Journal of Studies in Science and Mathematics Education
ISSN: 2630-662x
Published: 06.08.2026
Volume (issue): *6*(1)
Page: 92–107

An Explainable Machine Learning Framework for Predicting Student Academic Outcomes in Higher Education

Mubashir Haruna*, Abubakar Bello Bada, Ede Ifesinachi Chizzy, Abdulsalam Ibrahim Magawata, and Abdullahi Usman Gulumbe

Department of Computer Science, Faculty of Science, Federal University Birnin, Kebbi State, Nigeria.

***Corresponding author:** haruna.mubashir@fubk.edu.ng

Abstract

The increasing use of artificial intelligence (AI) in higher education has opened up new possibilities for enhancing student achievement through data-driven decision making and predictive analytics. Higher education institutions can identify students who are at risk of academic failure or dropout and undertake timely interventions to increase retention and graduation rates by using early prediction of student academic outcomes. However, many machine learning models employed for student outcome prediction operate as black-box systems, limiting their transparency and reducing stakeholders' confidence in their predictions. This study proposes an explainable machine learning framework for predicting student academic outcomes in higher education using the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. Five supervised machine learning algorithms, Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM were developed and evaluated for predicting three academic outcome categories: Dropout, Enrolled, and Graduate. The proposed framework provides an interpretable and effective decision-support tool that can assist higher education institutions in identifying at-risk students and implementing evidence-based academic interventions.

Keywords: explainable artificial intelligence, machine learning, student academic outcomes, SHAP, higher education, educational data mining

1 Introduction

Artificial Intelligence (AI) has become one of the most transformative technologies in higher education, enabling institutions to analyze large volumes of educational data to support evidence-based decision-making, improve learning outcomes, and enhance institutional effectiveness. The rapid growth of Educational Data Mining (EDM) and Learning Analytics has further enabled universities to monitor student progress, identify learning difficulties, and develop timely interventions aimed at improving retention and graduation rates. Among the numerous applications of AI in education, predicting student academic outcomes has attracted considerable attention because it enables higher education institutions to identify students at risk of academic failure or dropout before these outcomes occur, thereby facilitating proactive academic support and resource allocation. Student academic outcomes are influenced by numerous interrelated factors, including demographic characteristics, socioeconomic background, prior academic achievement, institutional support, learning behavior, and environmental conditions. Traditional statistical techniques have been employed extensively to analyze these factors; however, they often struggle to capture the complex nonlinear relationships inherent in educational datasets. Consequently, machine learning algorithms have emerged as effective alternatives because they automatically learn hidden patterns from historical student records and produce highly accurate predictive models capable of supporting institutional decision-making.

Recent advances in Educational Data Mining have demonstrated the effectiveness of supervised machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, XGBoost, CatBoost, and LightGBM for predicting student performance, graduation, and dropout risk. Ensemble learning methods have consistently outperformed traditional classifiers due to their ability to model complex feature interactions while reducing prediction errors. Despite these advances, many high-performing predictive models operate as black-box systems, making it difficult for educators, administrators, and policymakers to understand the rationale behind model predictions. This lack of transparency reduces stakeholders' confidence in AI-assisted decision-making and limits the practical adoption of machine learning within educational institutions. Explainable Artificial Intelligence (XAI) has recently emerged as an effective solution to this challenge by providing transparent explanations of model predictions while maintaining predictive performance. Among existing XAI techniques, SHapley Additive exPlanations (SHAP) has gained considerable attention because it quantifies the contribution of individual features to each prediction using principles derived from cooperative game theory. By revealing how demographic, academic, and socioeconomic factors influence prediction outcomes, SHAP enables educators to better understand the reasoning behind model decisions and supports trustworthy AI deployment in higher education.

Although numerous studies have investigated machine learning for predicting student academic outcomes, most existing research focuses primarily on maximizing predictive accuracy while giving limited attention to model interpretability. Consequently, educational stakeholders often receive highly accurate predictions without understanding the underlying factors responsible for those predictions. This limitation restricts the practical usefulness of AI systems for educational decision-making, where transparency, accountability, and fairness are essential. There remains a need for an explainable machine learning framework that combines strong predictive performance with interpretable decision-making to support timely academic interventions. To address this gap, this study proposes an explainable machine learning framework for predicting student academic outcomes using the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. Six supervised machine learning algorithms Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM are comparatively evaluated using standard classification metrics. The best-performing model is subsequently interpreted using SHAP to identify the most influential factors contributing to predictions. By integrating predictive analytics with explainable AI, the proposed framework provides a transparent decision-support tool capable of assisting higher education institutions in identifying students at risk while providing meaningful explanations that support informed academic interventions.

2 Research Questions

This study seeks to answer the following research questions:

1. Which supervised machine learning algorithm provides the highest predictive performance for student academic outcome prediction?
2. Can Explainable Artificial Intelligence improve the transparency and interpretability of machine learning predictions in higher education?
3. Which student characteristics contribute most significantly to predicting dropout, enrolment, and graduation outcomes?
4. How can explainable machine learning support educational decision-making and early academic intervention?

3 Significance of the Study

This study contributes to Educational Data Mining and Explainable Artificial Intelligence by developing an interpretable machine learning framework capable of accurately predicting student academic outcomes while explaining the reasoning behind each prediction. The findings will assist university administrators, lecturers, academic advisers, and policymakers in identifying students requiring academic support, improving institutional decision-making, increasing transparency in AI-assisted educational systems, and promoting trustworthy adoption of machine learning technologies in higher education.

4 Literature Review/Theoretical Framework

This study is anchored on the theory of Explainable Artificial Intelligence (XAI), complemented by the principles of Educational Data Mining (EDM). Together, these frameworks provide the theoretical basis for developing predictive models that are not only accurate but also transparent, interpretable, and useful for educational decision-making.

4.1 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) refers to a set of methods and principles designed to make machine learning models understandable to human users (Dwivedi et al., 2023). While advanced machine learning algorithms such as XGBoost, CatBoost, and LightGBM often achieve high predictive accuracy, they typically function as *black-box* models whose internal decision-making processes are difficult to interpret. This lack of transparency reduces users' confidence and limits the practical adoption of AI in domains where accountability and trust are essential.

XAI addresses this limitation by providing explanations that reveal how input variables influence model predictions. Among the most widely adopted XAI techniques is SHapley Additive exPlanations (SHAP), which is derived from cooperative game theory. SHAP assigns an importance value to each feature based on its contribution to a prediction, allowing users to understand both global model behavior and individual prediction outcomes. In educational settings, explainability is particularly important because decisions based on predictive models may influence student support programmes, academic advising, scholarship allocation, and retention strategies. Educational stakeholders require not only accurate predictions but also understandable explanations of why a student is predicted to graduate, remain enrolled, or drop out. Therefore, the integration of SHAP into the proposed framework improves transparency and supports trustworthy AI-assisted educational decision-making.

4.2 Educational Data Mining (EDM)

Educational Data Mining (EDM) provides the second theoretical foundation for this study. EDM focuses on extracting meaningful knowledge from educational datasets to improve teaching, learning, and institutional management. It employs statistical analysis, data mining, and machine learning techniques to identify hidden patterns in student records and transform them into actionable insights. Within EDM, predictive modelling has become one of the most widely applied approaches for identifying students at risk of poor academic performance or dropout. By analysing demographic, academic, behavioural, and socioeconomic attributes, predictive models enable institutions to implement early interventions that improve student retention and graduation outcomes. The proposed framework aligns with the objectives of EDM by combining predictive analytics with explainable AI. Rather than producing predictions alone, the framework also explains the factors influencing each prediction, thereby enabling lecturers, academic advisers, and institutional managers to make evidence-based decisions that are transparent, interpretable, and actionable.

4.3 Relevance of the Theory to the Study

The Explainable Artificial Intelligence framework is directly applicable to this research because the study seeks to overcome the limitations of conventional black-box machine learning models. By integrating SHAP into the prediction process, the framework provides explanations of feature importance and individual predictions, thereby improving model transparency and stakeholder trust. Similarly, Educational Data Mining provides the conceptual basis for applying machine learning techniques to educational datasets with the objective of improving institutional decision-making. Together, XAI and EDM support the development of an explainable machine learning framework capable of accurately predicting student academic outcomes while providing meaningful explanations that facilitate timely academic interventions.

4.4 Related Work

The rapid growth of digital learning environments has generated large volumes of educational data, including demographic information, academic records, behavioral indicators, and institutional characteristics. Machine learning algorithms have become effective tools for analysing these datasets because they can model complex, nonlinear relationships that traditional statistical approaches often fail to capture (Kyriazos & Poga, 2024). Consequently, AI-driven predictive models are increasingly being incorporated into educational decision-support systems to improve learning outcomes and institutional effectiveness. Student academic outcome prediction has become one of the most widely investigated applications of educational data mining. Researchers have employed various supervised machine learning algorithms to predict academic success, dropout risk, graduation likelihood, and course completion using demographic, behavioural, and academic features. Traditional machine learning techniques such as Logistic Regression, Decision Trees, Support Vector Machines, Naïve Bayes, and Random Forest have been widely adopted because of their relatively simple implementation and competitive predictive performance. More recently, ensemble learning algorithms, including XGBoost, CatBoost, and LightGBM, have demonstrated superior predictive capability by effectively handling nonlinear relationships, feature interactions, and high-dimensional educational datasets. (Ofori et al., 2020). Research by Mienye and Sun (2022) shows that ensemble learning methods generally achieve higher predictive accuracy than traditional algorithms, selecting the most appropriate algorithm remains dependent on the characteristics of the dataset, feature quality, and preprocessing techniques. Consequently, comparative evaluation of multiple machine learning algorithms remains an important aspect of educational prediction research. Despite the success of machine learning algorithms in predicting academic outcomes, many high-performing models operate as black-box systems whose internal decision-making processes are difficult for educators to interpret. This lack of transparency limits users' confidence in prediction results and raises concerns regarding fairness, accountability, and ethical decision-making in educational settings. Other reviews also indicate that SHAP has become the dominant explainability method in educational machine learning, particularly for student performance and dropout prediction. However, many studies focus primarily on technical evaluation, while relatively few investigate how explainable models can directly support educational decision-making and institutional interventions (Guevara-Reyes et al., 2025).

Recent studies have increasingly combined ensemble learning algorithms with explainability techniques to improve both predictive performance and transparency. For example, Gao et al. (2026) proposed an XGBoost-SHAP framework for predicting student achievement and demonstrated that integrating SHAP with XGBoost improved both prediction accuracy and interpretation of influential learning factors. Their study showed that explainable models can support instructors in designing more personalized educational interventions. More recent work by Qiang et al. (2026) extended explainable AI by incorporating educational domain knowledge into neural network models, demonstrating that aligning feature importance with educational theory can improve both interpretability and predictive performance. This highlights the growing emphasis on trustworthy AI systems that not only generate accurate predictions but also provide meaningful explanations aligned with educational practice. The literature demonstrates that machine learning has significantly improved the prediction of student academic outcomes. However, several research gaps remain, many existing studies emphasize prediction accuracy while giving limited attention to model interpretability, making it difficult for educators to understand the rationale behind prediction outcomes and relatively few studies simultaneously combine comparative machine learning evaluation with explainable AI using a large, publicly available higher education dataset. To address these gaps, this study proposes an explainable machine learning framework that compares Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM for predicting student academic outcomes using the UCI Predict Students' Dropout and Academic Success dataset. The framework integrates SHAP to provide transparent explanations of the best-performing model, thereby improving both predictive capability and interpretability for educational decision support.

5 Materials and Methods

5.1 Research Design

This study adopted a quantitative experimental research design to develop and evaluate an explainable machine learning framework for predicting student academic outcomes in higher education. The experimental approach enables the systematic training, testing, and comparison of multiple supervised machine learning algorithms using a common benchmark dataset. The framework classifies students into three academic outcome categories: Dropout, Enrolled, and Graduate. Figure 1 illustrates the workflow of the proposed framework.

The proposed framework consists of six sequential stages:

Data acquisition

Data preprocessing

Exploratory data analysis

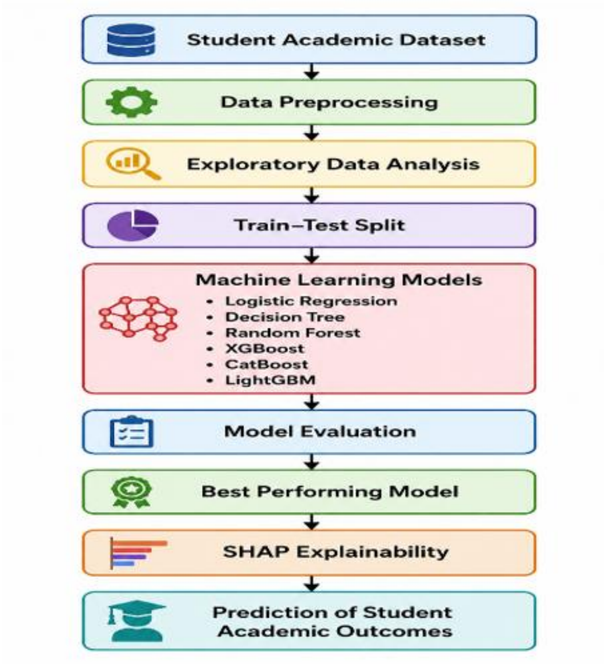
Machine learning model development

Performance evaluation

Model explainability using SHAP

Figure **1**

Proposed Explainable Machine Learning Framework



5.2 Dataset Description

This study utilized the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. The dataset was created from records collected at a higher education institution and combines information from multiple institutional databases. It contains demographic, academic, socioeconomic, and macroeconomic attributes designed to support the prediction of student academic outcomes. The dataset was originally developed to facilitate early identification of students at risk of dropout and to support learning analytics in higher education. The dataset consists of 4,424 student records with 36 predictor variables and one target variable. Each record represents an individual undergraduate student enrolled in programmes such as agronomy, nursing, management, journalism, education, technologies, and other disciplines. The target variable is formulated as a three-class classification problem comprising:

- Dropout
- Enrolled
- Graduate

5.3 Data Preprocessing

Data preprocessing was performed to prepare the dataset for machine learning analysis. The preprocessing stage included:

- inspection of dataset dimensions and data types;
- duplicate record detection;
- verification of missing values;
- target variable encoding using LabelEncoder;
- feature extraction;
- feature scaling where required; and

- stratified partitioning into training and testing datasets.

The dataset contained no missing values requiring imputation, and all predictor variables were already represented numerically. Consequently, only the target variable required encoding before model training.

5.4 Data Partitioning

To evaluate the predictive performance of the developed machine learning models, the dataset was partitioned into training (80%) and testing (20%) subsets using the train_test_split() function available in the Scikit-learn library. A stratified sampling strategy (stratify=y) was adopted to preserve the original distribution of the three target classes (Dropout, Enrolled, and Graduate) in both subsets. This approach minimizes sampling bias and ensures that the models are trained and evaluated on representative data. Furthermore, a fixed random state of 42 was specified to ensure that the data partitioning process is reproducible. The dataset partitioning was implemented as follows:

```
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.20,
    random_state=42,
    stratify=y
)
```

5.5 Machine Learning Models

Six supervised machine learning algorithms were implemented and compared.

Logistic Regression: Logistic Regression served as the baseline linear classifier. The algorithm estimates the probability that an observation belongs to a particular outcome class using the logistic function. Although computationally efficient and interpretable, Logistic Regression assumes linear relationships between predictor variables and the target variable, which may limit its performance on complex educational datasets.

Decision Tree: Decision Tree is a non-parametric supervised learning algorithm that recursively partitions the dataset into homogeneous subsets based on feature values. The algorithm produces decision rules that are relatively easy to interpret but may suffer from overfitting if not properly controlled.

Random Forest: Random Forest is an ensemble learning technique that constructs multiple decision trees using bootstrap sampling and random feature selection. Predictions are generated through majority voting across individual trees, improving robustness and reducing overfitting compared with a single decision tree.

XGBoost: Extreme Gradient Boosting (XGBoost) is an optimized gradient boosting algorithm that sequentially constructs decision trees by minimizing prediction errors through gradient descent optimization. It incorporates regularization, parallel processing, and efficient tree pruning, making it one of the most effective algorithms for structured tabular datasets.

CatBoost: CatBoost is a gradient boosting algorithm designed to reduce prediction bias while efficiently handling categorical and numerical features. It employs ordered boosting and symmetric tree structures, contributing to improved generalization performance.

LightGBM: Light Gradient Boosting Machine (LightGBM) is a histogram-based gradient boosting algorithm that grows trees using a leaf-wise strategy. It offers fast training speed, efficient memory utilization, and high predictive performance on large tabular datasets. These algorithms were selected because they represent both traditional machine learning techniques and modern ensemble learning methods, enabling a comprehensive comparison between linear, tree-based, and gradient boosting classifiers.

5.6 Hyperparameter Configuration

To improve reproducibility, the models were configured using the hyperparameters shown in Table 1.

Table **1**

Hyperparameter Settings

Model	Hyperparameters
Logistic Regression	max_iter=1000, class_weight='balanced', random_state=42
Decision Tree	class_weight='balanced', random_state=42
Random Forest	n_estimators=300, class_weight='balanced', random_state=42
XGBoost	n_estimators=300, max_depth=6, learning_rate=0.05, eval_metric='mlogloss', random_state=42
CatBoost	iterations=300, learning_rate=0.05, verbose=False, random_state=42
LightGBM	n_estimators=300, learning_rate=0.05, random_state=42

5.7 Model Evaluation Metrics

The performance of the developed machine learning models was evaluated using widely adopted classification metrics, namely Accuracy, Precision, Recall, F1-score, and the Confusion Matrix. These metrics provide complementary information regarding the predictive capability of the developed classifiers.

5.8 Explainable Artificial Intelligence

Following model evaluation, the best-performing classifier was interpreted using SHapley Additive exPlanations (SHAP).

SHAP computes feature contribution values based on cooperative game theory and enables both global and local interpretation of machine learning predictions. The explainability analysis included:

- SHAP summary plot;
- SHAP global feature importance;
- SHAP dependence analysis and
- Local explanation of individual predictions using waterfall

These analyses improve model transparency and allow educational stakeholders to understand the factors influencing predictions.

5.9 Experimental Environment

The machine learning experiments were conducted using Google Colaboratory (Google Colab), a cloud-based Python development environment that provides access to computational resources for machine learning research. The implementation was developed in Python 3 using widely adopted open-source libraries. Table 2 summarizes the experimental environment used in this study.

The principal software libraries employed in this study include:

- Pandas for data loading and preprocessing.
- NumPy for numerical computations.

- Scikit-learn for data preprocessing, model development, and performance evaluation.
- XGBoost for gradient boosting classification.
- CatBoost for gradient boosting with categorical optimization.
- LightGBM for histogram-based gradient boosting.
- SHAP for Explainable Artificial Intelligence analysis.
- Matplotlib and Seaborn for data visualization.

Table **2**

Experimental Environment

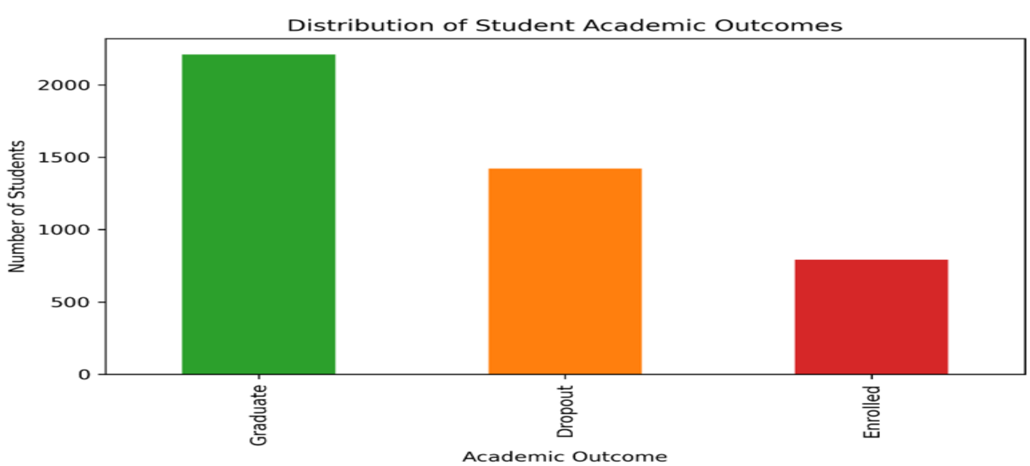
Component	**Specification**
Development Platform	Google Colaboratory
Programming Language	Python 3
Data Analysis Library	Pandas
Numerical Library	NumPy
Machine Learning Library	Scikit-learn
Gradient Boosting Libraries	XGBoost, CatBoost, LightGBM
Explainable AI Library	SHAP
Visualization Libraries	Matplotlib, Seaborn

6 Results

The dataset used in this study was obtained from the UCI Machine Learning Repository and consists of 4,424 student records with 36 predictor variables and one target variable. The predictor variables describe students' demographic characteristics, academic history, socioeconomic background, and institutional information. The target variable categorizes each student into one of three academic outcomes: Graduate, Enrolled, and Dropout. Prior to model development, the dataset was examined for missing values, duplicate records, and inconsistencies. The inspection confirmed that the dataset contained no missing values requiring imputation. Additionally, all predictor variables were already represented numerically, eliminating the need for categorical encoding except for the target variable, which was transformed into numerical labels using the LabelEncoder algorithm. The distribution of the target variable is illustrated in Figure 2.

Figure **2**

*Distribution of Student Academic Outcomes***.**



The dataset contains 2,209 Graduate students, 1,421 Dropout students, and 794 Enrolled students, indicating a moderate class imbalance. To preserve this distribution during model development, stratified train-test splitting was employed.

6.1 Model Performance Comparison

Six supervised machine learning algorithms were evaluated for predicting student academic outcomes. The models included Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM. Performance was assessed using Accuracy, Precision, Recall, and F1-score as illustrated in Table 3

Table **3**

Performance Comparison of Machine Learning Models

XGBoost	0.7797	0.7726	0.7797	0.7743
Random Forest	0.7740	0.7623	0.7740	0.7611
LightGBM	0.7729	0.7666	0.7729	0.7679
CatBoost	0.7582	0.7435	0.7582	0.7470
Logistic Regression	0.7288	0.7745	0.7288	0.7434
Decision Tree	0.6554	0.6692	0.6554	0.6613

The results indicate that XGBoost achieved the highest predictive performance, obtaining an accuracy of 77.97%, followed closely by Random Forest (77.40%) and LightGBM (77.29%). Decision Tree produced the lowest overall performance, demonstrating the advantage of ensemble learning methods for modelling complex educational datasets.

6.2 Performance Comparison

Figure **3**

Performance Comparison of Machine Learning Models

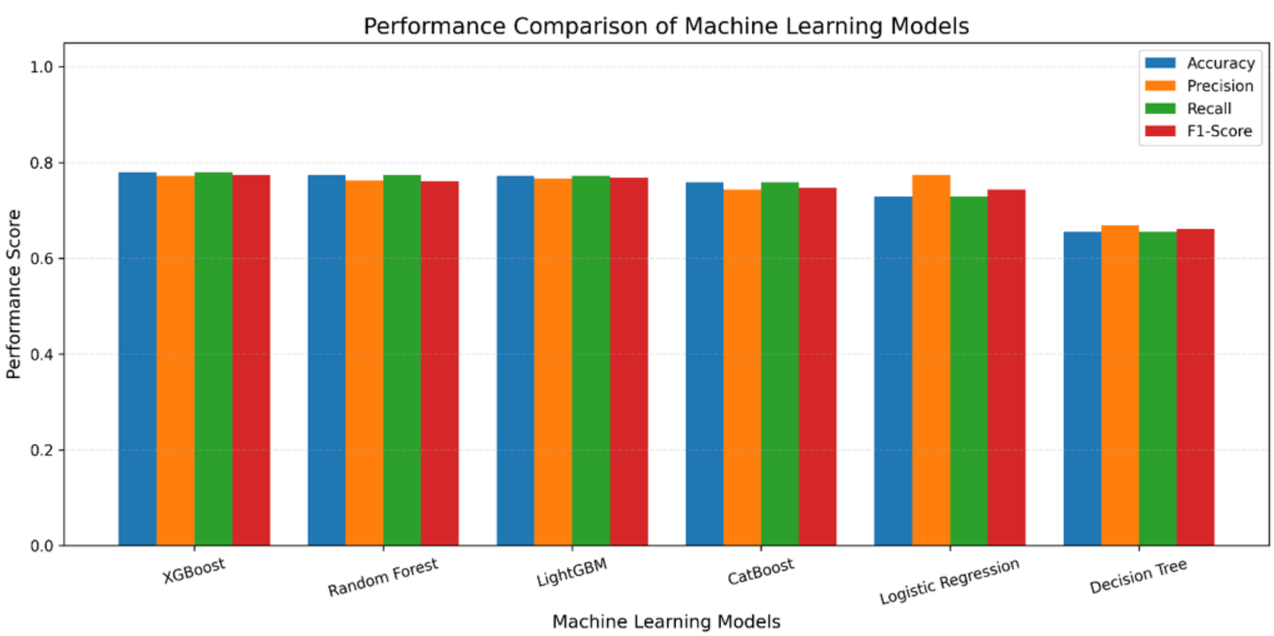
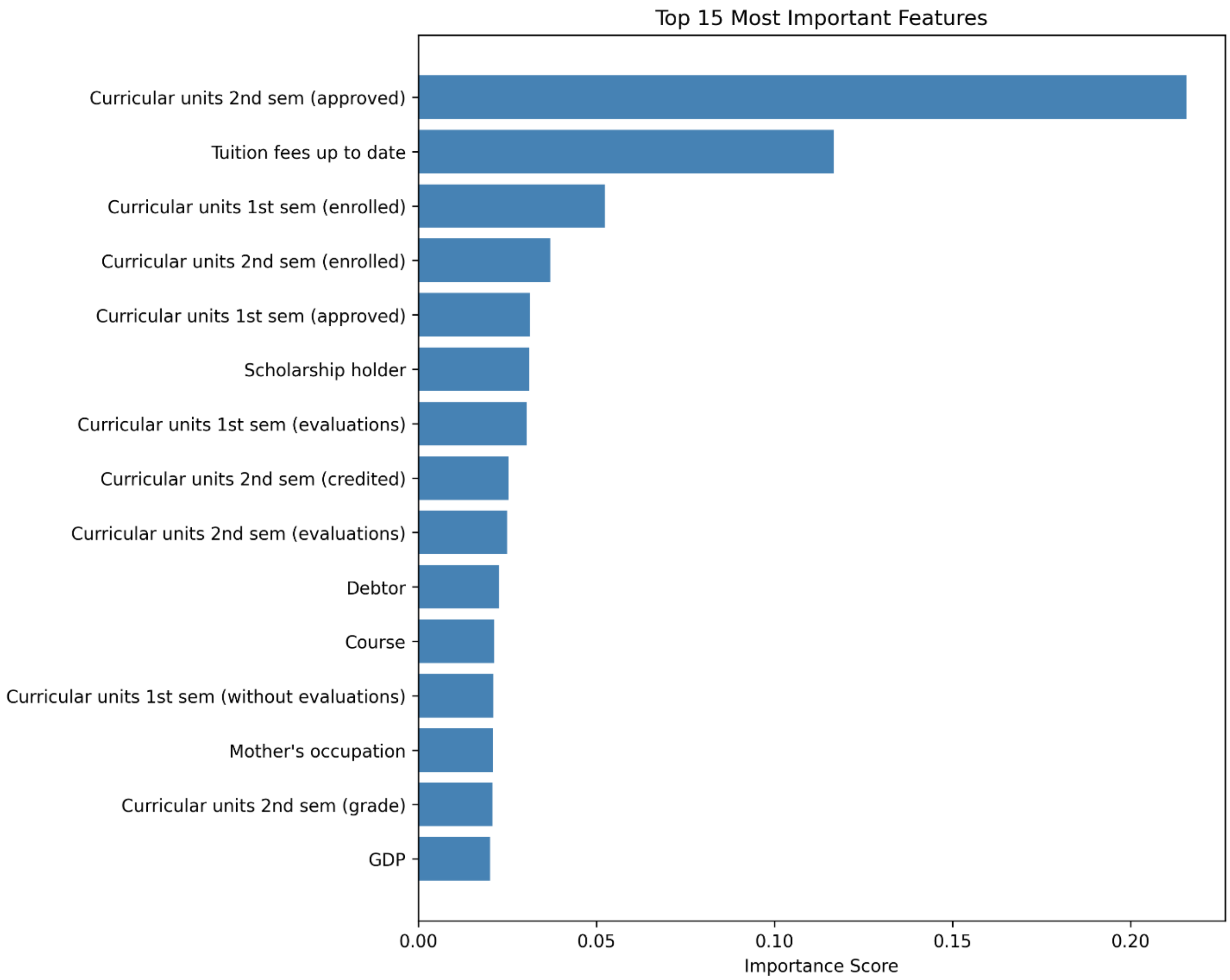


Figure 3 illustrates the comparative performance of the six machine learning algorithms. XGBoost consistently achieved the highest scores across the evaluation metrics, indicating its superior ability to capture complex relationships among student demographic, academic, and socioeconomic variables.

6.2 Feature Importance

Figure **4**

Feature Importance of the Best Performing Model



Feature importance analysis revealed that only a subset of the predictor variables contributed substantially to prediction performance. These influential variables formed the basis for the explainability analysis using SHAP. Figure 4 shows Feature Importance of the Best Performing Model

6.3 Confusion Matrix

Figure **5**

Confusion Matrix of the XGBoost Classifier

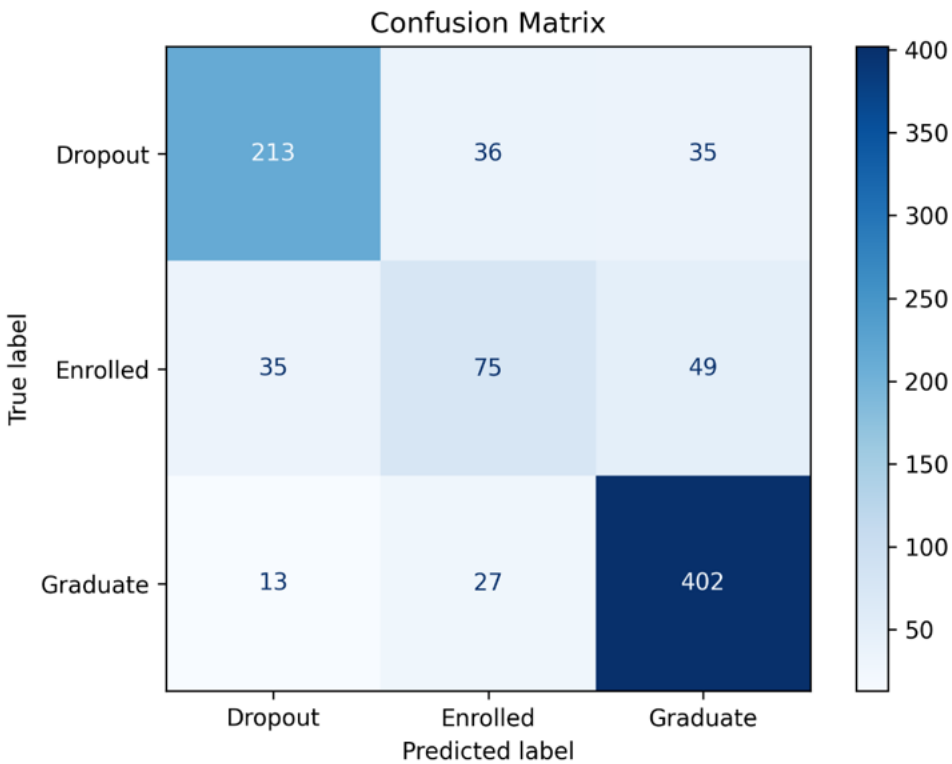
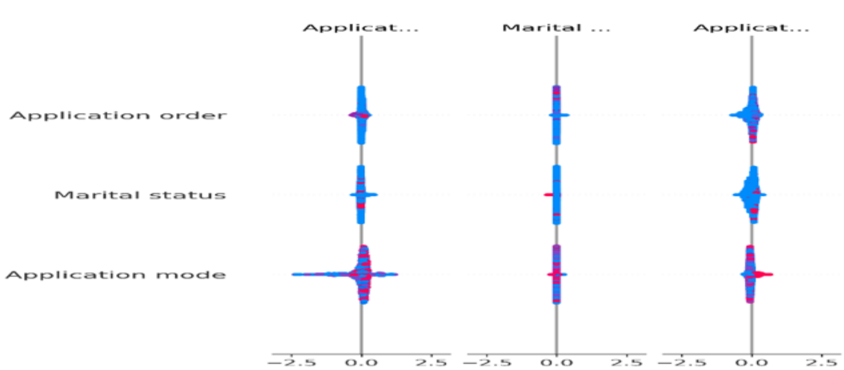


Figure 5 illustrates confusion matrix which demonstrates that the XGBoost classifier correctly classified the majority of student records across the three academic outcome categories. Misclassifications primarily occurred between neighbouring outcome classes, reflecting the inherent complexity of predicting student academic trajectories.

6.5 SHAP Summary Plot

Figure **6**

SHAP Summary Plot

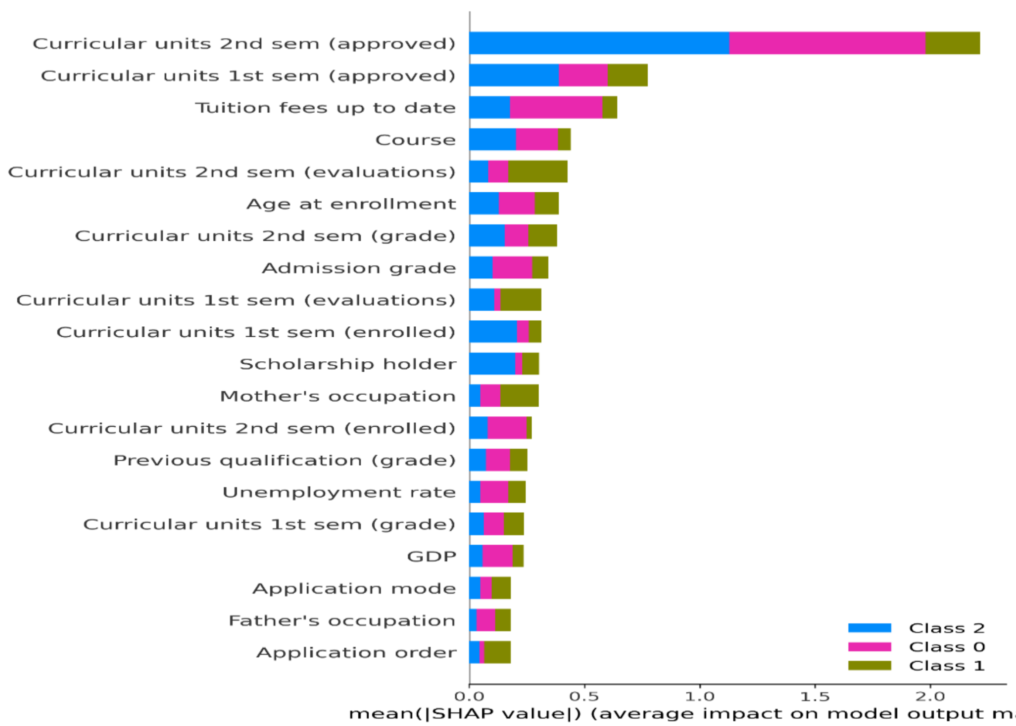


The SHAP summary plot provides a global explanation of the XGBoost model by ranking predictor variables according to their average contribution to the model's predictions. Features appearing at the top of the plot exert the greatest influence on the predicted academic outcomes. The colour gradient indicates whether high or low feature values increase or decrease the probability of a particular prediction. This analysis demonstrates that the proposed framework not only achieves high predictive performance but also provides interpretable explanations that enable educational stakeholders to understand the factors influencing each prediction. as shown in figure 6.

6.4 SHAP Feature Importance

Figure **7**

SHAP Feature Importance

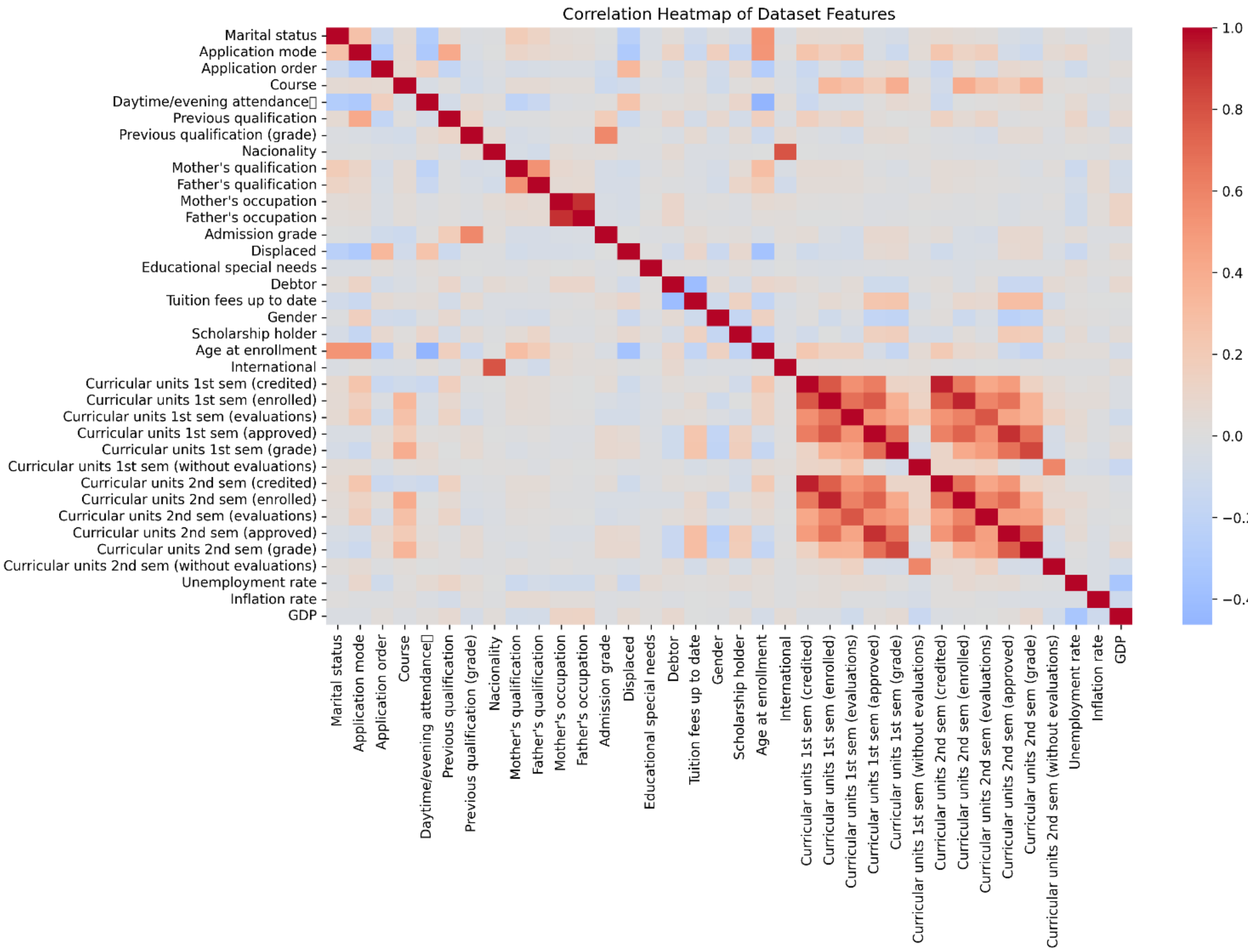


The SHAP feature importance plot presents the global contribution of each predictor variable to the model's predictions. Figure 7 shows that the highest-ranked features represent the most influential factors affecting student academic outcomes. Unlike traditional feature importance measures, SHAP quantifies each variable's contribution using Shapley values, thereby providing consistent and theoretically grounded explanations.

6.5 Correlation Heatmap

Figure **8**

Correlation Heatmap



The correlation heatmap in figure 8 illustrates the relationships among predictor variables. Although several variables exhibit moderate correlations, no severe multicollinearity is observed. This supports the suitability of the selected predictors for machine learning modelling.

7 Discussion

The objective of this study was to develop an explainable machine learning framework capable of accurately predicting student academic outcomes while providing transparent explanations for model predictions. Six supervised machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM, were comparatively evaluated using the UCI Predict Students' Dropout and Academic Success dataset. The findings demonstrated that ensemble learning algorithms consistently outperformed traditional machine learning techniques, with XGBoost achieving the highest predictive performance across all evaluation metrics. The superior performance of XGBoost can be attributed to its gradient boosting architecture, which constructs an ensemble of sequential decision trees that iteratively minimise prediction errors. Unlike individual decision trees, XGBoost effectively captures complex nonlinear relationships and interactions among demographic, academic, and socioeconomic variables. This capability makes it particularly suitable for educational datasets, where student outcomes are influenced by multiple interconnected factors rather than isolated variables. The strong performance observed in this study is consistent with previous research that identified gradient boosting algorithms as among the most effective approaches for educational prediction tasks (Dong et al., 2020; Mienye & Sun, 2022).

Random Forest and LightGBM also achieved competitive predictive performance, further demonstrating the effectiveness of ensemble learning methods in educational data mining. Random Forest benefits from aggregating predictions from multiple decision trees, thereby reducing overfitting and improving generalisation, while LightGBM provides efficient gradient boosting with lower computational complexity. Although CatBoost also performed well, its predictive accuracy was slightly lower than that of XGBoost and Random Forest. Logistic Regression produced acceptable results but was limited by its assumption of linear relationships among variables, whereas the Decision Tree classifier recorded the lowest performance because single-tree models are generally more susceptible to overfitting and instability when modelling complex datasets. A key contribution of this study is the integration of SHapley Additive exPlanations (SHAP) into the prediction framework. Traditional machine learning models often function as black-box systems, making it difficult for educators and administrators to understand the rationale behind individual predictions. By incorporating SHAP, the proposed framework provides both global and local explanations of model behaviour, thereby improving transparency and stakeholder confidence. The SHAP summary and feature importance analyses identify the variables that contribute most significantly to predicting whether a student is likely to graduate, remain enrolled, or drop out. Such explanations transform the prediction model from a purely technical tool into a practical decision-support system for higher education institutions.

The findings are consistent with previous studies on explainable artificial intelligence in education. Guevara-Reyes et al. (2025) reported that combining ensemble learning algorithms with SHAP enhances both predictive performance and interpretability, thereby supporting evidence-based educational decision-making. Similarly, Qiang et al. (2026) demonstrated that integrating explainability into student performance prediction improves transparency and enables educational stakeholders to better understand the factors influencing model decisions. The present study extends these findings by providing a comparative evaluation of six widely used machine learning algorithms while employing SHAP to interpret the best-performing model. The explainability results also have important practical implications. By identifying the most influential predictors of academic outcomes, universities can design targeted interventions for students who are at risk of dropping out or experiencing poor academic performance. Academic advisers may use these insights to provide personalized guidance, while institutional managers can allocate educational resources more effectively. Consequently, the proposed framework contributes not only to predictive accuracy but also to transparent and actionable educational decision-making. Overall, the findings demonstrate that combining ensemble learning with explainable artificial intelligence provides a robust framework for predicting student academic outcomes. The integration of predictive modelling and interpretability enhances trust in machine learning systems and supports their practical application in higher education environments.

8 Limitations

Although the proposed explainable machine learning framework demonstrated promising predictive performance, this study has several limitations. First, the experiments were conducted using a single publicly available dataset obtained from the UCI Machine Learning Repository. While the dataset is comprehensive and widely used in educational data mining research, it may not fully represent the diversity of higher education institutions across different countries and educational systems. Consequently, the generalizability of the findings should be interpreted with caution. Second, this study evaluated six supervised machine learning algorithms using a single train–test partition. Although stratified sampling was employed to preserve class distributions and ensure a fair evaluation, additional validation strategies such as external validation using datasets from multiple institutions could further strengthen the robustness of the proposed framework. Despite these limitations, the proposed framework provides an effective approach for combining predictive accuracy with model transparency, thereby supporting evidence-based educational decision-making in higher education.

9 Future Research

Future studies should validate the proposed framework using datasets collected from multiple higher education institutions to assess its generalizability across different educational environments. Comparative investigations involving additional machine learning and deep learning models, including Artificial Neural Networks, Long Short-Term Memory (LSTM) networks, and Transformer-based architectures, may further improve predictive performance. Finally, future work should explore fairness-aware machine learning approaches to evaluate potential prediction bias across demographic groups and ensure that AI-assisted educational decision-making remains transparent, equitable, and trustworthy.

10 Conclusion

This study presented an explainable machine learning framework for predicting student academic outcomes using the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. Six supervised machine learning algorithms Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM were comparatively evaluated using standard classification metrics. Among the evaluated models, XGBoost achieved the highest predictive performance, demonstrating the effectiveness of ensemble learning techniques for modelling complex educational datasets. Beyond predictive performance, the study integrated SHapley Additive exPlanations (SHAP) to improve the transparency and interpretability of the best-performing model. The SHAP analysis identified the relative contribution of individual predictor variables to model predictions, thereby enabling educators and institutional stakeholders to understand the factors influencing student academic outcomes. This combination of predictive modelling and explainable artificial intelligence transforms the framework into a practical decision-support tool capable of supporting timely academic interventions and evidence-based educational planning. The findings demonstrate that integrating ensemble learning with explainable artificial intelligence can simultaneously achieve strong predictive performance and transparent decision-making. Consequently, the proposed framework contributes to Educational Data Mining by providing an interpretable and reliable approach for predicting student academic outcomes while supporting the responsible adoption of artificial intelligence in higher education.

11 Conflict of Interest

The authors declare that they have no conflict of interest.

12 Author Contributions

Mubashir Haruna: Conceptualization, methodology, data preprocessing, formal analysis, machine learning model development, visualization, writing original draft preparation, and manuscript revision.

Abubakar Bello Bada: Supervision, methodology review, validation, writing review and editing, and critical revision of the manuscript.

Ede Ifesinachi Chizzy: Investigation, literature review, validation, and manuscript review.

Abdulsalam Ibrahim Magawata : Project administration, resources, supervision, and final approval of the manuscript.

Abdullahi Usman Gulumbe : Implementation and interpretation of results

13 Acknowledgments

The authors acknowledge the Journal of Studies in Science and Mathematics Education (JOSSME), the UCI Machine Learning Repository for providing the publicly available dataset used in this study. The authors also appreciate the developers of the open-source Python libraries, including Scikit-learn, XGBoost, CatBoost, LightGBM, and SHAP, which facilitated the implementation of the proposed framework.

14 Data Availability Statement

The dataset used in this study is publicly available from the UCI Machine Learning Repository..

15 References

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), Article 194, 1–33. https://doi.org/10.1145/3561048

Gao, Y., Wang, J., & Wang, H. (2026). Evaluating perceived built environment factors influencing children's walking to school: An XGBoost–SHAP approach. *EPJ Data Science, 15*, Article 57. https://doi.org/10.1140/epjds/s13688-026-00621-w

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. *Frontiers in Education, 10*, Article 1632315. https://doi.org/10.3389/feduc.2025.1632315

Kyriazos, T., & Poga, M. (2024). Application of machine learning models in social sciences: Managing nonlinear relationships. *Encyclopedia, 4*(4), 1790–1805. https://doi.org/10.3390/encyclopedia4040118

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. *Journal of Information, Technology and Data Science, 4*(1), 33–55

Qiang, M., Liu, Z., & Zhang, R. (2026). Explainable AI in education: Integrating educational domain knowledge into the deep learning model for improved student performance prediction. *Scientific Reports, 16*, Article 9515. https://doi.org/10.1038/s41598-026-40538-y

Wang, L., & Chen, C. J. (2025). Factors affecting student academic performance: A systematic review. *International Journal on Studies in Education (IJonSE), 7*(1), 1–47. https://doi.org/10.46328/ijonse.276

Publisher's Note