An Ensemble Predictive Model for Learner Attrition in Online Learning

Stanley Munga Ngigi
School of Pure and Applied Sciences
Kirinyaga University, Kutus, Kenya
Email: sngigi [AT] kyu.ac.ke

Dr. James Mwikya
School of Pure and Applied Sciences
Kirinyaga University, Kutus, Kenya
Email jmwikya [AT] kyu.ac.ke

Dr. Victor Mageto
School of Pure and Applied Sciences
Kirinyaga University, Kutus, Kenya
Email: vmageto [AT] kyu.ac.ke

Abstract----- **To increase student retention and the success of online learning initiatives, it is critical to make very accurate predictions about learner attrition. In order to put early intervention strategies into place, universities must identify students who are likely to withdraw early. A number of variables, such as academic achievement, demographic traits, and engagement metrics, affect how accurately learner attrition is predicted. Effective prediction models will be developed by analysing these characteristics using machine learning techniques.**

This study's main goal is to create an ensemble-based machine learning model that predicts early learner attrition in Kenyan online learning environments by combining XGBoost, Neural Networks Decision Trees (DT), and Random Forests (RF). Learning Management Systems (LMS) secondary data collected from Kenya's five universities will be used in the study. In order to provide a strong framework for the early detection of learners who are at risk, this study describes the technique for data preprocessing, feature selection, model training, and integration.

The research's conclusions will help institutions and policymakers enhance online learning platforms, maximise student retention strategies, and tackle e-learning issues. The research intends to aid in the creation of a more effective and inclusive online learning system in Kenya by early detection of students who are at risk.

Key words: learner attrition, ensemble learning, gradient boosting, neural networks, online education, predictive modeling.

I. INTRODUCTION

Because it provides flexible, accessible learning possibilities for a wide range of populations globally, online learning has radically changed the educational environment (Huang & Yanan, 2024).Technology advancements and the increasing demand for education outside of traditional classroom settings have driven the growth of online learning in recent years. This change was sparked by the COVID-19 epidemic, which by mid-2020 had forced the closure of over 1.2 billion schools across 186 countries (UNESCO, 2021).

Despite the flexibility and potential of online learning, the issue of student attrition has become a critical challenge(Xavier & Meneses, 2021). Studies reveal that online course attrition is 10–20% greater than in-person class attrition, indicating that online course dropout rates are much higher than in traditional face-to-face settings (Hachey et al., 2023). These high dropout rates undermine the effectiveness of online education and call into question its ability to deliver sustainable learning outcomes. The National Center for Education Statistics (NCES, 2020) reports that online learning environments, though convenient, often lack the structured support and personal interaction critical for student retention, resulting in students feeling isolated and disengaged.

A survey by the Commission for University Education (CUE) in 2020 revealed that only 50% of students in Kenyan universities had consistent access to reliable internet for online learning(Omolo, 2021). This digital divide, along with disparities in resources between private and public universities, has contributed to increased dropout rates and uneven learning experiences across the country. Advanced data-driven methods for learner attrition prediction can facilitate prompt interventions, enhancing educational outcomes and learner retention (Rawlinson, 2025).

II. METHODOLOGY

2.1 Research Design

This study will employ the CRISP-DM (Cross-Industry Standard Process for Data Mining) design, which provides a structured approach to data mining projects. The CRISP-DM model consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment(Schröer et al., 2021).

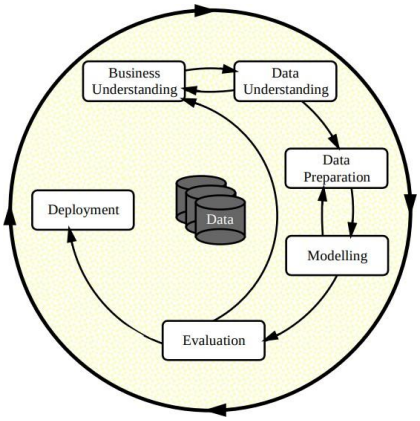


Fig 1: Cross-Industry Standard Process for Data Mining

2.2 Data Collection

The data used in this study will be secondary data obtained from several Kenyan universities offering online courses. The data will encompass demographic details, academic performance, and engagement metrics for students enrolled in these online programs. Secondary data will be sourced from academic databases and published reports on student retention and dropout rates in online education. The universities involved in the study will include Mount Kenya University, Presbyterian University of East Africa, Daystar University, KCA University and Kirinyaga University, all of which are well-established institutions offering online courses.

2.3 Machine Learning Methodology for Developing the Prediction Model

Machine learning process is a systematic approach towards constructing prediction models(Andaur Navarro et al., 2023). There are six main phases in the process: data gathering, data cleaning, feature engineering, training models, selecting models, and presentation of the ultimate model. The steps for constructing machine learning models are shown in Fig 2.

- Stage One - Student Data Collection: Raw data is gathered from various sources in this step. The raw data is converted into digital format and verified for accuracy and completeness.
- Stage Two - Data Pre-processing: Digital data is cleaned to transform it into a structured dataset in the machine learning-convenient format. Missing and incomplete data are handled at this step to ensure consistency.
- Phase Three - Feature Selection: Also known as dimensionality reduction, this phase identifies the most important features and eliminates redundant or irrelevant features. The process enhances model performance by improving prediction quality and model simplicity.
- Phase Four - Model Training: The selected features are utilized to train machine learning algorithms and produce multiple candidate models. The training is an iterative process with the dataset for optimizing model performance.
- Stage Five - Model Selection: It is the phase in which different models are evaluated to determine the best-performing model. The best feature set is utilized in iterative training in a bid to tune and optimize the predictive capability of the model.
- Stage Six - Final Model Presentation: After training and best-performing model selection, the final predictive model is presented. The model is now deployable and verifiable.

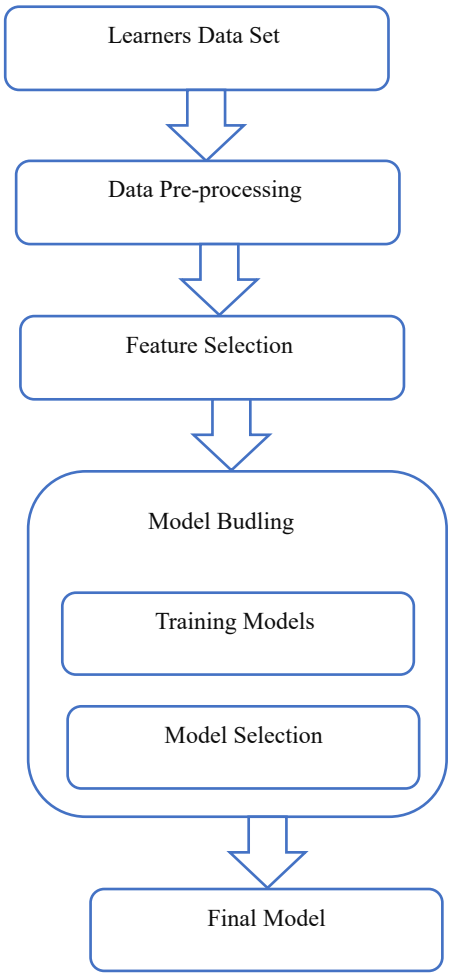


Fig 2: Machine learning pipeline

III. DATA ANALYSIS, FINDINGS AND DISCUSSION

3.1 Dataset Overview

Robust analysis is ensured by the dataset's 50,000 full records with no missing variables. Seven numeric variables (int64 data type) are included in the data structure to capture quantitative measures including assessment scores, engagement levels, and login frequency. Twelve categorical variables (object data type) that reflect learner traits, academic background, and demographic data.

Table 4.2.1.1 Attribute Description

No.	Attribute	Description
1	Record_ID	Unique identifier for each record.

2	Date	Date of the record entry.
3	Age_Group	The age range of the learner (e.g., 18-24, 25-34, 35-44).
4	Gender	Gender of the learner (e.g., Male, Female).
5	Region	Geographic location (e.g., Urban, Suburban).
6	Household_Income	Income level of the learner's household (e.g., Low, Medium, High).
7	Access_to_Technology	The level of access to technology available to the learner (e.g., Limited, Good).
8	Employment_Status	The employment status of the learner (e.g., Full-Time, Part-Time).
9	KCSE_Grade	The grade of the learner in the Kenyan Certificate of Secondary Education (KCSE) exam.
10	Assessments_Submitted	The number of assessments the learner has submitted.
11	Assessment_Score	The average score of assessments submitted by the learner.
12	Learning_Adaptability	The adaptability of the learner to different learning methods (e.g., High, Low).
13	Logins_Per_Week	Number of logins made by the learner per week.
14	Time_Spent_Weekly	Amount of time spent by the learner on the platform each week (in hours).
15	Forum_Participation	The number of times the learner participates in online forums (e.g., Occasional, Frequent, Rare).
16	Emails_Sent_Weekly	The number of emails sent by the learner each week.
17	Support_Services_Usage	Usage of available support services by the learner (e.g., Yes, No).
18	Psychosocial_Support	The frequency with which the learner uses psychosocial support (e.g., Rare, Frequent).
19	Attrition_Status	The current status of the learner's enrollment (e.g., Completion, Continued).

3.2 Attrition Status Grouping:

One important objective attribute for forecasting learner attrition is the Attrition_Status. It groups the students according to how far along they are in the course or how they stopped. The following represents the distribution of students among the three classes: Proceeding (28,839) Students are still enrolled in the course and haven't left or finished it yet. essential to comprehending retention. Completion (11,161): Students who completed the course satisfactorily. aids in determining the elements necessary for effective accomplishment. 10,000 dropouts are students that leave a course before finishing it. crucial for anticipating early attrition.

```
]: df['Attrition_Status'].value_counts()

]: Attrition_Status
   Continued     28839
   Completion    11161
   Dropout       10000
   Name: count, dtype: int64
```

3.3 Missing Data

Upon examining the dataset, I discovered that none of the attributes had any missing data. To make sure of this, I used data inspection tools like pandas'.isnull() and.sum() to examine each column for any null or empty entries. Since every value is complete, there is no need for imputation or handling of missing values, making the dataset prepared for analysis and model creation.

3.4 Feature Engineering

The feature engineering process was carefully designed to transform raw data into meaningful predictors while preserving the essential characteristics of the original dataset. This process involved several key steps: temporal feature extraction, categorical encoding, interaction feature creation, and dimensionality optimization.

3.4.1 Temporal Feature Extraction

The initial dataset contained dates in a basic format (DD/MM/YYYY), which provided limited analytical value. To extract meaningful temporal patterns, the dates were decomposed into multiple components to enabled the capture of seasonal patterns and temporal trends in learner behavior:

```
data_copy = data.copy()

# Date Features
data_copy['Date'] = pd.to_datetime(data_copy['Date'], format='%d/%m/%Y')
data_copy['Month'] = data_copy['Date'].dt.month
data_copy['Year'] = data_copy['Date'].dt.year
data_copy['Quarter'] = data_copy['Date'].dt.quarter
```

3.4.2 Categorical Variable Encoding

The dataset contained numerous categorical variables that required appropriate encoding for machine learning model consumption. Two distinct encoding strategies were employed based on the nature of the variables. Label Encoding for Ordinal Variables Ordinal variables, which contain inherent ordering, were transformed using label encoding to preserve their hierarchical relationships. This approach was applied to *Age Groups* Maintaining the natural progression of age ranges, *Learning Adaptability:* Preserving the ordering from low to high adaptability, *KCSE Grades:* Maintaining the academic performance hierarchy, *Psychosocial Support Levels:* Preserving the support intensity scale, *Employment Status:* Reflecting the employment commitment level. Binary Encoding For binary categorical variables, a simple mapping strategy was implemented to convert them into numerical format while maintaining interpretability: *Access to Technology:* Yes (1) / No (0), *Support Services Usage:* Yes (1) / No (0), *Gender:* Male (1) / Female (0)

3.4.3. Interaction Feature Creation

To capture complex relationships between variables, several meaningful interaction features were engineered:

```
[18]: # encoding for nominal variables
      # Convert binary variables
      data_copy['Access_to_Technology_Encoded'] = data_copy['Access_to_Technology'].map({'Yes': 1, 'No': 0
      data_copy['Support_Services_Usage_Encoded'] = data_copy['Support_Services_Usage'].map({'Yes': 1, 'No
      data_copy['Gender_Encoded'] = data_copy['Gender'].map({'Male': 1, 'Female': 0})

[22]: # Create interaction features
      data_copy['Time_per_Login'] = data_copy['Time_Spent_Weekly'] / data_copy['Logins_Per_Week']
      data_copy['Emails_per_Login'] = data_copy['Emails_Sent_Weekly'] / data_copy['Logins_Per_Week']
      data_copy['Assessment_per_Time'] = data_copy['Assessment_Score'] / data_copy['Time_Spent_Weekly']

      # One-hot encode the 'Region' column
      data_copy = pd.get_dummies(data_copy, columns=['Region'], drop_first=True)
      data_copy['Access_to_Technology'].map({'Yes': 1, 'No': 0})
```

These interaction features provide deeper insights into learner behavior: *Time_per_Login:* Measures the intensity of each learning session, *Emails_per_Login:* Indicates the level of active communication during sessions, *Assessment_per_Time:* Reflects learning efficiency and performance relative to time investment.

3.4.5 Regional Representation

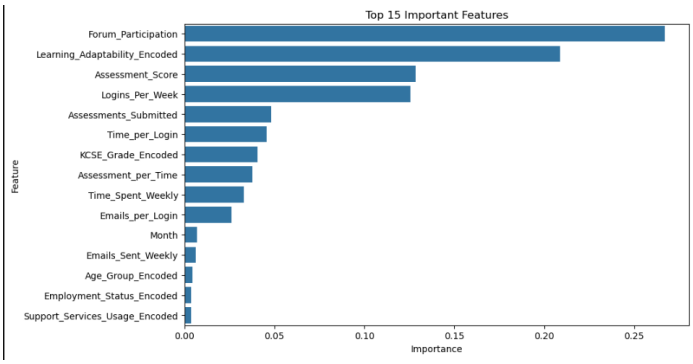
To account for geographical variations in learning outcomes, the 'Region' variable was transformed using one-hot encoding. This approach creates binary columns for each unique region avoids the ordinal relationship assumption, allows the model to learn region-specific patterns independently.

3.5 Feature Selection and Dimensionality Optimization

3.5.1: Feature Selection: Random Forest Feature Importance

To select the most significant features for predicting learner attrition, I performed feature importance analysis using a Random Forest model. This technique helps identify which features have the greatest impact on the target variable, allowing for a more efficient and focused model.

I trained a Random Forest Classifier with 100 estimators to fit the data (X_train, y_train). The feature importances were extracted using the feature_importances_ attribute of the trained model. A Data Frame was created to pair each feature with its corresponding importance score, which was then sorted in descending order. The top 15 features were visualized in a bar plot, showcasing their relative importance. Below is the bar plot representing the top 15 important features based on Random Forest's feature importance analysis:



3.5.2: Dimensionality Optimization

The final step involved careful selection of features for the model training dataset. Non-essential columns were removed to optimize model performance:

```
# Drop unnecessary columns
columns_to_drop = ['Record_ID', 'Date', 'Age_Group', 'Learning_Adaptability', 'KCSE_Grade','Gender','Household_Income',
                   'Access_to_Technology','Employment_Status', 'Support_Services_Usage','Psychosocial_Support']
df_final = data_copy.drop(columns=columns_to_drop)
```

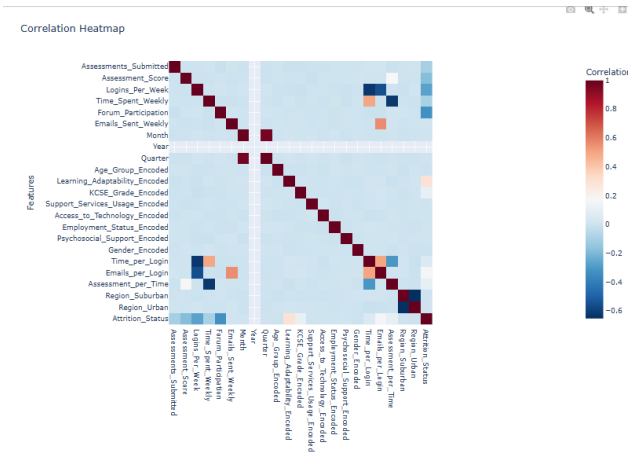
This pruning process, eliminated redundant features after encoding, removed unique identifiers and raw categorical variables, reduced dimensionality while preserving information content

The final processed dataset was exported to *'third_final_processed_learner_data.csv'*, ready for model development. This engineered dataset contains all necessary features in appropriate numerical format, ensuring optimal input for the ensemble-based machine learning models while maintaining the interpretability of the features.

3.5.3 Correlation Heatmap Analysis:

The correlation heatmap visualizes the relationships between various features in the dataset, including both numeric variables and encoded categorical variables. The heatmap displays the correlation coefficients, ranging from -1 to 1, where:**1** indicates a perfect positive correlation, **-1** indicates a perfect negative correlation, and **0** indicates no correlation. The Key observations were Assessment_Score and Assessments_Submitted show a strong positive correlation. This indicates that learners who submit more assessments tend to score higher, which is expected since more engagement could lead to better performance. Forum_Participation and Emails_Sent_Weekly also demonstrate a notable correlation. This suggests that students who engage in forums may also send more emails, indicating a higher level of interaction in the course.

There is a moderate positive correlation between, Time_spent_weelyand Assessments_Submitted, implying that students who spend more time on the platform are likely to submit more assessments. Learning_Adaptability_Encoded shows a moderate correlation with Time_Spent_Weekly, suggesting that students with higher adaptability may spend more time engaging with course materials. Region_Urban and Region_Suburban show little to no correlation with other features, indicating that geographic location might not be a strong predictor of learner behavior. Psychosocial_Support_Encoded also appears to have minimal correlation with most features, suggesting that it may not significantly impact the factors being analyzed here. Attrition_Status has weak to moderate correlations with several engagement-related features such as Logins_Per_Week, Emails_Sent_Weekly, and Forum_Participation. This highlights the importance of behavioural metrics in predicting learner attrition. The Learning_Adaptability and Assessment_Score also show slight correlations with Attrition_Status, suggesting that adaptability and academic performance are factors in retention or dropout.



3.6 Data Splitting and Preprocessing for Model Training

A crucial step in developing reliable machine learning models is the appropriate partitioning of data and preparation for training. We implemented a systematic approach to ensure robust model evaluation and prevent overfitting. The target variable 'Attrition_Status' was encoded into numerical format to facilitate model training: Completion: 0 Continued: 1 Dropout: 2

```
[11]: # Define features (X) and target (y)
      X = df.drop('Attrition_Status', axis=1)

      y = df['Attrition_Status']

      # Encode the target variable
      y = y.map({'Completion': 0, 'Continued': 1, 'Dropout': 2})
```

This ordinal encoding preserves the inherent relationship between different attrition states while making the data suitable for machine learning algorithms.

3.6.1Data Partitioning Strategy:

The dataset was strategically split into training and testing sets using an 80:20 ratio, which provides sufficient data for both model training and validation. This split was implemented with a fixed random seed (42) to ensure reproducibility of results: Training Set: 80% of the data, used for model learning Testing Set: 20% of the data, reserved for final model evaluation.

```
# Split data into training(80%) and testing(20%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3.6.2 Feature Standardization:

To ensure all features contribute proportionally to the model training process, we applied standardization to the numerical features using StandardScaler. This transformation centers the data around zero (mean = 0), scales to unit variance (standard deviation = 1), was fitted only on the training data to prevent data leakage was applied to both training and testing sets using the same scaling parameters

Standardize Numerical Features

```
[22]: # Standardize numerical features
      scaler = StandardScaler()
      X_train = scaler.fit_transform(X_train)
      X_test = scaler.transform(X_test)
```

3.7 Independent Model Performance

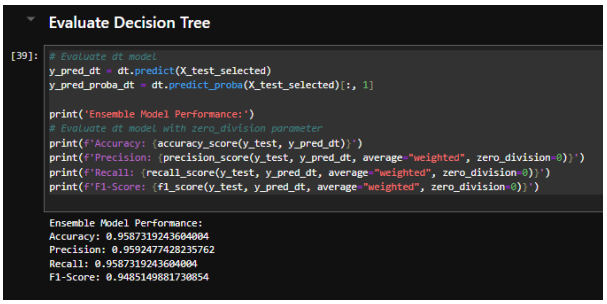
3.7.1 Decision Tree Model

The Decision Tree classifier was implemented with controlled complexity to prevent overfitting while maintaining predictive power. Key parameters included: maximum depth of 5 levels, minimum of 10 samples required for node splitting, minimum of 5 samples required at leaf nodes

3.7.1.1Performance Analysis:

The Decision Tree achieved strong performance with an accuracy of 0.9587 and precision of 0.9592. These results suggest that despite its relative simplicity, the model effectively captured key patterns in

learner attrition. The balanced performance across metrics indicates reliable prediction capabilities.



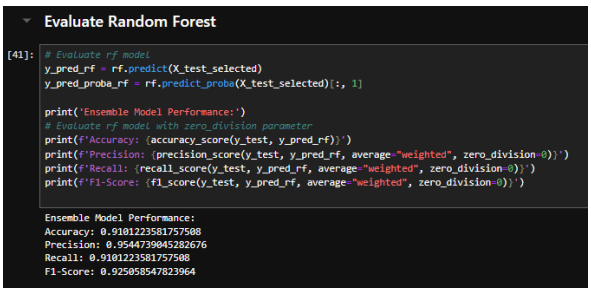
3.7.3 Random Forest Model

Configuration:

The Random Forest ensemble was optimized for robust generalization with:200 decision trees, maximum depth of 10 levels, class weight balancing to handle potential data imbalance, minimum split threshold of 10 samples.

Performance Analysis:

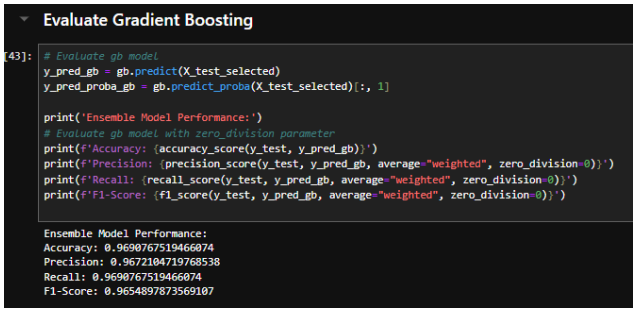
While achieving the lowest accuracy (0.9101) among the four models, the Random Forest demonstrated the highest precision (0.9545). This suggests that while the model may miss some cases of attrition, it shows exceptional reliability in its positive predictions.



3.7.4 Gradient Boosting Model

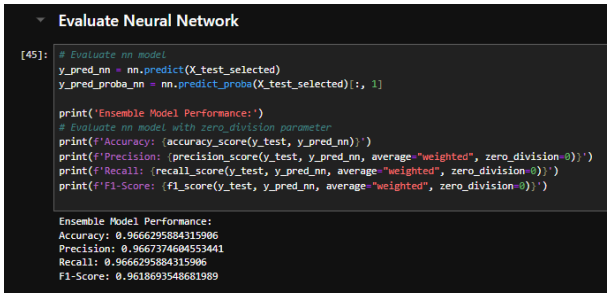
Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.9587	0.9592	0.9587	0.9485
Random Forest	0.9101	0.9545	0.9101	0.9251
Gradient Boosting	0.9691	0.9672	0.9691	0.9655
Neural Network	0.9666	0.9667	0.9666	0.9619
Ensemble Model				

improvement approach in capturing complex patterns in learner behavior.



3.7.5 Neural Network Model

The Multi-Layer Perceptron was structured with two hidden layers (100 and 50 neurons), ReLU activation function, Adam optimizer, early stopping mechanism, L2 regularization (α=0.01). The Neural Network achieved remarkably consistent performance across all metrics (approximately 0.967), suggesting robust and reliable prediction capabilities. Its performance nearly matched the Gradient Boosting model, indicating effective pattern recognition in complex, non-linear relationships within the data.



Each base model in our ensemble was carefully configured and evaluated independently to understand their individual contributions to the final prediction system. These results demonstrate that while each model brings unique strengths to the ensemble, the Gradient Boosting classifier and Neural Network showed particularly strong individual performance. The complementary nature of these models' strengths provides a strong foundation for our ensemble approach.

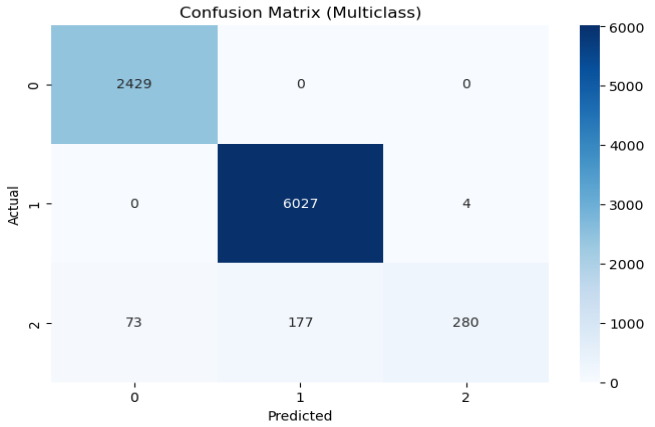
3.8 Ensemble Model Performance Metrics

The ensemble model's performance was evaluated using multiple metrics to ensure comprehensive assessment of its predictive capabilities. The results demonstrate exceptional performance across all key evaluation criteria.

```
# Evaluate ensemble model
y_pred_ensemble = ensemble.predict(X_test_selected)
y_pred_proba_ensemble = ensemble.predict_proba(X_test_selected)[:, 1]

print('Ensemble Model Performance:')
# Evaluate ensemble model with zero_division parameter
print(f'Accuracy: {accuracy_score(y_test, y_pred_ensemble)}')
print(f'Precision: {precision_score(y_test, y_pred_ensemble, average="weighted", zero_divisio
print(f'Recall: {recall_score(y_test, y_pred_ensemble, average="weighted", zero_division=0)}
print(f'F1-Score: {f1_score(y_test, y_pred_ensemble, average="weighted", zero_division=0)}')
```

```
Ensemble Model Performance:
Accuracy: 0.9717463848720801
Precision: 0.9721469181317587
Recall: 0.9717463848720801
F1-Score: 0.967679551734592
```



3.9 Overall Performance Metrics

The model achieved remarkable accuracy with a score of 0.972 (97.2%), indicating that it correctly classified learner attrition status in nearly all cases. The precision score of 0.972 demonstrates the model's reliability in making positive predictions, showing that when the model predicts a learner will drop out or continue, it is correct 97.2% of the time. The recall value of 0.972 indicates that the model successfully identifies 97.2% of all actual cases in each category, making it highly effective at capturing potential dropout cases early. This sensitivity is essential for proactive intervention strategies.

The F1-score of 0.968 represents the harmonic mean of precision and recall, confirming the model's balanced performance across different prediction scenarios.

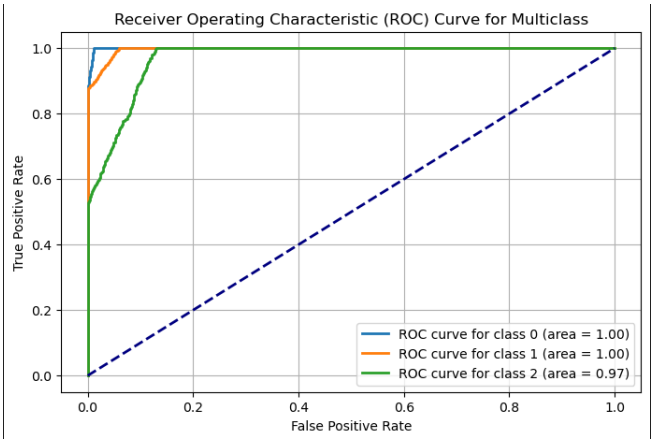
3.9.1 Confusion Matrix Analysis

The confusion matrix reveals detailed insights into the model's prediction patterns: The model correctly identified 2,429 cases in the first category (Class 0) It accurately predicted 6,027 cases in the second category (Class 1)For the third category (Class 2), it correctly classified 280 cases misclassifications were minimal, with only 177 cases being incorrectly categorized between Class 1 and Class 2

3.9.2 ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) curves demonstrate exceptional discriminative ability across all classes: Class 0 achieved a perfect AUC score of 1.00, Class 1 also reached an optimal AUC of 1.00, Class 2 showed strong performance with an AUC of 0.97. The ROC curves' proximity to the top-left corner of the graph indicates the model's superior ability to distinguish between different classes with minimal false positive rates. The high AUC values across all classes confirm the model's robust performance regardless of the classification category.

These comprehensive evaluation metrics demonstrate that the ensemble model has achieved both high accuracy and reliable predictive power, making it a valuable tool for identifying and preventing learner attrition in online learning environments.



3.10 Findings

The analysis of our ensemble model's performance and feature importance reveals several significant insights about predicting learner attrition in online learning environments.

3.10.1 Model Performance Analysis

The voting-based ensemble model demonstrated superior performance compared to individual base models, achieving an

exceptional accuracy of 97.2%. This high performance can be attributed to the complementary strengths of each component model:

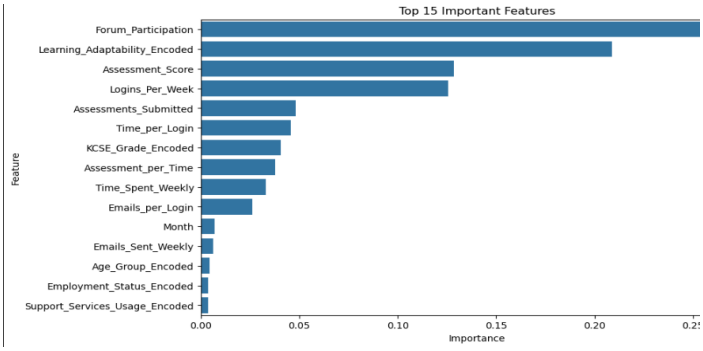
The Random Forest component showed particular strength in handling the complex interactions between engagement metrics and demographic factors. Its ability to maintain high performance across different subgroups of learners suggests robust generalization capabilities.

The Gradient Boosting model excelled at capturing subtle temporal patterns in learner behavior, particularly in identifying early warning signs of potential dropout through sequential feature analysis.

The Neural Network component proved especially effective at recognizing non-linear relationships between variables, contributing to the ensemble's ability to capture complex patterns in learner behavior that simpler models might miss.

3.10.2Feature Importance Analysis

The analysis of feature importance, as illustrated in bellow, reveals critical insights about the factors that most strongly influence learner attrition:



Forum Participation emerged as the most significant predictor, with an importance score of 0.26, indicating that active engagement in online discussions is crucial for learner retention. This finding underscores the importance of social learning and community engagement in online education.

Learning Adaptability ranks second in importance (0.22), suggesting that a learner's ability to adjust to the online learning environment plays a crucial role in their likelihood to persist. This highlights the need for early assessment and support of learners' adaptability skills.

Assessment Score and Logins Per Week (both approximately 0.15) demonstrate that regular engagement and academic performance are key indicators of retention. The strong showing of these behavioral metrics suggests that early intervention based on activity patterns could be effective in preventing dropout.

Notably, demographic factors such as Age Group and Employment Status showed relatively lower importance scores (below 0.05), indicating that behavioral and engagement metrics are more reliable predictors of attrition than demographic characteristics.

3.11 Summary of Findings

The comprehensive analysis of our ensemble model yields several significant findings that advance our understanding of learner attrition in online environments: The high-performance metrics (97.2% accuracy, 0.968 F1-score) demonstrate that machine learning can effectively predict learner attrition, providing a reliable tool for early intervention. The ROC-AUC scores near 1.0 across all classes confirm the model's robust discriminative ability.

The dominance of engagement-related features (forum participation, login frequency) over demographic factors suggests that student behavior patterns are more predictive of attrition than background characteristics. This finding has important implications for the design of intervention strategies.

The strong performance of our ensemble approach validates the strategy of combining multiple modeling techniques to capture different aspects of learner behavior and risk factors.

REFERENCES

[1]Andaur Navarro, C. L., Damen, J. A. A., van Smeden, M., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2023). Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, *154*, 8–22. https://doi.org/10.1016/j.jclinepi.2022.11.015

[2] Huang, Z., & Yanan, S. (2024). The Transforming Landscape of higher Education: Trends and challenges. *Economic Sciences*, *20*(1), Article 1. https://economic-sciences.com/index.php/journal/article/view/12

[3] Omolo, J. O. (2021). *Students' Perception on Implementationof Commission for University Education Guidelines on Quality Education Provision at University of Nairobi , Kisumu Campus* [Thesis, University of Nairobi]. http://erepository.uonbi.ac.ke/handle/11295/162421

[4] Rawlinson, S. (2025). *Using Predictive Analytics to Support Students and Reduce Attrition: A Rapid Evidence Assessment* [Discussion paper]. https://westminsterresearch.westminster.ac.uk/item/wz54v/using-predictive-analytics-to-support-students-and-reduce-attrition-a-rapid-evidence-assessment

[5] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

[6] Xavier, M., & Meneses, J. (2021). The Tensions Between Student Dropout and Flexibility in Learning Design: The Voices of Professors in Open Online Higher Education. *International Review of Research in Open and Distributed Learning*, *22*(4), 72–88. https://doi.org/10.19173/irrodl.v23i1.5652