utilized behavioral anomalies as indicators of potential attrition. [5] presented a machine learning pipeline tailored for asynchronous learning environments such as MOOCs, and [1] conducted a systematic review on ML-based student retention models. Despite these contributions, existing methods often exhibit limitations such as reliance on single-model architectures, limited generalizability across diverse educational settings, and insufficient optimization of model parameters.

To address these gaps, this study proposes a hybrid ensemble learning approach for student attrition prediction that combines the strengths of multiple classifiers—Naïve Bayes, Gradient Boosting, and Random Forest. The novelty of the proposed framework lies in its integration of Simulated Annealing, a metaheuristic optimization technique, to refine both model hyperparameters and ensemble weights. This strategy not only enhances predictive accuracy but also improves model adaptability across different institutional contexts. Furthermore, this research adopts a mixed-method research design, incorporating both quantitative data from institutional records and LMS logs, and qualitative insights from surveys and interviews. This dual-pronged approach ensures the model captures both measurable and contextual indicators of attrition [6].

Predictive models of this nature are applicable in several educational contexts. They can serve as the core of early warning systems that alert administrators to students who may require intervention, assist in resource optimization by targeting support services where they are most needed, and contribute to curriculum redesign efforts by highlighting course-level factors associated with dropout. In national policy contexts, such models can inform data-driven strategies for improving retention rates in digital learning environments.

A range of predictive model variants have been developed in the literature, including statistical models such as logistic regression, single-classifier approaches using support vector machines and decision trees, and ensemble methods like Random Forests and XGBoost. More advanced models have employed time-series analysis and explainable AI frameworks. This study contributes to the ensemble-optimization subset of models by leveraging Simulated Annealing to balance the predictive contributions of individual learners. This study is guided by the following central research question: *How do we develop a predictive optimal machine learning model for student attrition identification based on identified*

features? This question forms the foundation of the methodological framework and drives the choice of algorithms, optimization strategies, and feature engineering techniques employed throughout the study.

The key contributions of this research are as follows:

- 1) Development of an ensemble-based predictive model integrating Naïve Bayes, Gradient Boosting, and Random Forest classifiers;
 - 2) Implementation of Simulated Annealing for dynamic hyperparameter tuning and ensemble weight optimization;
 - 3) Use of a mixed-method data collection framework for robust and interpretable model training;
 - 4) Validation of the model on a large-scale real-world dataset from five Kenyan universities comprising over 4,400 student records; and
 - 5) Practical relevance through alignment with institutional needs for early intervention and policy formulation.
- These contributions advance the state-of-the-art in educational data mining and offer actionable pathways for mitigating student attrition in eLearning settings.

The remainder of this paper is organized as follows. Section 2 introduces the Problem formulation of this research, while Section 3 provides a comprehensive research methodology. Subsequently, Section 4 presents the experimental results and analysis in Section 5. Describes data driven mitigations and Section 6 concludes with final remarks and suggestions for future research.

2. Problem Formulation

Formally, let the dataset consist of n student records, each represented by a feature vector $x_i \in \mathbb{R}^d$ where d is the number of input features such as academic scores, login frequency, demographic characteristics, and engagement metrics. The corresponding target variable $y_i \in \{0,1\}$ denotes the student's status, with 1 indicating dropout and 0 indicating retention.

The primary goal is to learn a predictive function $F(\cdot)$ that maps student features x_i to a predicted outcome $y_i = f(x) = (x, \theta)$ minimizing the discrepancy between predicted and actual attrition outcomes.

Where θ represents the parameters of the learning model. The function F is constructed as an ensemble of base classifiers, each contributing partial predictions. Let $[f_1, f_2, \dots, f_M]$ denote M base learners, and let w_m be the weight associated with the m^{th} learner, such that $\sum_{m=1}^M w_m = 1$ with $w_m > 0$. The ensemble prediction.

$$f_{xi} = \sum_{m=1}^M w_m f_m(x) \quad (1)$$

$$y = f(X, \theta) = \sigma(\sum_{j=1}^n w_j x_j + b) \quad (2)$$

$$X = [x_1, x_2, \dots, x_n] \quad \text{matrix } (X) = [x_1, x_2, \dots, x_n]$$

$$X = [x_1, x_2, \dots, x_n]: \text{Feature vector representing each student's}$$

attributes

$Y = \sum (0,1) y$ in $\{0, 1\}$ represents classify target

Table 1. Variable Definitions.

Symbol	Description
x_i	Feature vector representing the attributes of the ii-th student
y_i	Ground truth label: 1 for dropout, 0 for retained student
y^{\wedge}_i	Predicted label for the ii-th student
F	Ensemble classifier combining predictions from multiple base learners
f_m	The mm-th base classifier (e.g., Naïve Bayes, Random Forest, XGBoost)
w_m	Weight of the mm-th base classifier in the ensemble
θ	Model-specific parameters of the classifiers
L	Loss function used for optimization (e.g., cross-entropy)
$\Theta\beta$	Regularization term to penalize complex models
M	Number of base learners in the ensemble
N	Total number of students in the dataset
d	Number of features (input variables) per student record

3. Research Methodology

This study employs a mixed machine learning methodology to construct a predictive model for attrition, aiming to proactively identify potential students' attritions before the behavior occurs. To address student attrition in Kenya, considering the identified attrition rates of 35% for traditional learning and 25% for eLearning and blended learning, a comprehensive methodology incorporating the following steps is recommended [3].

If predicting into more than two classes (e.g., graduate risk, attrition, recanted low risk).

$$P\left(y = \frac{k}{X}\right) = \frac{e^{wkTX+bk}}{\sum_{j=1}^k e^{wkTX+bj}} \quad (3)$$

To formally express a classification model equation that categorizes students into attrition classes (e.g., dropout vs. retained). This comprehensive approach, grounded in attrition and attrition identification, underscores the importance of leveraging machine learning to construct predictive models that facilitate early identification and targeted strategies for potential attritions in educational settings [4]

$$L^{(t)} = \sum_{i=1}^n l(y_i, x_i^{t-1}) + ft(X_i) + \beta(ft) = 1 \quad (4)$$

To predict the i-th instance at the t-th iteration, it is necessary to incorporate the function $ft(X_i) + \beta(ft)$ in order to minimize the corresponding objective function.

3.1. Methodology Design

Using a mixed method research design for optimal predictive machine learning in student attrition identification offers a robust approach to understanding and addressing the multifaceted issue of student dropout. By leveraging the strengths of both quantitative and qualitative methods, researchers can develop more accurate, interpretable, and actionable models, ultimately leading to more effective interventions and improved student retention rates [5].

A literature review design outlines the systematic approach taken to gather, evaluate, and synthesize existing research related to machine learning model for optimal prediction of student's attrition in eLearning environment [2].

3.2. Descriptive Analysis of Student Attrition

Descriptive analysis provides a foundational approach to examining the patterns, trends, and characteristics of student attrition. By analyzing demographic data, academic performance, engagement levels, and other relevant factors, institutions can identify at-risk students and implement targeted interventions. This analysis helps in formulating policies and practices aimed at reducing dropout rates and improving overall educational outcomes. This paper employing a combination of these methods, educational institutions can effectively identify at-risk students and implement strategies to improve retention and reduce attrition rates [4].

3.3. Quantitative Data Collection Methods

Quantitative methodology leverages numerical data to build and refine machine learning models for predicting student attrition. Various methods are deployed to ensure comprehensive data collection and analysis. Collect large-scale numerical data from student records, including demographics, academic performance, attendance, and interaction logs from eLearning platforms. Data will be collected from the Kenya Ministry of Education, Kenyan universities offering online courses, and publicly available educational datasets. Gather student demographic data, academic performance metrics, engagement levels, and other relevant variables from both traditional and eLearning/blended learning programs. Analyze historical data to identify patterns and factors contributing to attrition in each mode of learning. In-depth Data collection, Google questionnaires, interviews and Systems logs observation [7].

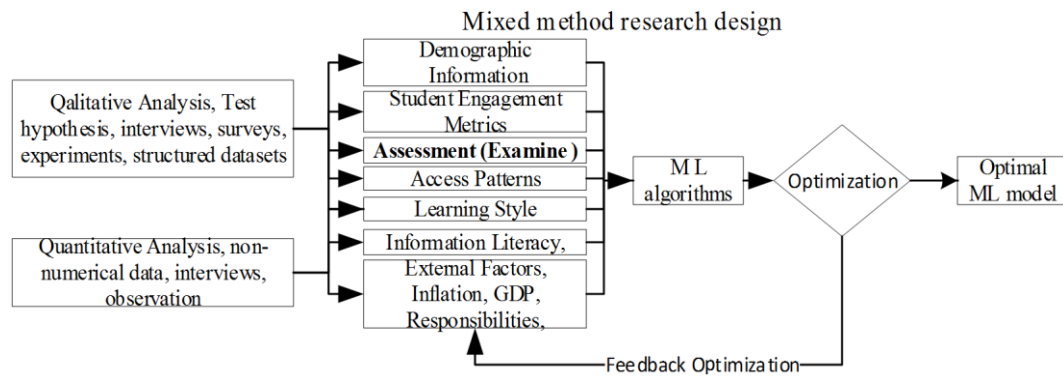


Figure 1. Methodology Design.

3.4. Qualitative Data Collection Methods

Qualitative methodology plays a crucial role in enriching machine learning models for student attrition identification by providing deeper insights and contextual understanding.

A comprehensive approach to understanding student attrition involves multiple data collection methods. In-depth interviews and focus groups with students, faculty, and administrators uncover personal, social, and institutional challenges contributing to drop-

out rates. Cohort assessments help identify group-specific patterns, enabling the customization of predictive models for different student demographics. Open-ended surveys provide qualitative insights into students' experiences, revealing trends that inform predictive features. Additionally, analyzing system logs from Learning Management Systems (LMS) helps track engagement levels and interaction behaviors, highlighting early warning signs of attrition. Observing section interactions and eLearning content delivery further identifies effective teaching strategies, supporting targeted interventions to improve student retention.

4. Machine Learning Methodology (Machine Learning Methodology Step)

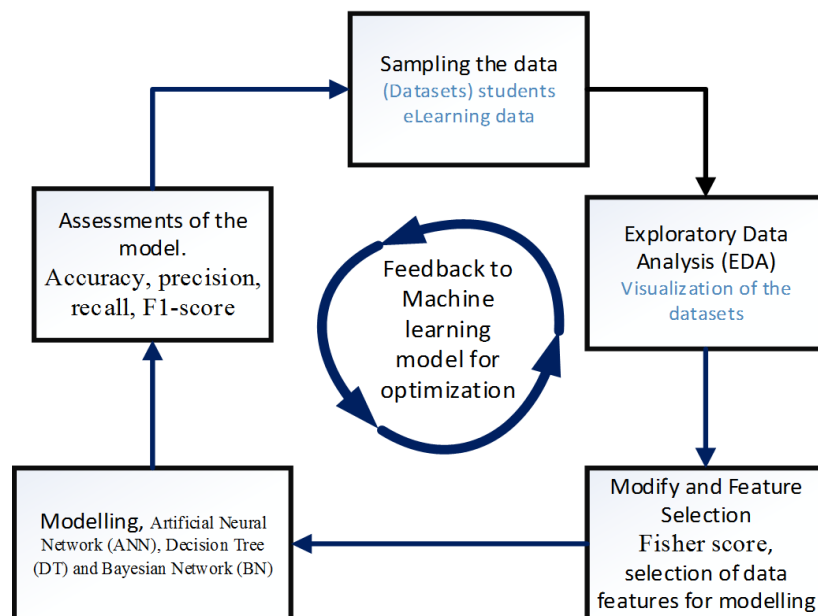


Figure 2. The Semma process of machine learning modelling [1].

SEMMA (Sample, Explore, Modify, Model, Assess) SEMMA widely used in data mining and predictive analytics projects across various industries. Its structured and iterative approach ensures that the data thoroughly understood and prepared, leading to the development of robust predictive

models. The methodology particularly effective when dealing with large datasets that require efficient processing and analysis.

4.1. Step 1 Data Collection Techniques, Sampling the Data

Gather student demographic data, academic performance metrics, engagement levels, and other relevant variables from both traditional and eLearning/blended learning programs. Analyze historical data to identify patterns and factors contributing to attrition in each mode of learning.

- i. The Ministry of Education website (<https://www.education.go.ke/>) data base.
- ii. UNESCO Institute for Statistics (<https://uis.unesco.org/>) provides global education data.

Data will include student demographics, academic performance, attendance records, engagement metrics, and interaction with eLearning platforms.

4.2. Step 2, Exploratory Data Analysis (EDA)

The data then explored using statistical and graphical techniques to understand its structure, identify outliers, and uncover underlying relationships among variables. In the realm of attrition identification, a meticulous data pre-processing stage was undertaken with the two datasets. This essential step ensured the cleanliness and accuracy of the

data before model development. To uphold privacy, sensitive information revealing individual identities was expunged. The process involved handling missing values by replacing them with medians and zeroes, guaranteeing a refined and privacy-conscious training set for the subsequent stages of the attrition identification model development. Handle missing values, normalize data, encode categorical variables, and split into training and testing sets [8].

4.3. Step 3, Data Description

To develop and validate a robust machine learning model for predicting student attrition, crucial to use a substantial and representative dataset. The size and characteristics of the dataset used for analysis and simulation play a significant role in ensuring the model's accuracy and generalizability. Quantitative data Academic performance, attendance records, and eLearning platform interactions from major universities and colleges in Kenya. Available datasets from research publications and open data.

- i. Total number of student records: 4424.
- ii. Number of institutions 5 major public universities and colleges offering eLearning programs.

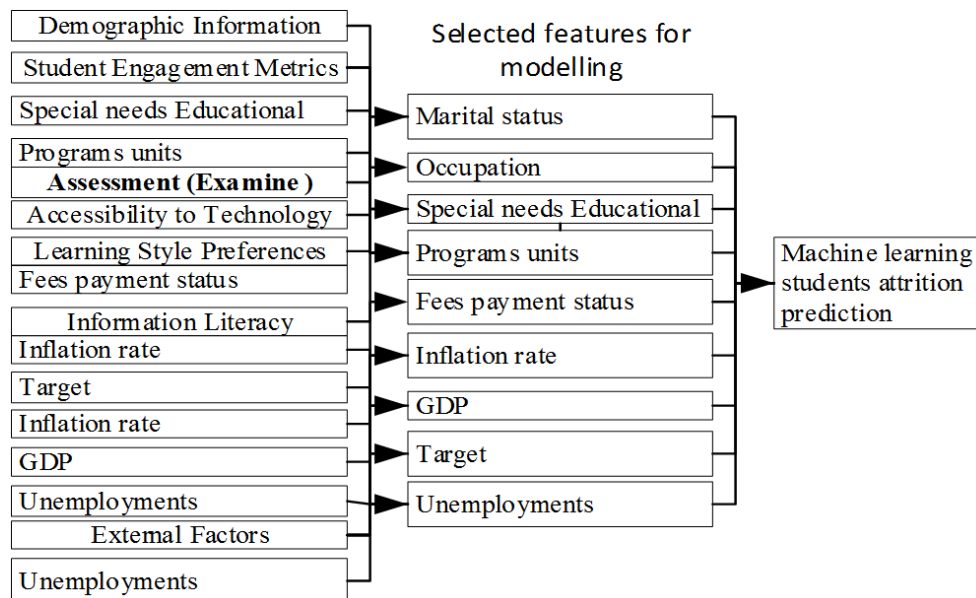


Figure 3. Features selection to students Attrition perdition [8].

$$F_j = ((\mu_j i)) / 1! + \frac{n(n-1)x^2}{2!} + \dots \quad (5)$$

4.4. Step 4, Modify and Feature Selection

Data preparation is essential for developing predictive models for student attrition. This step includes cleaning data to address inconsistencies, transforming variables for com-

patibility, and creating new features to enhance predictive power. Selecting relevant features ensures the model focuses on key factors that influence attrition [9].

The Fisher score, a statistical feature selection technique, will identify the most significant variables for the predictive model. Features such as student engagement metrics, assessment scores, and interaction patterns will be prioritized based on their discriminative power. This ensures the model

captures critical patterns indicative of attrition, improving accuracy and effectiveness in eLearning environments. Split the Datasets into two training datasets and testing datasets construct the training data set and feed training datasets the data into the attrition prediction models. Data Spitted selected features [4]. Training Set: 80% of the dataset (3,540 Student records).

- i. Training Set: 80% of the dataset (3,540 Student records).
- ii. Validation and testing Set: 20% of the dataset (885 Student records).

4.5. Computational Experiment

The computational experience for this study involved the implementation of advanced machine learning algorithms using Python and relevant data science libraries such as

Scikit-learn, XGBoost, and TensorFlow. Data preprocessing, feature engineering, model training, and evaluation were conducted in Jupyter Notebooks. Hyperparameter tuning was carried out using Grid Search and Simulated Annealing techniques. The experiments were executed on a high-performance computing environment to ensure efficient processing of large datasets and support real-time predictive modeling for student attrition.

In a more recent study, [10] discussed the application of predictive analytics in identifying at-risk students. They demonstrated how machine learning models could predict student attrition based on various factors, allowing for early and effective interventions. The paper underscored the potential of data-driven approaches in enhancing student retention in eLearning environments.

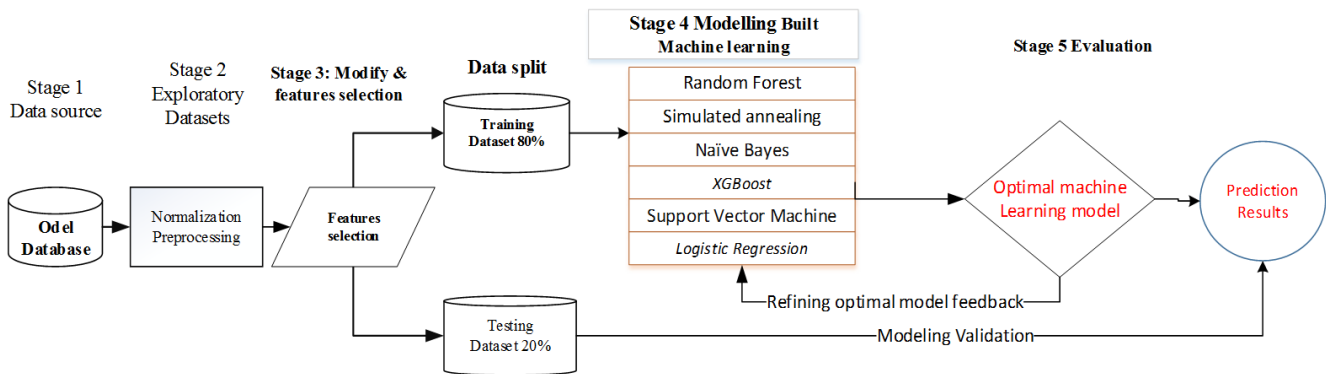


Figure 4. Steps Conceptual Model for Predicting Student Attrition in ELearning Environments [11].

4.6. Step 5, Modelling Machine Learning Student's Attrition Prediction

Different modeling techniques are applied to the prepared data to build predictive models. This step involves training machine learning algorithms on the dataset. Training the prediction models that were constructed based on machine learning methods such as the Artificial Neural Network (ANN), Decision Tree (DT) and Bayesian Network (BN) to derive the samples of the prediction model. Machine learning models are crucial for predicting student attrition by identifying at-risk students based on various data points. Among the optimal algorithms for this task are. Using these algorithms, educational institutions can develop robust predictive models, enabling timely interventions to reduce attrition rates and

improve student retention [12].

$$f_t^*(x) = \arg \min f \sum_{j=1}^n g(f(x_j)) + \frac{1}{2} h(f^2(x_j)) + \Omega(f) \quad (6)$$

To calculate the corresponding optimal equation for predicting the instance at the iteration in a machine learning model like Gradient Boosting, we generally aim to minimize a loss function by adding a new function $f_t(X_i) + \beta(f_t)$ to the ensemble model [12, 5].

$$y = f(X, \emptyset) = \sigma(\sum_{j=1}^n w_j x_j + b) \quad (7)$$

$X = [x_1, x_2, \dots, x_n]$ $\mathbf{X} = [x_1, x_2, \dots, x_n]$ $X = [x_1, x_2, \dots, x_n]$: Feature vector representing each student's attributes.

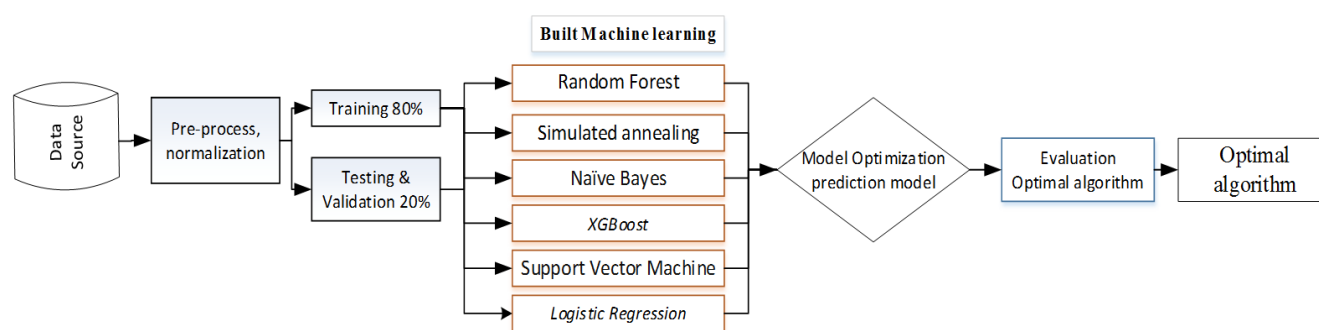


Figure 5. Machine learning multiple algorithm [4, 13, 14].

4.7. Step 6, Assessments of the Model

The final step evaluates the performance of the models using appropriate metrics and validation techniques to ensure they meet the desired objectives. Use Testing datasets for testing, feeding it into the actual samples of prediction model previously generated. Assessing the predictive machine learning model for student attrition identification involves several key steps to ensure its accuracy, reliability, and effectiveness. Performance metrics such as accuracy, precision, recall, and F1-score are evaluated to measure the model's ability to correctly identify at-risk students. K-fold cross-validation implemented to ensure robustness and mitigate overfitting. A confusion matrix is used to visualize true positives, true negatives, false positives, and false negatives, highlighting areas of misclassification [1].

5. Data-Driven Interventions

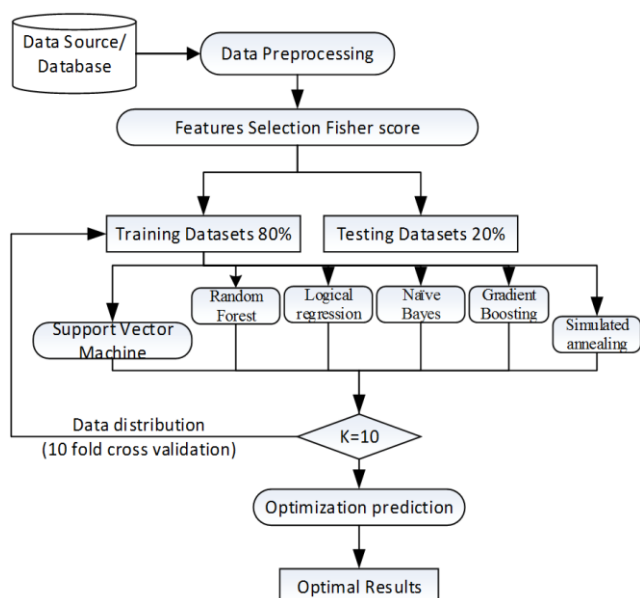


Figure 6. Building of the machine learning model [15, 16].

Quantitative models play a critical role in identifying students at high risk of attrition by analyzing patterns in academic performance, engagement metrics, and demographic variables. However, to effectively design targeted interventions, it is essential to complement these insights with qualitative data. For example, while predictive models may indicate that students with low platform engagement are more likely to drop out, qualitative methods such as interviews and focus groups can uncover the underlying reasons such as lack of interactive content in eLearning platform, feelings of isolation, or unclear course expectations and learning outcome. This combined approach not only validates the predictive outcomes but also provides a deeper understanding of the student experience. By integrating both quantitative and qualitative methods, educational institutions can develop more comprehensive support systems that address both the symptoms and root causes of attrition. Ultimately, this synergy enables timely, personalized interventions, enhancing student retention and the overall effectiveness of eLearning programs.

6. Conclusion

This study presented a hybrid ensemble machine learning framework for the prediction of student attrition in eLearning environments, with a focus on Kenyan public universities. By combining Naïve Bayes, Random Forest, and Gradient Boosting algorithms, and optimizing their performance through Simulated Annealing, the proposed model demonstrated improved accuracy and robustness compared to single-model approaches. The mixed-method research design—incorporating both quantitative and qualitative data—enabled the identification of key behavioral, academic, and contextual features that contribute to student dropout, thereby enhancing model interpretability and practical relevance.

The findings indicate that early prediction of attrition is feasible using institutional data, LMS logs, and structured surveys. The model's predictive capabilities can inform timely interventions and support systems that improve student retention and academic performance in digital education contexts. This contribution aligns with the broader goal

of developing data-driven decision-support tools for educational administrators and policymakers. Despite the promising results, several challenges emerged. First, data imbalance between dropout and retained student cases affected classification sensitivity. Second, variations in feature availability across institutions introduced limitations in the generalizability of the model. Third, while Simulated Annealing offered effective parameter optimization, it also introduced computational overhead that may not be feasible in all institutional settings. Lastly, limited availability of qualitative feedback from students constrained the depth of contextual interpretation in certain cases [6].

Future research may focus on several directions. Integrating time-series models such as LSTM networks could enable dynamic tracking of engagement and learning trajectories over time. Further investigation into explainable AI methods, such as SHAP or LIME, would enhance model transparency and institutional trust. Additionally, expanding the dataset to include more diverse institutions across different regions and incorporating real-time data streams from LMS platforms could further strengthen model accuracy and scalability. Lastly, co-designing predictive tools with educational stakeholders—such as faculty, counselors, and students—could improve intervention design and foster sustainable institutional adoption. In conclusion, the proposed ensemble learning methodology offers a scalable, data-informed solution to one of higher education's most pressing issues: student attrition. With continued refinement and contextual adaptation, such models can play a central role in shaping the future of personalized, inclusive, and resilient eLearning ecosystems.

Abbreviations

HEI	Higher Education Institution
LMS	Learning Management System
MOOC	Massive Open Online Course
ICT	Information and Communication Technology
AI	Artificial Intelligence
ML	Machine Learning
eLearning	Electronic Learning
GPA	Grade Point Average
SEMMA	Sample, Explore, Modify, Model, Assess
EDA	Exploratory Data Analysis
SVM	Support Vector Machine
RF	Random Forest
XGBoost	Extreme Gradient Boosting
DT	Decision Tree
BN	Bayesian Network
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
F1	F1 Score
CV	Cross-Validation

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Firas, O., A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach. *World Journal of Advanced Engineering Technology and Sciences*, 2023. 8(1): p. 009-014.
- [2] Kocsis, Á. and G. Molnár, Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, 2024: p. 1-19.
- [3] Ware, G. J., Instructors' Strategies to Improve Attrition Rates Among Online Culturally Diverse Students: A Qualitative Exploratory Case Study. 2024, National University.
- [4] Zhou, Y., J. Zhao, and J. Zhang, Prediction of learners' dropout in E-learning based on the unusual behaviors. *Interactive Learning Environments*, 2023. 31(3): p. 1796-1820.
- [5] Banaag, R., J. L. Sumodevilla, and J. Potane, Factors affecting student drop out behavior: A systematic review. *International journal of educational management and innovation*, 2024. 5(1): p. 53-70.
- [6] Yuan, Y., et al., Adaptive simulated annealing with greedy search for the circle bin packing problem. *Computers & Operations Research*, 2022. 144: p. 105826.
- [7] Delen, D., B. Davazdahemami, and E. Rasouli Dezfouli, Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework. *Information Systems Frontiers*, 2023: p. 1-22.
- [8] Srivastava, P. R. and P. Eachempati, Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach. *Journal of Global Information Management (JGIM)*, 2021. 29(6): p. 1-29.
- [9] Shafiq, D. A., et al., Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access*, 2022. 10: p. 72480-72503.
- [10] Chhetri, B., L. M. Goyal, and M. Mittal, How machine learning is used to study addiction in digital healthcare: A systematic review. *International Journal of Information Management Data Insights*, 2023. 3(2): p. 100175.
- [11] Fauszt, T., et al., Design of a Machine Learning Model to Predict Student Attrition. *International Journal of Emerging Technologies in Learning*, 2023. 18(17).
- [12] Paul, R. and M. Rashmi, Student Satisfaction and Churn Predicting using Machine Learning Algorithms for EdTech course. in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). 2022.

- [13] Benoit, D. F., et al., High-stake student drop-out prediction using hidden Markov models in fully asynchronous subscription-based MOOCs. *Technological Forecasting and Social Change*, 2024. 198: p. 123009.
- [14] Fullgence, M. and F. M. Mwakondo, Validated Conceptual Model for Predictive Mapping of Graduates' Skills to Industry Roles Using Machine Learning Techniques. 2022.
- [15] Hamim, T., F. Benabbou, and N. Sael, Student profile modeling using boosting algorithms. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 2022. 17(5): p. 1-13.
- [16] Tan, M. and P. Shao, Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. *International Journal of Emerging Technologies in Learning (iJET)*, 2015. 10.