Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Optimized Stacking Ensemble Model for Forecasting Student Dropout

Jackson Kisuule

jkisuule89@gmail.com

Victoria University

Anthony Bua

Victoria University

Research Article

Keywords: Stacking ensemble, K-Nearest Neighbors, Bayesian Network

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10827036/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Optimized Stacking Ensemble Model for Forecasting Student Dropout

Kisuule Jackson[1,2,*],Bua Anthony[1]
[1]Faculty of science and technology, Victoria University, Kampala
[2]The Unity University, Hargeisa, Somaliland

[*]*Corresponding author:* jkisuule89@gmail.com

Abstract

Student dropout remains a major challenge that affects academic success, institutional performance, and resource utilization. Existing methods of identifying at-risk students are often manual, slow, and less effective for timely intervention. This study aimed at developing an optimized stacking ensemble model for forecasting student dropout. The study used an optimized stacking ensemble model whose architecture integrated K-Nearest Neighbors and Bayesian Network classifiers, enabling the model to learn complementary decision patterns while reducing prediction variance and improving generalization. Hyperparameter optimization, SMOTE-based class balancing, feature engineering, and five-fold cross-validation were incorporated to improve model robustness. The model was evaluated using accuracy, precision, recall, F1-score, and specificity. Results showed that the stacking ensemble outperformed the existing individual classifiers and the soft voting ensemble, achieving 85.9% accuracy, precision of 76.4%, recall of 81.3%, F1-score of 78.8%, and specificity of 81.3% providing an effective early-warning tool to support timely interventions and improve student retention.

Keywords: Stacking ensemble; K-Nearest Neighbors; Bayesian Network.

1 Introduction

Decreasing student attrition rates is one of the main objectives of most higher education institutions [1]. According to recent analysis, dropout rates in developing regions remain persistently high, with projections indicating that up to 40% of university entrants may fail to graduate without targeted interventions [2]. However, to achieve this goal, universities need to accurately identify and focus their efforts on students most likely to quit their studies before they graduate .This has given rise to a need to implement a stacking ensemble model for forecasting which students will eventually drop out [1].

Educational institutions are increasingly adopting data-driven technologies to improve academic performance, enhance student support services, and promote institutional sustainability. Among these technologies, Educational Data Mining (EDM) and Machine Learning (ML) have emerged as effective approaches for analyzing educational data and generating predictive insights that support informed decision-making. These techniques enable institutions to identify hidden patterns in student records and develop predictive models that facilitate timely academic interventions and improve student retention [3, 4]. Student dropout has become one of the most significant challenges facing higher education institutions worldwide. High dropout rates negatively affect institutional performance, reduce graduation rates, increase financial losses, and limit students' opportunities for academic and professional success. Although universities continue to implement academic advising, financial support, and student counseling programs, many students still withdraw before completing their studies. This demonstrates the need for intelligent predictive systems capable of identifying students who are at risk of dropping out before withdrawal occurs [4,5]. The increasing availability of educational data has shifted research from traditional statistical approaches toward intelligent machine learning and deep learning models capable of automatically identifying students at risk of dropping out [6].

[7] and [4] articulate that the causes of student dropout are multidimensional and involve academic, financial, personal, institutional, and social factors. Poor academic performance, financial constraints, inadequate institutional support, dissatisfaction with academic programs, attendance patterns, and personal circumstances have consistently been identified as major contributors to student attrition. Since these factors interact differently across insti-

tutions, developing institution-specific predictive models is essential for improving prediction accuracy and supporting effective intervention strategies. Existing approaches for identifying at-risk students largely depend on manual monitoring and periodic academic reviews, making it difficult to detect vulnerable students early enough for meaningful intervention. Consequently, there is a need for an intelligent prediction framework that utilizes historical student data to forecast dropout risk and support proactive decision-making by university administrators.

Early studies on student dropout prediction primarily relied on educational theories and statistical models. [8] introduced the Student Integration Theory, which argues that students' academic and social integration significantly influences persistence in higher education. [9] later proposed the Student Attrition Model, emphasizing behavioral intentions, institutional commitment, financial circumstances, and environmental factors as determinants of student persistence.

Although these theories have contributed significantly to understanding student attrition, they mainly provide conceptual explanations rather than predictive mechanisms capable of automatically identifying at-risk students. Consequently, researchers increasingly adopted statistical techniques such as Logistic Regression to estimate dropout probabilities from institutional datasets [10].

Machine learning algorithms have become widely adopted because they can learn nonlinear relationships among demographic, behavioral, academic, and socio-economic variables without relying on restrictive statistical assumptions.

[11] applied Logistic Regression to predict student dropout using institutional academic records. Although the method provided interpretable results, its linear decision boundary limited predictive performance when modeling complex educational datasets.

KNN has also been employed for educational prediction due to its simplicity and competitive predictive performance. [12] reported satisfactory results using KNN, although computational complexity increased considerably as dataset size expanded.

Bayesian Networks have received considerable attention because they model uncertainty among educational variables. [13] demonstrated that Bayesian

Networks estimate dropout probabilities effectively under incomplete information. However, constructing optimal network structures becomes increasingly challenging in complexity for high-dimensional educational datasets.

Recent advances in Artificial Intelligence have encouraged researchers to investigate deep learning techniques for educational prediction. Artificial Neural Networks (ANNs) automatically learn hierarchical representations from educational datasets without requiring extensive feature engineering.

[14] developed ANN models capable of accurately identifying students likely to withdraw from university. Similarly, [15] demonstrated that Long Short-Term Memory (LSTM) networks effectively capture temporal relationships in students' academic progression.

Ensemble learning combines predictions generated by multiple models to improve classification accuracy and robustness. Bagging approaches such as Random Forest reduce prediction variance by aggregating multiple Decision Trees, whereas boosting algorithms including AdaBoost, Gradient Boosting, LightGBM, and XGBoost sequentially improve weak learners.

More recently, stacking ensemble learning has emerged as one of the most effective approaches for educational prediction. [16] demonstrated that stacking heterogeneous machine learning models significantly improves student dropout prediction by leveraging the complementary strengths of multiple classifiers. Similarly, [17] reported that stacking ensembles consistently outperform individual machine learning models in terms of prediction accuracy and robustness. [18] studied a model for predicting dropout of higher education students that aimed at exploring models to facilitate management in formulating assertive policies to promote student retention at higher education institutions. The study proposed an ensemble model which included three data mining techniques of which two considered white box (decision tree and logistic regression), and one black box (ANN) to predict the dropout of undergraduate students. The challenge of the ensemble model was that all selected models had similar performance, homogeneously classifying the same students. They suggested that in order to improve the power of their approach, new models(techniques/algorithms) such as Bayesian classifier, KNN and support vector machines could be incorporated into the ensemble model for better accuracy, performance and robustness of the same specific problem. This study was relevant to the present research as it aimed to incorporate the suggested models to underscore the importance of accuracy,

performance and robustness which were core evaluation metrics in predicting student dropout.

According to [16], the authors acknowledged that their study was limited by the small dataset size (only 110 students), which prevented the use of more sophisticated deep learning models such as neural networks. They suggested that future work should collect data from more students enrolled in subsequent course cycles and consider additional features designed with machine learning requirements in mind. This study was relevant to the present research as it demonstrated that ensemble methods outperform individual classifiers in dropout prediction and confirmed the importance of feature selection and data quality, particularly in small-sample contexts common in developing regions. [12], the authors noted that models trained on individual algorithms showed performance variability, and that combining these algorithms in an ensemble framework could improve prediction stability and generalization. This study was relevant to the present research because it provided empirical evidence for selecting SVM, KNN, and Bayesian networks as complementary base learners for the ensemble model, particularly given their proven effectiveness in resource-constrained contexts similar to Somaliland. [18], the authors noted that the complexity of hyperparameter optimization using advanced genetic algorithms requires substantial computational resources and technical expertise, which may limit applicability in institutions with constrained digital infrastructure. This study was relevant to the present research because it provided methodological guidance on data preprocessing and feature selection strategies essential for developing an optimized ensemble model, and highlights the importance of balancing predictive accuracy with computational feasibility for deployment.

2 Methods

The goal of this study is to develop and evaluate an optimized stacking ensemble model for forecasting student dropout and use the dataset collected from The Unity University, Hargeisa, Somaliland to evaluate the performance. The proposed methodology addresses the limitations of existing student dropout prediction approaches by integrating KNN and Bayesian Network classifiers within a stacking ensemble model to improve prediction accuracy, generalization, and robustness. To achieve this objective, a systematic machine learning pipeline was implemented, comprising data collection,

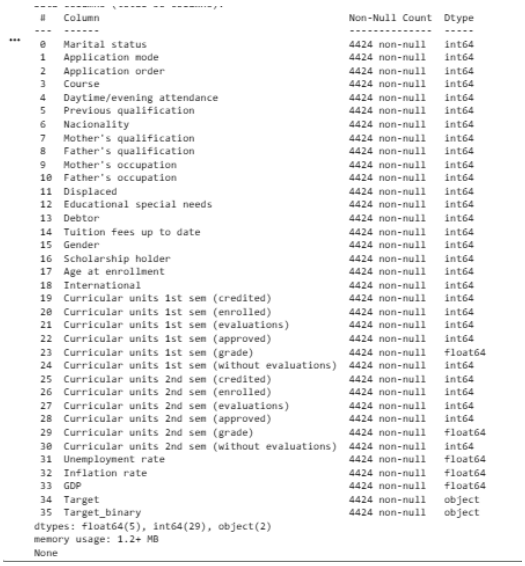
data preprocessing, feature engineering, class balancing using the Synthetic Minority Oversampling Technique (SMOTE) [19], hyperparameter optimization, model development, and performance evaluation using historical student records. The developed model was compared with the individual base learners and a soft voting ensemble using standard classification performance metrics, including accuracy, precision, recall, F1-score, and specificity.

2.1 Dataset

The dataset used in this study consists of historical student academic records collected from The Unity University, Hargeisa, Somaliland, covering the academic period from 2021 to 2025. The dataset was obtained from the university's student information management system and contains demographic, academic, and institutional attributes that influence student retention and dropout. Before model development, the dataset underwent several preprocessing operations to improve data quality and ensure reliable prediction performance. Categorical variables were transformed into numerical representations through label encoding, and continuous features were normalized to ensure comparable scales across all variables. Feature engineering was further performed to derive meaningful attributes that improve the predictive capability of the proposed model.

Student dropout datasets are naturally imbalanced because the number of continuing students is considerably larger than the number of students who discontinue their studies. To address this challenge, SMOTE was applied during model training to balance the class distribution and minimize prediction bias towards the majority class. This preprocessing step enabled the stacking ensemble model to learn representative patterns from both dropout and non-dropout classes.

Following preprocessing, the dataset was partitioned into training and testing subsets to facilitate model development and unbiased performance evaluation. The training data were used to build and optimize the base learners and the Gaussian Naive Bayes meta-learner, while the testing data were reserved for evaluating the generalization ability of the proposed stacking ensemble model using standard classification performance metrics.



```
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Marital status                                  4424 non-null   int64
 1   Application mode                                4424 non-null   int64
 2   Application order                               4424 non-null   int64
 3   Course                                          4424 non-null   int64
 4   Daytime/evening attendance                      4424 non-null   int64
 5   Previous qualification                          4424 non-null   int64
 6   Nacionality                                     4424 non-null   int64
 7   Mother's qualification                          4424 non-null   int64
 8   Father's qualification                          4424 non-null   int64
 9   Mother's occupation                             4424 non-null   int64
 10  Father's occupation                             4424 non-null   int64
 11  Displaced                                       4424 non-null   int64
 12  Educational special needs                       4424 non-null   int64
 13  Debtor                                          4424 non-null   int64
 14  Tuition fees up to date                         4424 non-null   int64
 15  Gender                                          4424 non-null   int64
 16  Scholarship holder                              4424 non-null   int64
 17  Age at enrollment                               4424 non-null   int64
 18  International                                   4424 non-null   int64
 19  Curricular units 1st sem (credited)             4424 non-null   int64
 20  Curricular units 1st sem (enrolled)             4424 non-null   int64
 21  Curricular units 1st sem (evaluations)          4424 non-null   int64
 22  Curricular units 1st sem (approved)             4424 non-null   int64
 23  Curricular units 1st sem (grade)                4424 non-null   float64
 24  Curricular units 1st sem (without evaluations)  4424 non-null   int64
 25  Curricular units 2nd sem (credited)             4424 non-null   int64
 26  Curricular units 2nd sem (enrolled)             4424 non-null   int64
 27  Curricular units 2nd sem (evaluations)          4424 non-null   int64
 28  Curricular units 2nd sem (approved)             4424 non-null   int64
 29  Curricular units 2nd sem (grade)                4424 non-null   float64
 30  Curricular units 2nd sem (without evaluations)  4424 non-null   int64
 31  Unemployment rate                               4424 non-null   float64
 32  Inflation rate                                  4424 non-null   float64
 33  GDP                                             4424 non-null   float64
 34  Target                                          4424 non-null   object
 35  Target_binary                                   4424 non-null   object
dtypes: float64(5), int64(29), object(2)
memory usage: 1.2+ MB
None
```

Fig. 1: Dataset Description

2.2 Model Development

Individual learning algorithms capture only specific patterns within educational data, resulting in inconsistent predictive performance across different student populations. To overcome these limitations, this study developed an optimized stacking ensemble model that combines the complementary strengths of KNN and Bayesian Network classifiers within a unified prediction framework. The KNN classifier effectively captures local similarity patterns among students, while the Bayesian Network models probabilistic relationships between demographic, academic, and socio-economic factors associated with student dropout. Their combined predictions provide richer information to the meta-learner, enabling more accurate and robust classification than either individual model.

The development process of the proposed model consisted of four major stages, namely data preprocessing, base learner training, ensemble integration, and model evaluation, as illustrated in Figure 2.

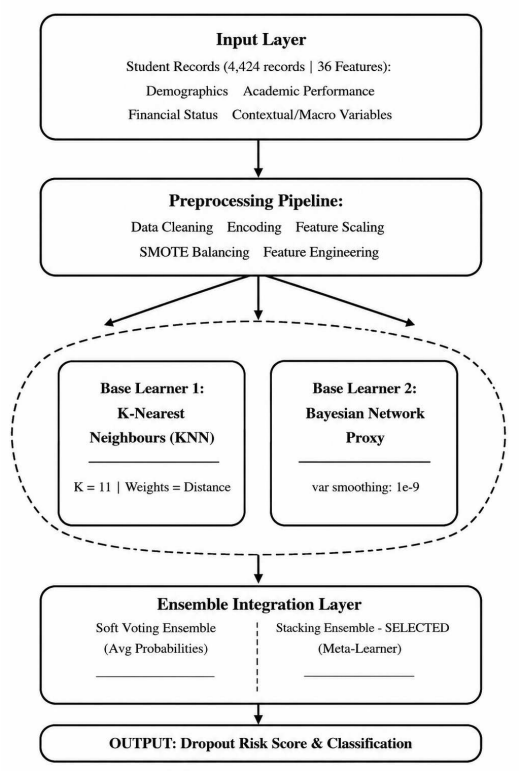


Fig. 2: Model Architecture

2.3 Mathematical Model Formulation

Each student record was represented as a feature vector x of n = 36 dimensions. After preprocessing (encoding, scaling, feature engineering), each student i was;

$$x^{(i)} = \left[x_1, x_2, x_3, \ldots, x_n\right]^T \in \mathbb{R}^n \tag{2.1}$$

where $x^{((i))}$ $(i = 1..., 36)$ is a standardized numeric feature, T is the transpose operator, R is real numbers. The target label y is in {1=Dropout, 0=Not_Dropout}.

Ensemble learning extends supervised learning by combining multiple base learners to obtain a model whose predictive performance exceeds that of the individual classifiers. Instead of relying on a single algorithm, heterogeneous learners capture complementary patterns within the data, while a meta-learner integrates their predictions to produce the final classification.

The proposed model therefore employs KNN and Bayesian Network classifiers as base learners whose prediction probabilities are combined through a Gaussian Naive Bayes meta-learner within a stacking ensemble architecture [20, 21].

KNN classified a student based on similarity to the K=11 nearest training records, using Euclidean distance as the similarity measure

$$\hat{P}_{\text{KNN}}(\text{Dropout} \mid x) = \frac{\sum_{x_j \in N_K(x)} w_j \, \mathbf{1}[y_j = \text{Dropout}]}{\sum_{x_j \in N_K(x)} w_j} \tag{2.2}$$

Where, $N_K(x)$ was the set of the $K = 11$ nearest neighbors of x in the training data, $\mathbf{1}[y_j = \text{Dropout}]$ was an indicator that equals 1 if neighbor j was a dropout, and 0 otherwise and closer neighbors (smaller $d(x, x_j)$) received larger weights w_j, so they influenced the probability estimate more strongly.

The Bayesian Network classifier (implemented as Gaussian Naive Bayes) used Bayes' Theorem to compute the posterior probability of each class given a student's feature vector $X = (x_1, x_2, \ldots, x_n)$

$$P(y \mid X) = \frac{P(X \mid y)\, P(y)}{P(X)} \tag{2.3}$$

Where; $P(y \mid X)$ was the posterior probability of class y given features X, $P(X \mid y)$ was the likelihood of observing features X under class y, $P(y)$ was the prior probability of class y and $P(X)$ was the marginal likelihood (normalizing constant).

For the Dropout class, the BN (Gaussian Naive Bayes) dropout probability for a new student x was

$$\hat{P}_{\text{BN}}(\text{Dropout} \mid x) = \frac{P(\text{Dropout}) \prod_{k=1}^{n} P(x_k \mid \text{Dropout})}{P(X)} \tag{2.4}$$

After training, each base learner produced a calibrated dropout probability for every student, using cross validated out of fold predictions so that the

meta learner never sees "cheated" predictions. For student i, the output vector was,

$$z^{(i)} = \left[\hat{P}_{\text{KNN}}(D \mid x^{(i)}), \hat{P}_{\text{BN}}(D \mid x^{(i)})\right]^{\text{T}} \in [0,1]^2 \tag{2.5}$$

Where $\hat{P}_{\text{KNN}}(D \mid x^{(i)})$ was the calibrated dropout probability from the KNN, $\hat{P}_{\text{BN}}(D \mid x^{(i)})$ was the calibrated dropout probability from the Bayesian Network. D denoted the Dropout class and the superscript (i) indexed students, $[0,1]^2$ indicated that all two components were probabilities between 0 and 1.The 2 dimensional vector $z^{(i)}$ replaced the original feature vector $x^{(i)}$ as the input to the Level 1 meta learner. The meta dataset was,

$$\mathcal{Z} = \left\{\left(z^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N} \tag{2.6}$$

constructed using 5 fold cross validation on the training set so that each $z^{(i)}$ is based on out of fold predictions from the base models. The stacked ensemble converted the vector z into a dropout probability using;

$$\hat{P}_{\text{Stack}}(\text{Dropout} \mid x) = \sigma(\eta(z)) = \frac{1}{1+\exp(-\eta(z))} \tag{2.7}$$

with the linear predictor;

$$\eta(z) = \beta_0 + \beta_1 \hat{P}_{\text{KNN}}(\text{Dropout} \mid x) + \beta_2 \hat{P}_{\text{BN}}(\text{Dropout} \mid x) \tag{2.8}$$

Where, $\hat{P}_{\text{KNN}}(\text{Dropout} \mid x)$ was the KNN dropout probability, $\hat{P}_{\text{BN}}(\text{Dropout} \mid x)$ was the Bayesian Network dropout probability, and β_0,β_1 ,β_2 were meta learner coefficients learned from data giving a final stack ensemble model as;

$$\hat{P}_{\text{Stack}}(\text{Dropout} \mid x) = \frac{1}{1+\exp\left(-\left[\beta_0 + \beta_1 \hat{P}_{\text{KNN}} + \beta_2 \hat{P}_{\text{BN}}\right]\right)} \tag{2.9}$$

Using this probability, the final class label was obtained via a threshold rule:

$$\hat{y} = \begin{cases} 1(\text{Dropout}), & \text{if } \hat{P}_{\text{Stack}}(\text{Dropout} \mid x) \geq 0.5, \\ 0(\text{Not_Dropout}), & \text{otherwise.} \end{cases}$$

2.4 Simulation

The simulation of the proposed optimized stacking ensemble model was carried out using Python and machine-learning libraries including Scikit-learn, Pandas, NumPy, and Imbalanced-learn. The dataset was preprocessed through cleaning, encoding, feature scaling, feature engineering, and SMOTE balancing before model training. The processed dataset was divided into 70% training data and 30% testing data. K-Nearest Neighbors and Gaussian Naive Bayes were trained as the base learners, while Gaussian Naive Bayes was used as the stacking meta-learner.

Since KNN and Gaussian Naive Bayes are non-iterative machine-learning algorithms, model training did not involve epochs as used in neural networks. Instead, model optimization was performed using hyperparameter tuning and five-fold cross-validation. For KNN, different values of K were evaluated, and the best-performing configuration was $K = 11$ with distance weighting. The Bayesian Network proxy was optimized using the variance-smoothing parameter. The final stacking ensemble combined the prediction probabilities from the two base learners and generated the final dropout classification.

The simulation was executed on a computer with at least intel(R) Pentium(R) CPU 4415U @ 2.30GHz 2.30 GHz processor, at least 4 GB RAM, running on any operating system. Figure 3.[1] shows the simulation of the stacked ensemble model

```
stacking = StackingClassifier(
    estimators=[('knn', best_knn), ('nb', best_nb)],
    final_estimator=GaussianNB(),
    stack_method='predict_proba',
    cv=5,
    n_jobs=-1
)
```

Fig. 3: Simulation of the Optimized Ensemble Model

2.5 Performance Metrics and Evaluation

The models were evaluated using standard classification performance metrics, including Accuracy, Precision, Recall, F1-score, and Specificity.
TP refers to True Positives, TN - True Negatives, FP - False Positives, and FN - False Negatives

[1]Source code: https://github.com/jkisuule/ensemble_model

Accuracy measures the overall proportion of correctly classified student records.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

,
Precision determines the proportion of students predicted to drop out who actually dropped out.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall measures the model's ability to correctly identify students at risk of dropping out.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score provides a balanced measure of Precision and Recall.

$$\text{F1-score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

Specificity evaluates the model's ability to correctly classify students who are not at risk of dropping out.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Confusion matrices were employed to provide a detailed evaluation of the classification behavior of each model as seen in figure 7.

3 Results and Discussions

3.1 Results

3.1.1 Descriptive Statistics

The dataset consists of 4,424 student records with demographic, academic, financial, and macroeconomic attributes used as predictor variables. The statistical summary confirms the suitability of the dataset for supervised machine learning and provides an overview of the distribution of the study variables.

Table 1: Summary of Dataset Characteristics

Dataset Characteristic	**Value**
Total observations	4,424
Predictor variables	36
Target variable	Student Dropout
Study period	2021–2025
Training data	70% (3,096 records)
Testing data	30% (1,328 records)
Missing values	0
Duplicate records	Removed during preprocessing

3.1.2 Correlation Analysis

The correlation matrix shows both positive and negative associations among academic, demographic and economic variables, providing useful insights for feature selection and model development.

3.1.3 Exploratory Data Analysis

The analysis illustrates the distribution of student dropout status together with the behavior of important academic and demographic variables. These visualizations provide an understanding of the data characteristics before model training.

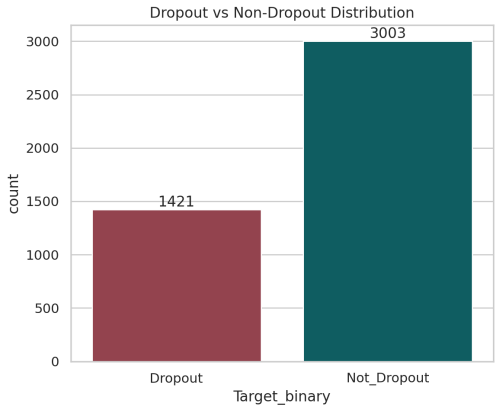


Fig. 4: Dropout Prediction

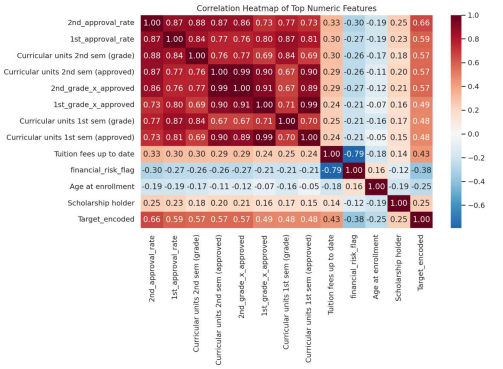
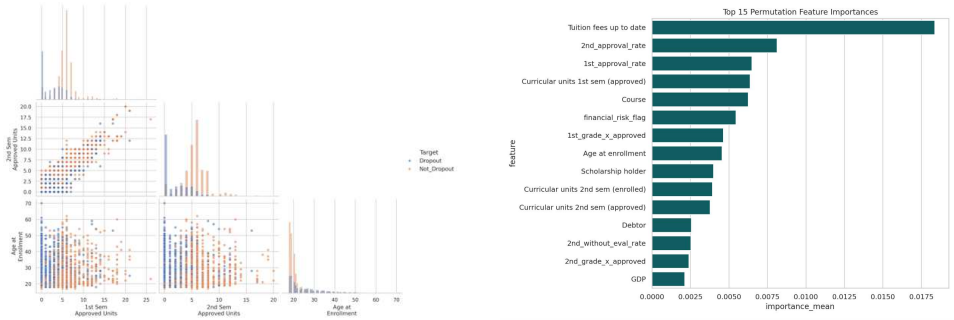


Fig. 5: Correlation matrix of the predictor variables



(a) Pair-plot of Key Predictor Variables

(b) Feature Importance Ranking

Fig. 6: Exploratory Data Analysis and Feature Importance

3.1.4 Feature Importance Analysis

The results indicate that academic performance variables, particularly the number of approved curricular units and cumulative academic achievement, contributed most significantly to the prediction process, while demographic and macroeconomic variables provided additional complementary information.

3.1.5 Model Performance Evaluation

The proposed stacking ensemble achieved the highest overall classification performance, outperforming the individual KNN, Bayesian Network, and Soft Voting Ensemble models. The improvement demonstrates the effectiveness of integrating heterogeneous classifiers through stacking ensemble learning.

Table 2: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)
KNN	82.7	70.8	78.5	74.4	78.5
Bayesian Network	84.6	75.8	76.3	76.1	76.3
Soft Voting Ensemble	85.2	75.8	79.4	77.6	79.4
Stacking Ensemble	**85.9**	**76.4**	**81.3**	**78.8**	**81.3**

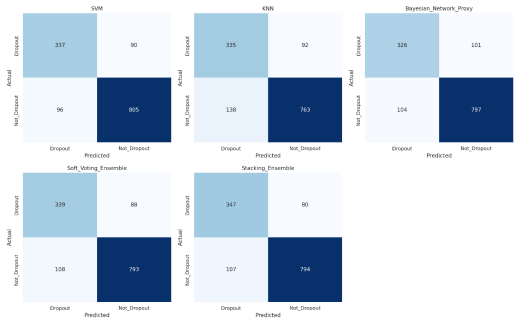


Fig. 7: Confusion Matrices for the Evaluated Models

3.1.6 Confusion Matrix Analysis

The stacking ensemble correctly classified a greater proportion of both dropout and non-dropout students while reducing false positive and false negative classifications compared with the baseline models.

3.2 Discussion

Class distribution analysis showed that 67.88% of the students did not drop out, while 32.12% dropped out, indicating a moderately imbalanced dataset that required careful model training. Correlation and exploratory analyses revealed meaningful relationships among the predictor variables, while feature importance analysis identified academic performance, approved curricular units, tuition fee status, and financial indicators as the most influential factors affecting student dropout. Performance evaluation further showed that the proposed stacking ensemble outperformed the K-Nearest Neighbors, Bayesian Network, and Soft Voting Ensemble models, achieving the highest accuracy (85.9%), precision (76.4%), recall (81.3%), F1-score (78.8%), and specificity (81.3%). The confusion matrix analysis also demonstrated that the proposed model reduced false classifications and improved the identification of students at risk of dropping out. Overall, these findings confirm that integrating more than one machine learning algorithms within a stacking framework produces more accurate and robust predictions than individual classifiers, providing a practical decision-support tool for implementing timely interventions aimed at improving student retention in higher education institutions.

4 Conclusion

This study developed and evaluated an optimized stacking ensemble model for forecasting student dropout using institutional data collected from The Unity University between 2021 and 2025. Following data preprocessing, feature engineering, class balancing, and model optimization, the proposed model was compared with KNN, Bayesian Network, and Soft Voting Ensemble classifiers. The experimental results demonstrated that the stacking ensemble achieved the best overall performance, attaining an accuracy of 85.9%, precision of 76.4%, recall of 81.3%, F1-score of 78.8%, and specificity of 81.3%. The findings further established that academic performance indicators, tuition fee status, debtor status, and other institutional variables are important predictors of student dropout, demonstrating the value of integrating heterogeneous machine learning algorithms for educational prediction tasks. Consequently, the proposed model provides a reliable decision-support tool that can assist universities in identifying students at risk of dropping out and implementing timely interventions to improve student retention and graduation outcomes. Future research should validate the proposed model using datasets from multiple higher education institutions, investigate the inclusion of additional behavioural and learning management system variables, compare the model with more advanced machine learning and deep learning techniques, and evaluate the long-term impact of machine learning-driven interventions on reducing student dropout rates in real educational environments.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BN	Bayesian Network
GPA	Grading Point Average
KNN	K-Nearest Neighbors
ML	Machine Learning
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine

Acknowledgments

Not applicable

Author Contributions

Kisuule Jackson

Contributed to the study conceptualization, software development, formal analysis, visualization, and the writing, and editing of the manuscript.

Bua Anthony

Contributed to the study conceptualization, methodology, technical analysis, interpretation of the results, and the critical review and editing of the manuscript.

Funding

This work is not supported by any organization or institution.

Data Availability Statement

The dataset generated during and/or analyzed during the current study is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval

Permission to access and use the institutional student data for this study was granted by The Unity University, Hargeisa, Somaliland, through the Office of the Academic Registrar[2]. The data were used strictly for academic research

[2] https://github.com/jkisuule/ensemble_model

purposes in accordance with applicable ethical, confidentiality, and institutional data protection requirements. Personally identifiable information was removed from the dataset prior to analysis, and the study was conducted using anonymized historical student records.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. K. Hoyos Osorio and G. Daza Santacoloma, "Predictive model to identify college students with high dropout rates," *Revista electrónica de investigación educativa*, vol. 25, 2023.

[2] A. Valencia-Arias, S. Chalela, M. Cadavid-Orrego, A. Gallegos, M. Benjumea-Arias, and D. Y. Rodríguez-Salazar, "University dropout model for developing countries: A colombian context approach," *Behavioral Sciences*, vol. 13, no. 5, p. 382, 2023.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 10, no. 3, p. e1355, 2020.

[4] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Science and Management*, vol. 8, no. 1, pp. 72–85, 2025.

[5] T. Abed, R. Ajoodha, and A. Jadhav, "A prediction model to improve student placement at a south african higher education institution," in

2020 International SAUPEC/RobMech/PRASA Conference, pp. 1–6, IEEE, 2020.

[6] L. H. Baniata, S. Kang, M. A. Alsharaiah, and M. H. Baniata, "Advanced deep learning model for predicting the academic performances of students in educational institutions," *Applied Sciences*, vol. 14, no. 5, p. 1963, 2024.

[7] R. Bean, *Comparative industrial relations: An introduction to cross-national perspectives*. Routledge, 2021.

[8] A. Chrysikos, E. Ahmed, and R. Ward, "Analysis of tinto's student integration theory in first-year undergraduate computing students of a uk higher education institution," *International journal of comparative education and development*, vol. 19, no. 2-3, pp. 97–121, 2017.

[9] S. Akter and M. O. F. Bhuiyan, "Dropout rate in online education during covid-19: A study of university students in bangladesh,"

[10] B. Duro, A. Gomes, F. B. Correia, A. R. Borges, and J. Bernardino, "Machine learning and deep learning for dropout prediction in higher education: A review," *Computers*, vol. 15, no. 3, p. 164, 2026.

[11] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, 2020.

[12] M. A. Hassan, A. H. Muse, and S. Nadarajah, "Predicting student dropout rates using supervised machine learning: Insights from the 2022 national education accessibility survey in somaliland," *Applied Sciences*, vol. 14, no. 17, p. 7593, 2024.

[13] T. V. Ramaano, "Developing a bayesian network model to predict students' performance based on the analysis of their higher education trajectory," Master's thesis, University of the Witwatersrand, Johannesburg (South Africa), 2024.

[14] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student retention using educational data mining and predictive analytics: a systematic literature review," *Ieee Access*, vol. 10, pp. 72480–72503, 2022.

[15] G. Deeva, J. De Smedt, and J. De Weerdt, "Educational sequence mining for dropout prediction in moocs: Model building, evaluation, and benchmarking," *IEEE Transactions on Learning Technologies*, vol. 15,

no. 6, pp. 720–735, 2022.

[16] J. Pecuchova and M. Drlik, "Predicting students at risk of early dropping out from course using ensemble classification methods," *Procedia Computer Science*, vol. 225, pp. 3223–3232, 2023.

[17] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, p. 105992, 2020.

[18] M. Rebelo Marcolino, T. Reis Porto, T. Thompsen Primo, R. Targino, V. Ramos, E. Marques Queiroga, R. Munoz, and C. Cechinel, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Scientific Reports*, vol. 15, no. 1, p. 9840, 2025.

[19] Z. Xu, D. Shen, Y. Kou, and T. Nie, "A synthetic minority oversampling technique based on gaussian mixture model filtering for imbalanced data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 3740–3753, 2022.

[20] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[21] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.