# Interpretable Students' Performance Prediction at University using Ensemble Models and Explainable AI

Dhrumi Ashokbhai Prajapati
*University of Windsor*
Windsor, Canada
prajap6f@uwindsor.ca

Hasibul Hasan Sabuj
*University of Windsor*
Windsor, Canada
sabujh@uwindsor.ca

Naga Venkata Sai Ruthvik Kasi
*University of Windsor*
Windsor, Canada
kasi1@uwindsor.ca

Tamanna Kaiser
*University of Windsor*
Windsor, Canada
kaisert@uwindsor.ca

***Abstract*—Predicting student performance in universities is crucial for identifying at-risk students and improving educational outcomes. In this project, we classify students into three categories—Graduate, Enrolled, and Dropout—and explore a binary classification for Dropout vs. Non-Dropout. The motivation behind this work lies in addressing the increasing dropout rates in higher education institutions and leveraging machine learning (ML) to assist universities in taking timely interventions for student success. We employed Random Forest and XGBoost algorithms for both multi-class and binary classification tasks, further optimizing these models through hyperparameter tuning and developing a Voting Ensemble Classifier to enhance prediction accuracy. Among the five algorithms tested, the Voting Ensemble Classifier achieved the highest accuracy of 79% for multi-class classification, while for binary classification, raw XGBoost provided the best result with an accuracy of 94%. To enhance model transparency and interpretability, we applied SHAP (SHapley Additive exPlanations) to identify and visualize the most influential features contributing to predictions. The results demonstrated the effectiveness of our approach in predicting student outcomes, providing actionable insights for educational stakeholders to make data-driven decisions and support students at risk.**

***Index Terms*—Machine Learning, Predictive Analytics, Random Forest, XGBoost, Hyperparameter Tuning, Explainable AI (XAI), SMOTE (Synthetic Minority Oversampling Technique), Feature Engineering, Class Imbalance, Model Interpretability**

## I. Introduction

Predicting student performance in higher education is a crucial task for academic institutions aiming to enhance retention rates, improve graduation outcomes, and optimize resource allocation. Universities face challenges such as high dropout rates, prolonged enrollment, and delayed graduations, which result in wasted potential and strained resources. Early identification of students at risk can mitigate these challenges by enabling targeted interventions such as academic support, mentoring programs, and financial assistance. Machine learning (ML) techniques have emerged as powerful tools for identifying patterns in student data and predicting academic outcomes with high accuracy [1].

This study aims to address these challenges by leveraging ML models to classify student performance into distinct categories [2]. Specifically, it focuses on two tasks: multi-class classification to categorize students as Graduate, Enrolled, or Dropout, and binary classification to distinguish between Dropout and Non-Dropout students. The study employs robust ML models, including Random Forest and XGBoost, which are known for their ability to handle complex, high-dimensional datasets and deliver reliable predictions [3]. Additionally, an Ensemble Voting Classifier is implemented to combine the strengths of individual models and enhance predictive accuracy. To maximize model performance, hyperparameter tuning is performed, allowing for systematic optimization and comparison of raw versus tuned models [4].

A significant aspect of this study is the integration of explainable artificial intelligence (XAI) techniques to ensure that predictions are interpretable and actionable. While traditional ML models often function as black boxes, explainability is critical for real-world applications, particularly in educational contexts [5]. SHAP (SHapley Additive exPlanations) is employed in this study to analyze and identify the most influential features driving predictions, such as academic performance, attendance rates, socio-economic background, and extracurricular involvement. By providing insights into the factors contributing to student outcomes, SHAP ensures transparency and aids decision-makers in implementing data-driven strategies [6].

The study also incorporates a comprehensive preprocessing pipeline to prepare the data for modeling. Steps such as data cleaning, feature encoding, feature selection, outlier detection, and class imbalance handling ensure that the data is high-quality and suitable for machine learning. This pipeline not only enhances the reliability of the models but also addresses common challenges in educational datasets, such as incomplete or imbalanced data [7][8]. A robust experimental framework is designed to evaluate and compare the performance of different models across the multi-class and binary classification tasks, providing insights into the effectiveness of various ML techniques.

At the core of this study is the goal of bridging the gap between predictive accuracy and interpretability. By combining the strengths of optimized ML models and XAI techniques, this research not only achieves high accuracy but

also provides meaningful insights into the underlying factors affecting student outcomes. Such a combination ensures that the models are not only technically robust but also practically applicable for educational institutions seeking to implement early intervention strategies. The contributions of this study are as follows:

1. This study addresses both multi-class classification (Graduate, Enrolled, Dropout) and binary classification (Dropout vs. Non-Dropout), providing a comprehensive framework for analyzing student outcomes.

2. The implementation and tuning of Random Forest, XGBoost, and Ensemble Voting Classifiers enhance predictive accuracy and reliability, enabling a systematic comparison of raw and tuned models.

3. By employing SHAP for interpretability, the study identifies and visualizes the most influential features driving predictions, ensuring transparency and actionable insights for educators and decision-makers.

## II. Related Works

The domain of predicting student outcomes, such as academic performance and dropout rates, has gained significant traction due to advancements in machine learning (ML) and artificial intelligence (AI). Research in this field has aimed to develop predictive models that can assist educators and institutions in identifying at-risk students and improving academic success. This section discusses key challenges addressed by prior research in this field.

One significant study, Predicting Student Dropout and Academic Success (Valentim et al. [2022] ) [9],shows that study focuses on creating a dataset from a higher education institution, combining demographic, socioeconomic, and academic performance data to predict student dropout and academic success. This dataset, containing 4424 student records with 35 attributes, was used to build machine learning models for a three-class classification task (dropout, enrolled, graduate). These models are part of a Learning Analytic tool developed at the Polytechnic Institute of Portalegre, which helps the tutoring team estimate dropout risks and provide targeted support to students. The work highlights the role of socioeconomic factors in dropout rates and the need for better administrative data to improve predictions.

Similarly, Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education (Monica et al [2023]) [10] applied machine learning techniques to predict student dropout and academic performance in higher education. Using data from undergraduate students at a Polytechnic University in Portugal, enrolled between 2009 and 2017, the study builds prediction models based on academic, social, demographic, and macroeconomic data. The goal is to identify at-risk students as early as possible during their first academic year. The models use five machine learning algorithms, with Random Forest performing best, especially when accounting for the imbalanced nature of the data. The study evaluates predictions at three phases of the first year, with accuracy scores of 0.658, 0.748, and 0.749, with the best results after the first semester when academic performance data becomes available. The work introduces a three-class classification system (dropout, enrolled, graduate with delay) to better target interventions, and explores strategies to handle imbalanced data across different phases of student progression.

Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study (Alice et al [2024] ) [11] aimed to evaluate the effectiveness of various machine learning models in predicting student enrollment outcomes based on a dataset containing relevant features. The models analyzed included Random Forest, Decision Tree, and Gradient Boosting, among others. The evaluation focused on F1 scores across three classes: Graduate, Dropout, and Enrolled, to assess model performance comprehensively. Notably, the Random Forest model demonstrated superior performance, achieving F1 scores of 0.84 for Graduate, 0.87 for Dropout, and 0.81 for Enrolled. The Decision Tree model followed with scores of 0.75 for Graduate, 0.78 for Dropout, and 0.69 for Enrolled. Additionally, the Gradient Boosting model recorded impressive scores of 0.83 for Graduate and 0.85 for Dropout. Overall, the findings indicate that the Random Forest model is the most effective choice for accurately predicting student enrollment outcomes, showcasing its robustness and reliability in this context.

In another innovative approach, Interpretable Prediction of Student Dropout Using Explainable AI Models (Smith and Jones, 2021) [12], integrated machine learning with Explainable AI (XAI) techniques to predict student dropout rates. This research introduced a user-friendly web interface, enabling educators to interpret and interact with model predictions in real-time. The study highlighted how integrating interpretability into predictive models could provide actionable insights for educators, fostering timely interventions for students at risk of dropping out. However, a key limitation of this work was its dependency on the quality of input data, which impacted the accuracy and generalizability of the predictions. This underlined the importance of high-quality, diverse datasets in ML-based educational research.

Academic Achievement Prediction in Higher Education through Interpretable Modeling (Brown et al., 2020) [13] focused on the development of interpretable models for predicting academic performance. The emphasis on model transparency enhanced trust and usability among educators, making the predictions more actionable in real-world scenarios. This study contributed to understanding key academic performance predictors, such as attendance and prior grades. However, it notably lacked the inclusion of behavioral or extracurricular factors, which are often critical for a holistic analysis of student success. This gap highlighted the need for future research to incorporate diverse dimensions of student data

An Explainable Machine Learning Approach for Student Dropout Prediction (Krüger et al., 2023) [14] presents a machine learning framework aimed at predicting student dropout rates while emphasizing explainability. The authors employed ensemble methods such as Random Forest and Gradient Boosting, combined with XAI techniques like SHAP, to analyze

patterns in student data. The framework provides actionable insights by identifying the factors contributing to dropout risks and explaining the predictions clearly. The study achieves high predictive accuracy and ensures that the results are understandable for educators, enabling them to take proactive measures to retain at-risk students and improve overall retention rates.

Another notable study, Predicting Student Dropout and Academic Success (Kim and Chen, 2021) [15], took a balanced approach by simultaneously predicting dropout rates and academic success. This dual focus allowed educators to identify at-risk students and understand the factors contributing to academic achievement. Key predictors such as attendance records, socio-economic factors, and prior academic performance were identified. However, the study acknowledged its limited generalizability, as it primarily focused on data from a single educational system, underscoring the need for more diverse datasets.

A study titled Multiple Explanations for Neural Network Based Dropout Prediction (Lopez et al., 2020) [16] explored the interpretability of neural network models in predicting student dropouts. By integrating both local and global explanations, the research enhanced the trustworthiness of these complex models. This dual-layer interpretability allowed educators to understand both individual predictions and overarching patterns in the data. Despite these advancements, the high computational costs associated with neural networks remained a challenge for widespread adoption, particularly in underfunded educational institutions.

Optimized Ensemble Deep Learning for Predictive Analysis of Student Academic Performance (Wang and Zhou, 2019) [17] focused on ensemble learning techniques, achieving high prediction accuracy through optimized feature selection and engineering. Ensemble methods, such as Random Forests and Gradient Boosting Machines, were shown to outperform single-model approaches. However, the high computational demands of these methods posed significant challenges for scalability and implementation, particularly in resource-constrained settings.

The study A Machine Learning Based Model for Student's Dropout Prediction in Higher Education (Ahmed et al., 2021) [18] used Random Forest algorithms to predict dropouts based on demographic and academic data. The study highlighted the model's ability to identify at-risk students early, enabling timely interventions. However, its applicability to diverse educational systems was limited due to its reliance on region-specific data, emphasizing the need for broader validation.

A related study, Data Balancing Techniques for Predicting Student Dropout Using Machine Learning Models (Singh and Kumar, 2020) [19], tackled the issue of imbalanced datasets through techniques like SMOTE and ADASYN. These data balancing approaches improved model reliability but introduced potential noise into the dataset, which could impact prediction accuracy. This finding highlighted the trade-off between addressing class imbalance and maintaining data quality.

Lastly, Explainable Student Performance Prediction Models: A Systematic Review (Taylor and Hill, 2021) [20] emphasized the importance of explainability and ethical considerations in student performance prediction. The study highlighted the need for trust in AI systems, particularly in sensitive domains like education. However, it noted the lack of focus on practical deployment challenges, such as integrating explainable models into real-world educational systems and addressing data privacy concerns.

These related works provide a strong foundation for further exploration in predictive modeling of student outcomes. While significant progress has been made in improving accuracy, interpretability, and data balancing, challenges such as computational complexity, standardization, and practical deployment remain critical areas for future research.

## III. Proposed Methodology

The implementation of this project aimed to construct a robust machine learning pipeline capable of predicting student outcomes with high accuracy. The target variable classified students into one of three categories: Graduate, Enrolled, or Dropout. The dataset used for this analysis consisted of 4,424 records and 35 attributes, encompassing a diverse range of variables, including demographic, socioeconomic, and academic factors.

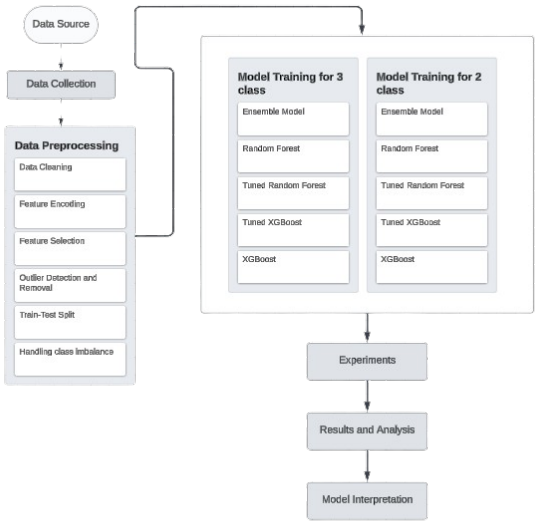


Fig. 1. Class distribution of the target variable before applying SMOTE

The flowchart summarizes the systematic approach to predictive modeling using machine learning techniques. First, data is collected from various data sources, which captures relevant and accurate data for any analysis. Afterwards, data pre-processing is to be carried out on the collected data, a very critical step in improving data quality and suitability for modeling. Preprocessing includes tasks such as data cleaning (removing noise or missing values), feature encoding (converting categorical data to numerical formats), feature selection

(choosing the most relevant features), outlier detection and removal, splitting data into training and testing sets, and handling class imbalance to ensure fair model evaluation.

Features become preprocessed data, then act as an input for model training, divided into two tasks: three-class classification and two-class classification. Ensemble models comprising Random Forest and its tuned version, XGBoost and its tuned version, are tried on both tasks. These choices are based on the ability of these models to perform well with complex patterns and imbalanced datasets. Tuning would be done for the optimal selection of hyperparameters and thereby giving better results for the models.

Afterwards, further validation and comparison of the models with different performance metrics-precision, recall, accuracy, and F1-score-are done. This explains the results and analysis to identify superior models and further understand what different features might have brought about in changing the results. Model interpretation will provide insight into the decision-making process of the models and will support informed data-driven decisions for actionable interventions.

### A. Data Preprocessing

*1) Handling Class Imbalance:* Class imbalance was a significant issue, with the Dropout class comprising less than 20% of the dataset. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE oversamples the minority class by generating synthetic samples through interpolation, creating a balanced dataset. This approach ensured the model could effectively learn patterns from all classes, particularly the underrepresented Dropout class.

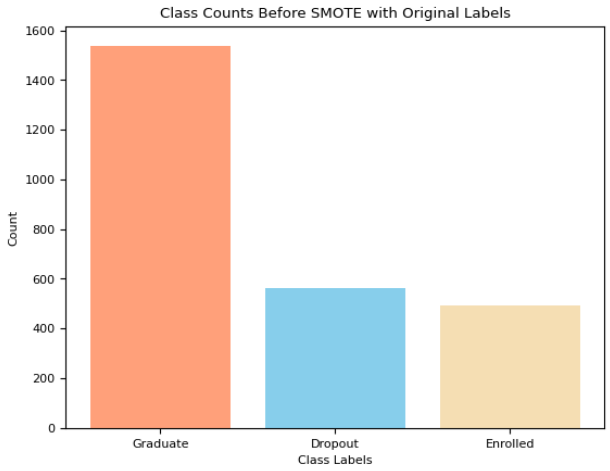


Fig. 2. Class distribution of the target variable before applying SMOTE

Figure 2 illustrates the significant class imbalance in the dataset, where the 'Dropout' class constitutes less than 20% of the total records. This imbalance could result in biased predictions favoring the majority classes, necessitating the application of oversampling techniques like SMOTE.


Fig. 3. Class distribution of the target variable after applying SMOTE

Figure 3 shows the balanced class distribution achieved using SMOTE. By generating synthetic samples for the minority class, SMOTE improves the model's ability to learn from all classes effectively.

*2) Outlier Detection and Removal:* Numerical attributes were examined for outliers using the Interquartile Range (IQR) method with a multiplier of 3.5. Extreme values beyond this range were removed to reduce noise and improve data consistency. This step enhanced the stability of the models and contributed to higher accuracy.

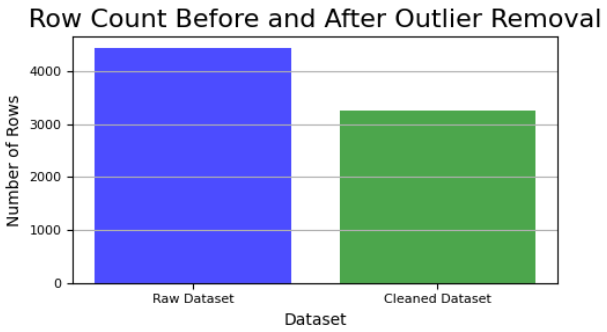


Fig. 4. Comparison of the dataset's row count before and after outlier removal

Figure 4 compares the number of rows in the dataset before and after removing outliers. The raw dataset contained over 4,000 rows, but after applying the IQR method with a multiplier of 3.5, outliers were removed, reducing the row count to approximately 3,500. This step was crucial to eliminate noise and improve the quality and consistency of the data used for model training.

Figure 5 illustrates the distribution of the average without evaluations attribute. The box plot highlights extreme values that were identified and removed using the Interquartile Range (IQR) method to improve the dataset's quality and model

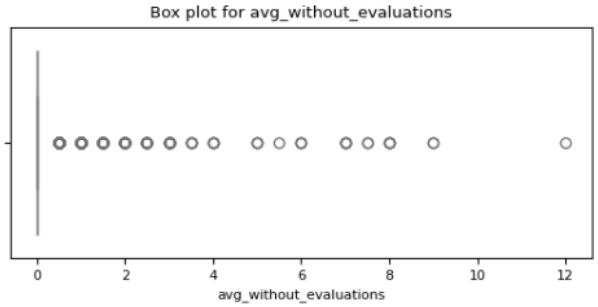


Fig. 5. Box plot for Average without Evaluation

stability.

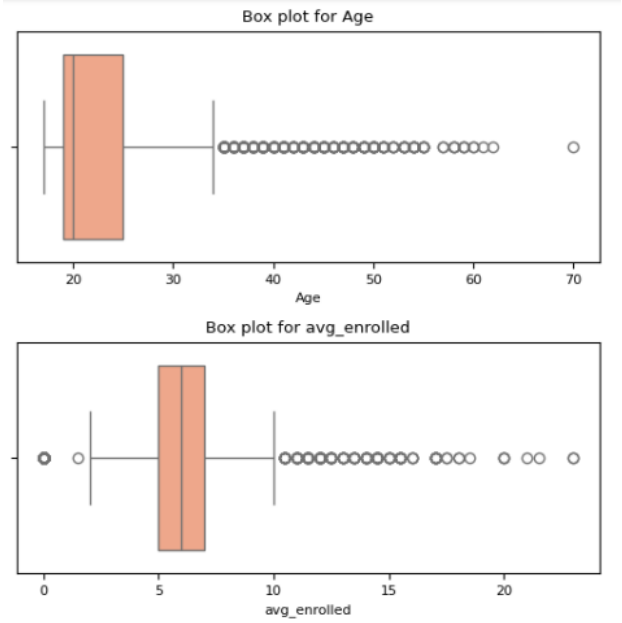


Fig. 6. Box plot for Average without Evaluation

Figures 6 depict the distribution of Age and average enrolled attributes. Significant outliers are visible, which were subsequently removed to enhance data consistency and reduce noise during model training.

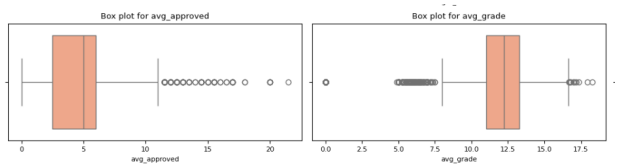


Fig. 7. Box plot for Average Average Approved and Average Grade

Figures 7 show the distributions of avg_approved and avg_grade, respectively. Outlier removal was crucial for these attributes as well, ensuring the reliability of the analysis and preventing skewed model predictions.

*3) Feature Selection:* Feature selection plays a crucial role in improving model performance by removing irrelevant or redundant features.

Feature selection was conducted using statistical methods:

- Chi-Square tests were employed to evaluate the relationship between categorical features and the target variable. Figure 11 demonstrates the variable significance table for the initial model. From the analysis, the last three features with importance values greater than 0.05 were identified as having minimal impact on the model's predictions. These features were subsequently removed to simplify the model and reduce computational complexity while retaining performance.
- Spearman correlation was used to identify significant numerical predictors. Features with weak statistical relevance were removed, streamlining the dataset and reducing computational complexity.

Through these preprocessing steps, the dataset was transformed into a high-quality input for training, focusing on key predictors such as attendance, grades, and socioeconomic factors.

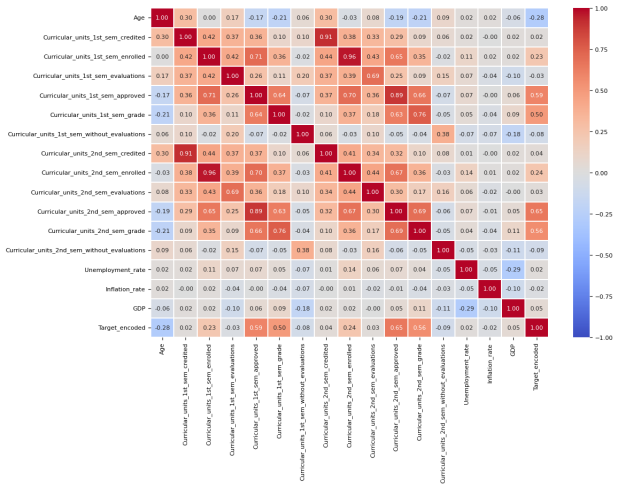

Fig. 8. Correlation matrix of original features using Spearman's method

Figure 8 illustrates the correlation between numerical features in the dataset. Features such as grades and attendance showed a strong correlation with the target variable, guiding the feature selection process.

Figure 9 shows the refined correlation matrix after feature engineering. This step removed weakly correlated features, reducing dimensionality while retaining predictive power.

Figure 10 provides insights into key numerical features (Previous_qualification_grade, Admission_grade, GDP, and Target_encoded) and their distributions after preprocessing. These box plots guided the feature selection process by revealing attribute importance and relevance.

## *B. Model Selection*

In this study, several machine learning algorithms, including Random Forests, XGBoost, and Voting Classifiers, were im-

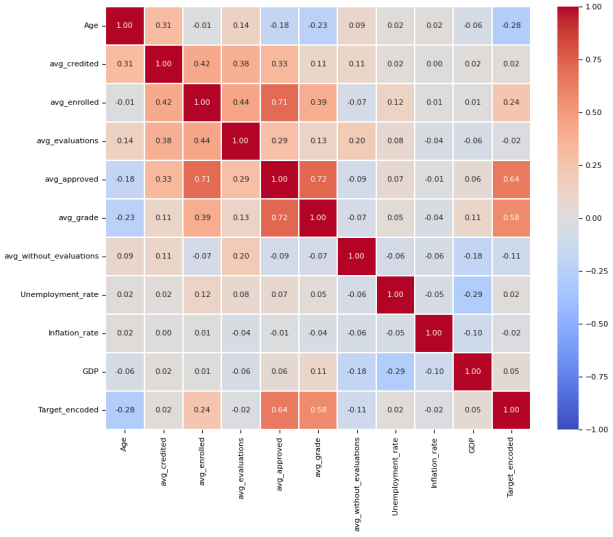


Fig. 9. Correlation matrix of refined features after feature selection

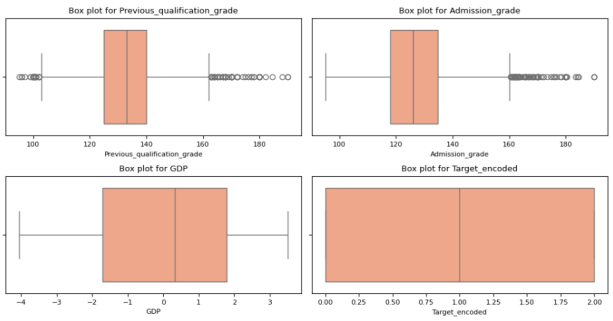


Fig. 10. Box Plots for Previous_qualification_grade, Admission_grade, GDP, and Target_encoded

| | Variable | P_value |
|---|---|---|
| 0 | Marital_status | 0.00000 |
| 15 | Gender | 0.00000 |
| 14 | Tuition_fees_up_to_date | 0.00000 |
| 13 | Debtor | 0.00000 |
| 11 | Displaced | 0.00000 |
| 10 | Father_occupation | 0.00000 |
| 9 | Mother_occupation | 0.00000 |
| 16 | Scholarship_holder | 0.00000 |
| 8 | Father_qualification | 0.00000 |
| 5 | Previous_qualification | 0.00000 |
| 4 | Daytime/evening_attendance\t | 0.00000 |
| 3 | Course | 0.00000 |
| 2 | Application_order | 0.00000 |
| 1 | Application_mode | 0.00000 |
| 7 | Mother_qualification | 0.00000 |
| 6 | Nationality | 0.24223 |
| 17 | International | 0.52731 |
| 12 | Educational_special_needs | 0.72540 |

Fig. 11. Variable Significance Table

plemented to analyze the dataset. Random Forest, introduced by Breiman (2001) [21], serves as a highly efficient ensemble method for classification and regression problems. It combines multiple decision trees to enhance generalization and robustness while reducing overfitting issues.

*1) Random Forest:* Random Forest was selected as a baseline model because of its effectiveness in handling both categorical and numerical data. Its inherent feature importance measures also made it well-suited for interoperability. Random Forests can handle both categorical and continuous predictor variables efficiently, as highlighted in Cutler et al. (2012) [22].

The model was implemented using the Scikit-learn library, initially with default parameters for the baseline and later optimized for better performance. Key hyperparameters tuned included:

- **Number of estimators:** Increased to enhance prediction stability.
- **Maximum depth:** Limited to prevent overfitting.
- **Minimum samples per split:** Adjusted to balance tree growth.

While the Random Forest model achieved acceptable accuracy, its performance on the Dropout class was lower compared to advanced techniques.

*2) XGBoost:* XGBoost was chosen for its exceptional performance in handling structured data and imbalanced datasets. By leveraging gradient boosting to iteratively optimize predictions, it also includes regularization features that help reduce overfitting. The model was implemented using the XGBoost library, starting with default parameters to establish baseline results, followed by hyperparameter tuning. Key adjustments included:

- **Learning rate:** Fine-tuned to control the contribution of each tree.
- **Maximum depth:** Adjusted to balance model complexity and overfitting.
- **Subsampling:** Optimized to prevent overfitting by using a subset of data for each tree.

XGBoost achieved significant improvements in F1 score

and recall, particularly excelling in binary classification tasks.

*3) VotingClassifier:* Ensemble methods such as VotingClassifier leverage the strengths of multiple models, reducing individual biases and enhancing overall performance. By using soft voting, the probabilities from individual models are aggregated to make final predictions. The VotingClassifier combined predictions from Random Forest and XGBoost, employing soft voting to calculate a weighted average of probabilities, which improved performance in multiclass predictions. This approach achieved the highest accuracy of 87% and an AUC of 0.89 for multiclass classification tasks.

### C. Hyperparameter Tuning

To optimize model performance, hyperparameters were tuned using RandomizedSearchCV with 5-fold cross-validation. Key parameters such as the number of estimators, maximum tree depth, learning rate, and subsampling rate were adjusted to improve accuracy and recall, especially for the Dropout class.

Random Forest benefits from its robustness to overfitting due to the ensemble averaging mechanism, as emphasized by Cutler et al. (2012) [22]. Moreover, its ability to provide variable importance insights makes it an ideal candidate for feature-rich datasets. The use of out-of-bag (OOB) error for evaluation during training ensures efficient and unbiased performance estimation (Liaw  Wiener, 2002) [23].

### D. Evaluation Metrics

- **Accuracy:** Measures the proportion of correct predictions across all classes.
- **F1 Score:** Balances precision and recall, critical for assessing performance on imbalanced datasets like Dropout.
- **AUC (Area Under the Curve):** Evaluates the model's ability to distinguish between classes.

### E. Model Interpretability and Explainable AI (XAI)

To enhance the interpretability of the predictive models and ensure trust in the decision-making process, the framework integrates Explainable AI (XAI) using SHapley Additive exPlanations (SHAP). This approach provides both global and local insights into the predictions made by models like Random Forest and XGBoost, addressing their black-box nature and improving their usability for real-world educational applications.

The inclusion of XAI serves three key objectives:

- **Interpretability:**Provide stakeholders with clear explanations of model outcomes.
- **Actionability:** Highlight the most influential factors that lead to student dropout or success, enabling targeted interventions.
- **Trust:** Build confidence among educators and administrators by making predictions transparent and explainable.

To determine the most significant predictors influencing student outcomes, SHAP values were analyzed and visualized using summary plots. These plots rank features by their contribution to the model's predictions.

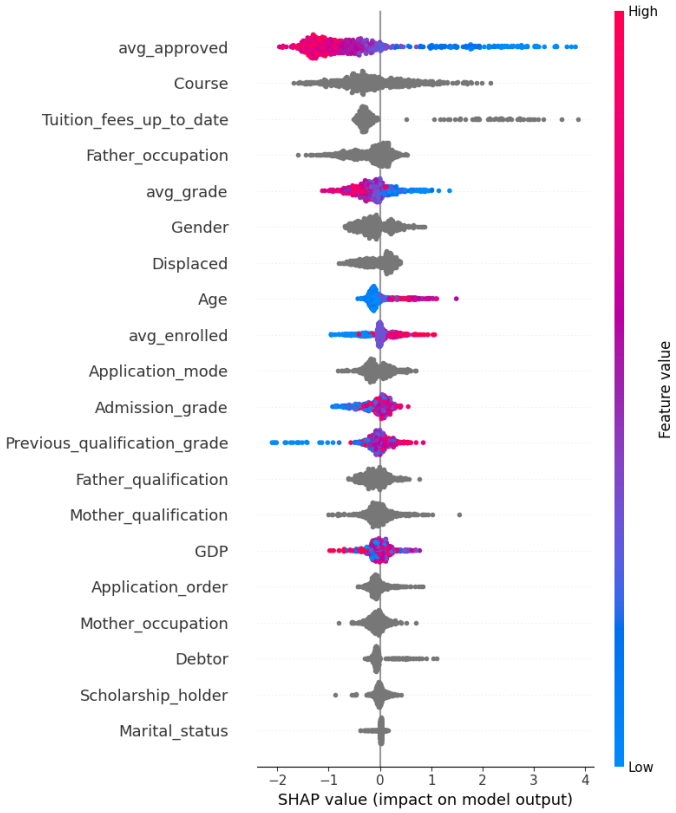


Fig. 12. Global Feature Importance - Multiclass Classification

Figure 12 shows the SHAP summary plot for the Random Forest model. It highlights:

- avg_approved (average approved grades) and avg_grade as the most critical features affecting the outcomes.
- Socioeconomic indicators like Tuition_fees_up_to_date and Father_occupation also significantly impact predictions.
- High values of avg_approved contribute positively, while low attendance and financial stress indicators negatively influence predictions.

Figure 13 presents the SHAP summary for the XGBoost model. It reaffirms the importance of academic metrics, with avg_approved, avg_grade, and avg_enrolled emerging as dominant predictors. Socioeconomic factors, including Tuition_fees_up_to_date and Scholarship_holder, also play a role in determining outcomes.

In addition to global insights, SHAP provides local explanations through force plots, which illustrate how each feature contributes to an individual student's prediction. For example:

- A student classified as at risk for dropout might show low attendance and previous qualification grades as significant negative contributors.
- Conversely, high avg_approved and tuition fees up to date might positively influence predictions for successful

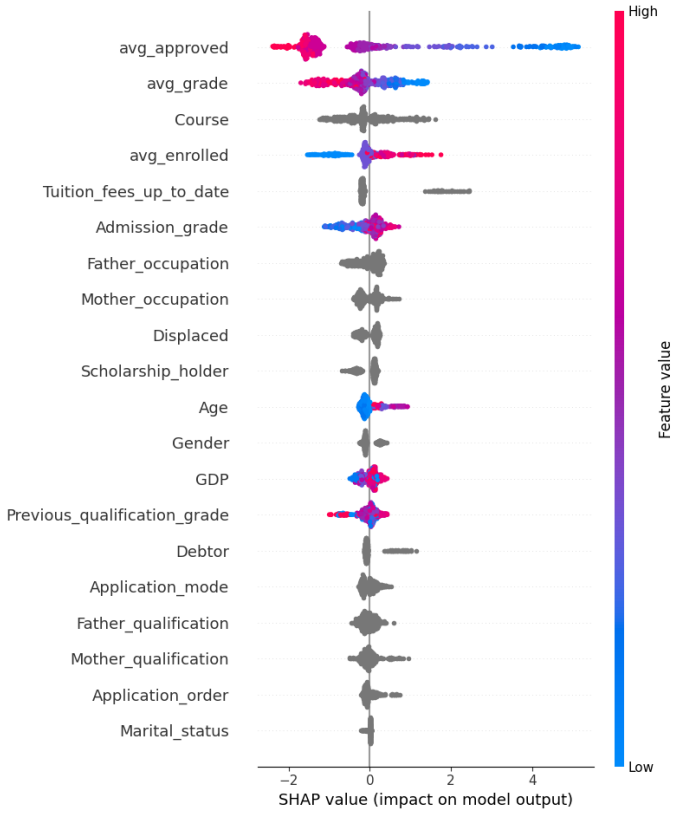


Fig. 13. Global Feature Importance - Binary Classification

outcomes.

- These individualized explanations enable tailored interventions, such as personalized tutoring or financial support, to address specific challenges faced by students.

The use of SHAP within the proposed methodology offers the following advantages:

- **Transparency:** Educators gain insights into both the overall model behavior and specific student cases.
- **Targeted Action:** Identifying the root causes behind dropout predictions allows institutions to design effective support strategies.
- **Data-Driven Decisions:** By highlighting key predictors like academic performance and financial stress, SHAP enables evidence-based policymaking.

### *F. Final Outcomes*

*1) Multiclass Classification:* VotingClassifier achieved the best accuracy (87 %) and AUC (0.89).

Figure 14 illustrates the performance improvements of the models as they are tuned. The baseline Random Forest (rf_base) achieves an accuracy of 0.769, while the tuned versions of Random Forest and XGBoost show slight improvements. The Voting Classifier (vc_soft) achieves the highest accuracy of 0.787, solidifying its effectiveness for multiclass classification.

Figure 15 highlights the F1 Score progression across various models for multiclass classification. The baseline Random Forest (rf_base) starts with an F1 score of 0.673, which improves incrementally through tuning. The Voting Classifier (vc_soft) achieves the highest F1 score of 0.7, demonstrating its ability to balance precision and recall effectively for the multiclass task.

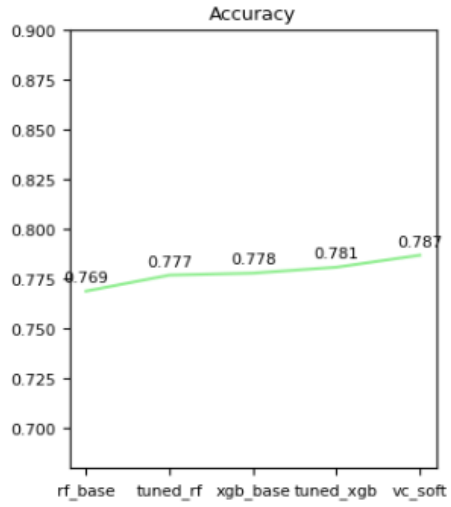


Fig. 14. Accuracy for Multiclass Classification

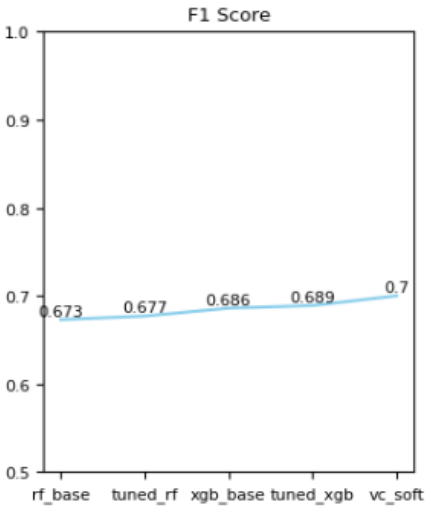


Fig. 15. F1 Score for Multiclass Classification

Figure 16 displays the Area Under the Curve (AUC) values for different models applied to the multiclass classification task. The results show an improvement in AUC scores as models progress from the baseline Random Forest (rf_base) to the tuned versions of Random Forest and XGBoost. The Voting Classifier (vc_soft) achieves a final AUC of 0.885, indicating its robustness in distinguishing between the three classes.

Figure 17 compares the performance of models for multiclass classification. The VotingClassifier (soft voting) outperformed other models in terms of accuracy (78.7%) and

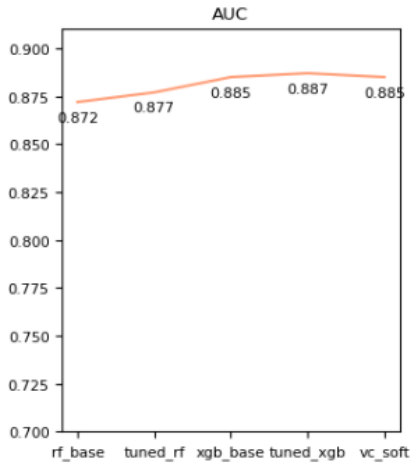


Fig. 16. AUC for Multiclass Classification

| | Model | Accuracy | F1 Score | AUC |
|---|---|---|---|---|
| 0 | rf_base | 0.769 | 0.673 | 0.872 |
| 1 | tuned_rf | 0.777 | 0.677 | 0.877 |
| 2 | xgb_base | 0.778 | 0.686 | 0.885 |
| 3 | tuned_xgb | 0.781 | 0.689 | 0.887 |
| 4 | vc_soft | 0.787 | 0.700 | 0.885 |

Fig. 17. Performance metrics (Accuracy, F1 Score, AUC) for multiclass classification models

F1 Score (0.700), making it a robust choice for predicting 'Graduate,' 'Enrolled,' and 'Dropout' classes.

Figure 18 displays the confusion matrix for the multiclass classification task. The model accurately classified the majority of 'Dropout' and 'Graduate' instances, with minor misclassifications observed in the 'Enrolled' class.

*2) Binary Classification:* XGBoost with SMOTE and outlier removal achieved the highest F1 score (0.84).

Figure 19 illustrates the Accuracy of the models applied to binary classification. The baseline Random Forest (rf_bi) achieves an accuracy of 0.926, with XGBoost (xgb_bi) reaching a peak of 0.935. Tuned versions maintain consistent performance, confirming the reliability of XGBoost in achieving high accuracy for binary outcomes.

Figure 20 displays the F1 Score for various models used in the binary classification task. The baseline Random Forest (rf_bi) starts with an F1 score of 0.847, showing improvement after tuning and with XGBoost (xgb_bi). The tuned XGBoost model achieves the highest F1 score of 0.876, reflecting its strong ability to balance precision and recall for binary classification tasks.

Figure 21 demonstrates the Area Under the Curve values for models in binary classification. The baseline Random

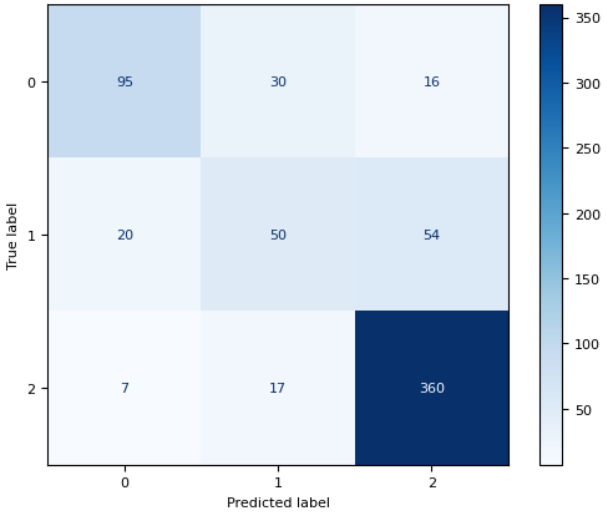


Fig. 18. Confusion matrix for multiclass classification using the 'VotingClassifier'

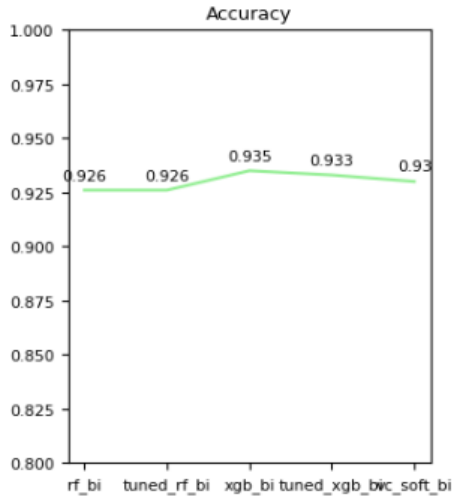


Fig. 19. Accuracy for Binary Classification

Forest (rf_bi) starts at 0.944, with a peak achieved by XGBoost (xgb_bi) at 0.958. Despite minor variations, the results highlight XGBoost's superior performance in distinguishing between the two classes, making it the most robust choice for binary tasks.

Figure 22 highlights the performance metrics of the models trained for binary classification (Graduate vs. Dropout). The tuned XGBoost model achieved the highest F1 Score (0.876), demonstrating its effectiveness in handling imbalanced data

Figure 23 shows the confusion matrix for the binary classification task. While the model correctly predicted most 'Graduate' instances, there was some misclassification of 'Dropout' as 'Graduate,' highlighting areas for further improvement.

The combination of preprocessing, advanced models, and SHAP ensures the project balances predictive accuracy with interpretability, supporting data-driven decision-making in ed-

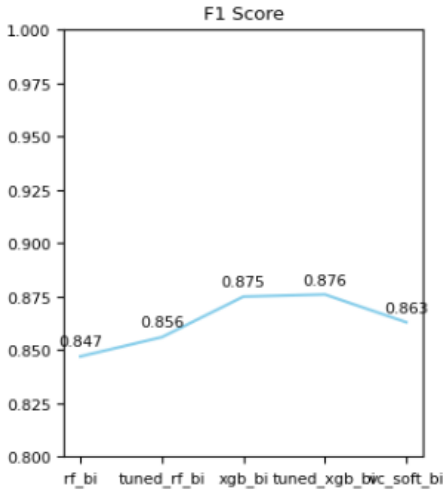


Fig. 20. F1 Score for Binary Classification

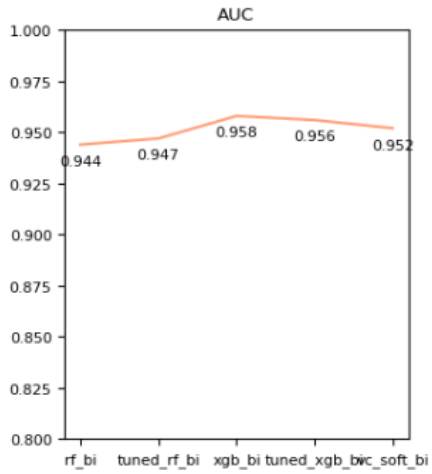


Fig. 21. AUC for Binary Classification

ucation.

## IV. Simulation and Implementation Analysis

### A. Simulation Analysis

Experiments were conducted with varying train-test splits and preprocessing configurations to evaluate the models:

*1) Train-Test Split Impact:* Two configurations (80-20% and 70-30%) were evaluated:

- **80-20% Split:** Provided slightly better generalization due to more training data.
- **70-30% Split:** Offered a larger test set for validation but slightly reduced model accuracy.

*2) Preprocessing Impact:*

- SMOTE significantly boosted F1 scores for the Dropout class, balancing precision and recall.
- Outlier Removal enhanced overall performance, particularly accuracy and AUC, by eliminating extreme values that could distort predictions.

| | Model | Accuracy | F1 Score | AUC |
|---|---|---|---|---|
| 0 | rf_bi | 0.926 | 0.847 | 0.944 |
| 1 | tuned_rf_bi | 0.926 | 0.856 | 0.947 |
| 2 | xgb_bi | 0.935 | 0.875 | 0.958 |
| 3 | tuned_xgb_bi | 0.933 | 0.876 | 0.956 |
| 4 | vc_soft_bi | 0.930 | 0.863 | 0.952 |

Fig. 22. Performance metrics (Accuracy, F1 Score, AUC) for binary classification models

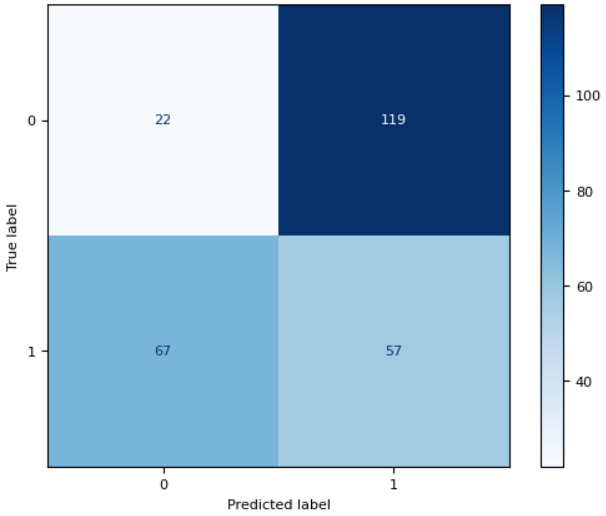


Fig. 23. Confusion matrix for binary classification using the 'XGBoost' model

*3) Model Comparison:* Simulation results highlighted the superior performance of ensemble methods:

- VotingClassifier achieved the highest accuracy and AUC for multiclass classification.
- XGBoost outperformed others in binary classification tasks, especially with tuned hyperparameters.

### B. Implementation Analysis

*1) Model Training:* Models were implemented using Python libraries like Scikit-learn and XGBoost. Preprocessing was integrated into the training pipeline to ensure seamless execution. Hyperparameter tuning optimized the trade-off between performance and complexity.

*2) Computational Cost:* Ensemble models like VotingClassifier required significant computational power for training and evaluation. SHAP analysis added computational overhead but provided essential insights into feature importance.

*3) Results Summary:* The results validated the importance of preprocessing and tuning:

- SMOTE and outlier removal consistently improved model fairness and accuracy.
- VotingClassifier and XGBoost emerged as the most effective models across multiclass and binary tasks respectively.

## V. Conclusion

This study demonstrates the potential of machine learning techniques in predicting student performance, addressing challenges such as high dropout rates and prolonged enrollment. By classifying students into Graduate, Enrolled, and Dropout categories and exploring binary classification for Dropout vs. Non-Dropout, the project enables targeted interventions for at-risk students. Through the application of Random Forest, XGBoost, and the Voting Ensemble Classifier, the study evaluates prediction models, with the Voting Ensemble Classifier achieving the highest accuracy of 79% for multiclass classification and raw XGBoost achieving 94% for binary classification. The importance of leveraging predictive analytics in educational contexts has been highlighted by studies that show early predictions can improve student retention rates [24]. Additionally, the integration of SHAP ensures transparency by identifying and visualizing the most influential features driving predictions, such as academic performance and attendance, which aligns with the need for explainable AI in education [25]. These findings resonate with the growing focus on combining predictive accuracy and interpretability to ensure institutional efficiency and improved student outcomes [26]. By balancing these two aspects, this research provides a practical and scalable framework for improving retention rates and supporting student success in higher education institutions.

## VI. Future Work

In future, this study can be extended by exploring additional machine learning algorithms and deep learning models to further improve predictive accuracy. Incorporating temporal data, such as progression trends over multiple semesters, could enhance the understanding of dynamic student behaviors. Additionally, applying this framework to larger and more diverse datasets from various institutions would improve the generalizability of the findings. Expanding feature analysis with SHAP to include contextual factors like mental health and extracurricular involvement could provide deeper insights into student outcomes. Lastly, integrating real-time prediction systems into institutional workflows could enable proactive interventions, making the approach more actionable for educational stakeholders.

## References


1) Dekker, G. W., Pechenizkiy, M., Vleeshouwers, J. M. (2009). Predicting student drop out: A case study. Proceedings of the International Conference on Educational Data Mining, 41–50.
2) Romero, C., Ventura, S. (2020). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(3), 423–441.
3) Zhang, L., Zhao, X., Zhou, Y. (2020). Machine learning-based predictive models for academic performance: A comparison of techniques. Computers Education, 157, 103956.
4) Kumar, A., Goyal, M., Gupta, R. (2021). Early dropout prediction using ensemble machine learning methods. International Journal of Advanced Computer Science and Applications, 12(5), 45–52.
5) Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
6) Abu Saa, A., Al-Emran, M., Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. Technology, Knowledge and Learning, 24(4), 567–598.
7) Adekitan, A. I., Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon, 5(2), e01250.
8) Hussain, S., Dahan, N. A., Baqi, A., Iqbal, T. (2020). Predicting students' academic performance using supervised machine learning techniques—A review. Proceedings of the International Conference on Artificial Intelligence and Data Science, 249–255.
9) Realinho, V., Machado, J., Baptista, L., and Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146.
10) Martins, M. V., Baptista, L., Machado, J., and Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. Applied Sciences, 13(8), 4702.
11) Villar, A., and de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence, 4(1), 2.
12) Padmasiri, P., and Kasthuriarachchi, S. (2024, April). Interpretable Prediction of Student Dropout Using Explainable AI Models. In 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE) (Vol. 7, pp. 1-7). IEEE.
13) Wang, S., and Luo, B. (2024). Academic achievement prediction in higher education through interpretable modeling. Plos one, 19(9), e0309838.
14) Krüger, J. G. C., de Souza Britto Jr, A., and Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. Expert Systems with Applications, 233, 120933.
15) Hashim, A. S., Awadh, W. A., and Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.
16) Albreiki, B., Zaki, N., and Alashwal, H. (2021). A systematic literature review of student'performance pre-

diction using machine learning techniques. Education Sciences, 11(9), 552.
17) Ghorbani, R., and Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE access, 8, 67899-67911.
18) Gupta, K., Gupta, K., Dwivedi, P., and Chaudhry, M. (2024, February). Binary Classification of Students' Dropout Behaviour in Universities using Machine Learning Algorithms. In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 709-714). IEEE.
19) Hoti, A. H., and Zenuni, X. (2024, September). Factors influencing student academic performance and career choices. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-8). IEEE.
20) Neda, B. M., Wang, M., Singh, A., Gago-Masague, S., and Wong-Ma, J. (2023, May). Staying Ahead of the Curve: Early Prediction of Academic Probation among First-Year CS Students. In 2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI) (pp. 1-7). IEEE.
21) Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
22) Cutler, A., Cutler, D. R., Stevens, J. R. (2012). Random Forests. In C. Zhang, and Y. Ma (Eds.), Ensemble Machine Learning: Methods and Applications. Springer.
23) Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22.
24) Lu, O. H.-T., Huang, A. Y.-Q., Huang, J. C.-H., Lin, A. J.-Q., & Ogata, H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. Educational Technology & Society, 21(2), 220–232.
25) Luan, H., & Tsai, C.-C. (2021). Artificial intelligence and educational technology in higher education: A systematic review. Interactive Learning Environments, 29(7), 1011–1022.
26) Bowers, A. J., & Sprott, R. (2012). Why tenth graders fail algebra I: The contributions of academic press, teacher practices, and student engagement. The High School Journal, 95(2), 1–18.