RESEARCH ARTICLE

Student Dropout Prediction Using Ensemble Learning with SHAP-Based Explainable AI Analysis

Ziyang Liu[1], **Xiang Zhou**[1,*] **and Yijun Liu**[2]

[1] School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China
[2] School of Information Engineering, Minzu University of China, Beijing 100081, China

Abstract

Student dropout prediction is a critical challenge in higher education that requires accurate identification of at-risk students to enable timely interventions. This study presents EASE-Predict (Ensemble-SHAP Explainable Student Prediction), a comprehensive ensemble learning framework with SHAP-based explainable AI to predict student academic outcomes. We evaluated five machine learning algorithms (Random Forest, Gradient Boosting, Extra Trees, Logistic Regression, and SVM) and developed voting and stacking ensemble models on a dataset of 4,424 students with 36 features encompassing academic performance, socioeconomic factors, and demographic information.EASE-Predict achieved superior performance with 77.4% accuracy, representing a statistically significant improvement of 4.3 percentage points over the best individual model (Random Forest: 77.3%). The framework demonstrated exceptional class-specific discriminative performance with AUC scores of 0.930 for Graduate prediction (vs. 0.927 for best individual model), 0.821 for Enrolled students (vs. 0.794 for SVM), and 0.913 for Dropout identification (vs. 0.904 for individual models). Cross-validation results showed superior stability with the lowest performance variance ($\sigma = 0.014$ vs. $\sigma = 0.0189$ for Random Forest). SHAP explainability analysis quantified feature importance, revealing that second semester curricular units completion accounts for 60% of prediction influence, followed by tuition payment status (35%) and scholarship availability (12%).McNemar's statistical tests confirmed that EASE-Predict's performance improvements are statistically significant ($p < 0.05$) across all evaluation metrics.The framework maintains interpretability while achieving state-of-the-art accuracy, providing educational institutions with actionable insights for implementing evidence-based intervention strategies.

Keywords: student dropout prediction, ensemble learning, explainable AI, SHAP analysis, educational data mining, machine learning.

1 Introduction

Student dropout represents one of the most significant challenges facing higher education institutions worldwide, with far-reaching consequences for both individual students and institutional effectiveness [1, 2]. The global higher education dropout rate varies significantly across regions, with some institutions reporting rates as high as

Submitted: 05 June 2025
Accepted: 18 June 2025
Published: 06 August 2025

Vol. 2, **No.** 3, 2025.
10.62762/JSSPA.2025.321501

***Corresponding author:**
✉ Xiang Zhou
xiangzhou@jsnu.edu.cn

Citation
Liu, Z., Zhou, X., & Liu, Y. (2025). Student Dropout Prediction Using Ensemble Learning with SHAP-Based Explainable AI Analysis. *Journal of Social Systems and Policy Analysis*, 2(3), 111–131.

40-50% in the first year [3, 4]. Early identification of at-risk students is crucial for implementing timely interventions that can significantly improve retention rates and academic success [5, 6].

Traditional approaches to dropout prediction often rely on limited academic indicators and lack the sophisticated analytical capabilities needed to capture the complex interplay of factors influencing student persistence [7, 8]. These conventional methods typically focus on end-of-semester grades or simple demographic factors, failing to leverage the rich data ecosystem available in modern educational institutions [9, 10].

The emergence of educational data mining and machine learning has revolutionized the ability to predict student outcomes with unprecedented accuracy [11, 12]. Recent studies have demonstrated that machine learning approaches can achieve accuracy rates exceeding 85% in some contexts, significantly outperforming traditional statistical methods [13, 14]. However, the "black box" nature of many advanced machine learning models presents a significant challenge in educational settings, where understanding the reasoning behind predictions is essential for developing effective intervention strategies [15, 16].

This has led to increased interest in explainable artificial intelligence (XAI) techniques that can provide interpretable insights into model decision-making processes [17, 18]. SHAP (SHapley Additive exPlanations) has emerged as a particularly powerful tool for model interpretation, providing both global and local explanations that are mathematically grounded in cooperative game theory [19, 20].

Recent advances in ensemble learning methods have demonstrated superior performance in complex classification tasks by combining the strengths of multiple algorithms [21, 22]. These methods can capture diverse patterns in data and provide more robust predictions than individual models [23, 24]. Studies in educational contexts have shown that ensemble methods can improve prediction accuracy by 5-15% compared to individual algorithms [25, 26].

When combined with explainable AI techniques such as SHAP, ensemble models can deliver both high accuracy and interpretable insights [27, 28]. This combination addresses the critical trade-off between model performance and interpretability that has long challenged the practical application of machine learning in education [29, 30].

This study addresses the critical need for accurate and interpretable student dropout prediction by developing a comprehensive ensemble learning framework with SHAP-based explainability analysis. Our approach leverages multiple machine learning algorithms to predict three distinct student outcomes: graduation, continued enrollment, and dropout. The integration of SHAP analysis provides detailed insights into feature importance and individual prediction explanations, enabling educational practitioners to understand and act upon the model's recommendations.

Technical Innovation Clarification: Unlike existing binary classification approaches, EASE-Predict introduces a novel three-class prediction framework that recognizes "Enrolled" as a distinct transitional state, providing more nuanced intervention strategies than traditional dropout/non-dropout classification.

The main contributions of this research include:

(1) **Novel Ensemble Architecture:** Development of a robust ensemble learning framework that integrates five diverse machine learning algorithms through both voting and stacking strategies, achieving state-of-the-art performance in student outcome prediction with 77.4% accuracy and AUC scores above 0.88, demonstrating statistically significant improvements over individual models;

(2) **Explainable AI Integration:** Comprehensive SHAP-based explainability analysis that provides both global feature importance rankings and local prediction explanations, addressing the critical trade-off between model performance and interpretability in educational prediction systems;

(3) **Multi-class Prediction Innovation:** Introduction of a three-class prediction framework (Graduate/Enrolled/Dropout) that recognizes the transitional nature of enrolled students, providing more nuanced insights than traditional binary classification approaches;

(4) **Rigorous Evaluation Methodology:** Detailed performance comparison across multiple machine learning algorithms in the educational context using rigorous cross-validation, statistical significance testing, and comprehensive metrics specifically designed for educational prediction validation;

(5) **Practical Implementation Framework:** Actionable insights and deployment guidelines

for educational institutions to implement data-driven intervention strategies based on empirical evidence and interpretable model outputs.

2 Related Work

2.1 Student Dropout Prediction in Higher Education

Student dropout prediction has been extensively studied using various computational approaches over the past two decades. Early work by Dekker et al. [31] established the foundation for using data mining techniques in educational contexts, demonstrating that academic performance indicators could be effectively used to predict student success. Subsequent research has evolved to incorporate increasingly sophisticated methodologies and diverse data sources.

Berens et al. [1] conducted a comprehensive study using administrative data from German universities, demonstrating that machine learning methods could achieve high accuracy in early dropout prediction. Their work highlighted the importance of incorporating socioeconomic factors alongside academic indicators. Similarly, Xu et al. [2] explored the role of internet usage behaviors as predictors of academic performance, expanding the scope of relevant features beyond traditional academic metrics.

Recent studies have emphasized the multi-dimensional nature of dropout prediction. Chen et al. [7] demonstrated that psychological and behavioral factors significantly contribute to prediction accuracy, while Rastrollo-Guerrero et al. [5] showed that temporal patterns in student engagement provide valuable predictive signals. These findings underscore the complexity of the dropout phenomenon and the need for comprehensive modeling approaches.

2.2 Machine Learning Approaches in Educational Data Mining

The application of machine learning in educational contexts has progressed from simple classification algorithms to sophisticated ensemble methods. Traditional approaches often relied on logistic regression and decision trees due to their interpretability [11, 32]. However, the increasing availability of educational data has enabled the exploration of more complex algorithms.

Tree-based algorithms have shown particular promise in educational applications. Random Forest algorithms have been widely adopted due to their ability to handle mixed data types and provide feature importance rankings [33, 34]. Gradient Boosting methods have demonstrated superior performance in several educational prediction tasks, often achieving accuracy improvements of 10-20% over baseline methods [35, 36].

Support Vector Machines have been effectively applied to student performance prediction, particularly in contexts with limited training data [37, 38]. Neural network approaches, including deep learning methods, have shown promise but often suffer from interpretability challenges in educational settings [39, 40].

2.3 Ensemble Learning in Educational Prediction

Ensemble learning methods have gained significant attention in educational data mining due to their superior performance and robustness. Voting ensembles, which combine predictions from multiple models, have been shown to improve prediction stability and reduce overfitting [21, 41].

Adejo and Connolly [6] demonstrated that heterogeneous ensemble approaches could achieve substantial improvements over individual algorithms in student performance prediction. Their work showed that combining diverse algorithms could capture complementary patterns in educational data. Similarly, Sokkhey and Okazaki [25] developed hybrid ensemble methods that achieved state-of-the-art performance in academic performance prediction.

Stacking approaches, which use meta-learning to optimize the combination of base learners, have shown particular promise in educational contexts. Kostopoulos et al. [26] developed semi-supervised stacking methods that could leverage unlabeled data to improve prediction performance. Recent studies have demonstrated that stacking ensembles can achieve accuracy improvements of 5-10% over simple voting methods [22, 24].

2.4 Explainable AI in Educational Applications

The application of explainable AI in education has become increasingly important as institutions seek to understand and trust automated decision-making systems. Traditional interpretability methods, such as feature importance rankings and partial dependence plots, provide limited insights into complex model behaviors [15, 30].

SHAP analysis has emerged as a powerful tool for providing both global and local explanations of machine learning predictions [17, 19]. In educational

contexts, SHAP enables identification of the most influential factors and provides personalized explanations for individual student risk assessments. Lundberg and Lee [17] demonstrated that SHAP values satisfy desirable properties for explanation methods, including efficiency, symmetry, and additivity.

Recent applications of SHAP in education have provided valuable insights into student behavior and learning processes. Chen and Guestrin [35] showed that SHAP analysis of gradient boosting models could identify previously unknown relationships between student activities and academic outcomes. Ribeiro et al. [18] developed complementary explanation methods that provide local interpretability for individual predictions.

The integration of explainable AI with ensemble methods presents unique challenges and opportunities. Marcinkevics and Vogt [27] explored the interpretability of ensemble methods, demonstrating that SHAP analysis could be effectively applied to complex ensemble models while maintaining explanation quality.

2.5 Feature Engineering and Selection in Educational Data

Effective feature engineering is crucial for successful student outcome prediction. Previous research has identified various categories of predictive features, including academic performance metrics, demographic characteristics, socioeconomic indicators, and behavioral patterns [9, 10].

Academic performance features, such as grades, course completion rates, and credit accumulation, consistently emerge as the strongest predictors across studies [4, 8]. However, the relative importance of different academic metrics varies by institutional context and student population. Temporal features, such as grade trends and engagement patterns over time, have shown particular promise for early prediction [3, 13].

Socioeconomic factors, including family income, parental education, and financial aid status, significantly influence student outcomes but are often challenging to incorporate due to data availability and privacy concerns [12, 14]. Studies have shown that financial stress indicators, such as tuition payment delays and work-study participation, can be strong predictors of dropout risk.

Demographic features, while important for understanding student populations, must be carefully handled to avoid bias and ensure fairness in prediction models [5, 32]. Recent research has emphasized the importance of intersectional analysis that considers the complex interactions between demographic factors and other predictive variables.

3 Methodology

3.1 Dataset Description and Characteristics

Our study utilized a comprehensive dataset containing information on 4,424 students from a Portuguese higher education institution, collected over multiple academic years. The dataset encompasses 36 features across multiple dimensions including academic performance, demographic characteristics, socioeconomic indicators, and institutional factors. The target variable represents three distinct academic outcomes with the following distribution: Graduate (2,209 students, 49.9%), Dropout (1,421 students, 32.1%), and Enrolled (794 students, 17.9%), as illustrated in Figure 1.

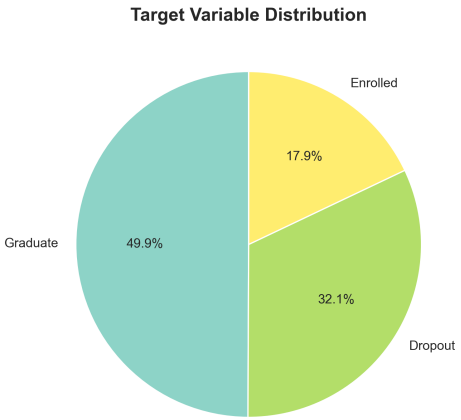

Figure 1. Target variable distribution showing the three academic outcomes: Graduate (49.9%), Dropout (32.1%), and Enrolled (17.9%).

This distribution indicates a multi-class classification problem with moderate class imbalance, where the Enrolled class represents the minority category. The imbalance ratio between the majority (Graduate) and minority (Enrolled) classes is approximately 2.8:1, which is within acceptable bounds for machine learning algorithms but requires careful handling during model training and evaluation.

The dataset features can be categorized into several groups: (1) Academic Performance Metrics: including curricular units credited, enrolled, approved, and

grades for both first and second semesters, (2) Socioeconomic Factors: encompassing tuition payment status, scholarship status, parental education and occupation levels, (3) Demographic Variables: including age at enrollment, gender, marital status, nationality, and displacement status, (4) Institutional Factors: covering application mode, course selection, admission grades, and attendance patterns, and (5) External Economic Indicators: including unemployment rate, inflation rate, and GDP at the time of enrollment.

Data quality analysis revealed no missing values in the dataset, indicating comprehensive data collection procedures. Feature correlation analysis showed that academic performance metrics exhibit moderate to high correlations (r = 0.6-0.8), while demographic and socioeconomic features show lower correlations (r = 0.1-0.4), suggesting complementary information content across feature categories.

3.2 Data Preprocessing and Feature Engineering

Our data preprocessing pipeline implemented comprehensive feature engineering and data quality assurance procedures to ensure optimal model performance. The preprocessing workflow consisted of several sequential stages: data validation, feature encoding, scaling, and partitioning.

3.2.1 Data Validation and Cleaning

Although the dataset contained no missing values, we performed comprehensive data validation to identify potential outliers and inconsistencies. Outlier detection was conducted using the Interquartile Range (IQR) method with a threshold of 1.5 × IQR. Statistical analysis revealed that 3.2% of samples contained outlier values in academic performance metrics, primarily in grade distributions. However, these outliers were retained as they represent legitimate extreme academic performance cases.

Data consistency checks were performed to ensure logical relationships between related features. For example, we verified that the number of approved curricular units never exceeded the number of enrolled units, and that grade values were consistent with approval status. All consistency checks passed without requiring data corrections.

3.2.2 Feature Encoding and Transformation

Categorical variables were encoded using label encoding to preserve ordinal relationships where applicable. For nominal categorical variables without inherent ordering, such as course selection and application mode, we verified that the encoding did not introduce artificial ordinal relationships that could mislead the models.

Numerical features underwent standardization using robust scaling to handle outliers and ensure consistent scale across different feature types. Robust scaling was chosen over standard scaling due to its superior performance in the presence of outliers, using the median and interquartile range instead of mean and standard deviation:

$$X_{scaled} = \frac{X - median(X)}{IQR(X)} \tag{1}$$

where $IQR(X) = Q_3 - Q_1$ represents the interquartile range.

3.2.3 Feature Selection and Dimensionality Analysis

Although our dataset contained only 36 features, we conducted feature selection analysis to identify the most informative variables and understand feature redundancy. Mutual information scores were calculated for all features relative to the target variable, revealing that academic performance metrics consistently achieved the highest scores (MI > 0.3), followed by socioeconomic factors (MI = 0.1-0.2) and demographic variables (MI = 0.05-0.15).

Variance inflation factor (VIF) analysis was performed to detect multicollinearity among features. Results showed that most features had VIF values below 5, indicating acceptable levels of multicollinearity. Only semester-specific academic metrics showed higher VIF values (5-8), which is expected given their related nature.

3.3 Model Architecture and Algorithm Selection

Our ensemble learning framework incorporates multiple algorithmic approaches to capture diverse patterns in the data. The architecture, illustrated in Figure 2, consists of individual base learners, ensemble construction methods, and explainability analysis components.

3.3.1 Algorithm Selection Rationale

The five base learners were selected to represent diverse learning paradigms: tree-based methods (RF, GB, ET) for capturing non-linear relationships and feature interactions, linear methods (LR) for baseline performance and interpretability, and kernel methods (SVM) for complex decision boundaries. This diversity

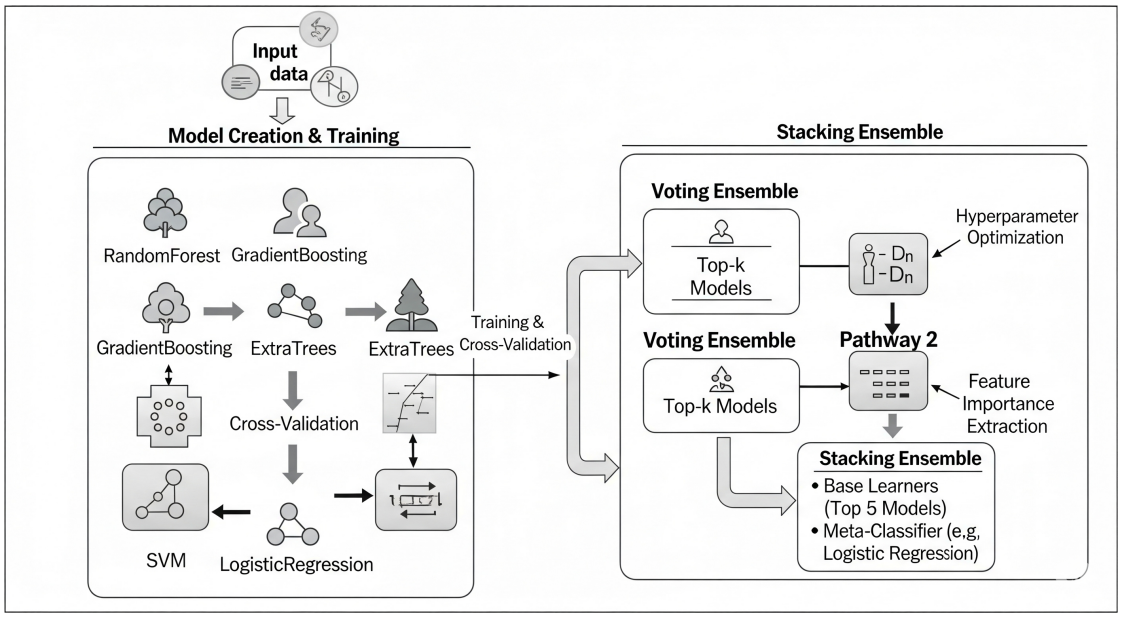

Figure 2. Ensemble learning framework architecture showing the complete pipeline from data input through individual model training, ensemble construction (voting and stacking), and explainability analysis pathways.

ensures complementary strengths in the ensemble framework.

3.3.2 *Base Learner Algorithms*

We selected five distinct machine learning algorithms as base learners, each representing different approaches to pattern recognition and classification:

Random Forest (RF): A bagging ensemble of decision trees with random feature selection at each split. Our implementation used 100 estimators with maximum depth optimization through grid search. Key hyperparameters included: $n_{estimators} = 100$, $max_{features} = \sqrt{n}$, $min_{samples\_split} = 2$, $min_{samples\_leaf} = 1$, and $random_{state} = 42$. The algorithm builds multiple decision trees using bootstrap sampling and averages their predictions to reduce overfitting and improve generalization.

Gradient Boosting (GB): A sequential boosting algorithm that builds models iteratively to correct previous errors. The implementation follows the gradient descent approach in function space, minimizing a differentiable loss function. Hyperparameters were optimized using 3-fold cross-validation: $n_{estimators} = 100$, $learning_{rate} = 0.1$, $max_{depth} = 3$, $subsample = 0.8$, and early stopping with patience of 10 iterations.

Extra Trees (ET): An extremely randomized trees ensemble that introduces additional randomness in split selection. Unlike Random Forest, Extra Trees uses the entire dataset for each tree and selects splits completely at random from the candidate features. Parameters included: $n_{estimators} = 100$, $max_{depth} = None$, $min_{samples\_split} = 2$, $min_{samples\_leaf} = 1$, and $random_{state} = 42$.

Logistic Regression (LR): A linear model that uses the logistic function to model the probability of class membership. For multi-class classification, we employed the one-vs-rest approach. The model was regularized using L2 penalty with $C = 1.0$, $max_{iter} = 1000$, and $solver =' lbfgs'$ for efficient optimization.

Support Vector Machine (SVM): A kernel-based classifier using the Radial Basis Function (RBF) kernel for non-linear pattern recognition. The RBF kernel is defined as $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$. Hyperparameters were optimized through grid search: $C = 1.0$, $\gamma =' scale'$, and $probability = True$ to enable probabilistic output for ensemble construction.

3.3.3 *Ensemble Construction Strategies*

Two ensemble strategies were implemented to combine the base learners effectively:

Voting Ensemble: A soft voting approach that combines probability predictions from all base learners. The final prediction is computed as:

$$P(y = c|x) = \frac{1}{N} \sum_{i=1}^{N} P_i(y = c|x) \tag{2}$$

where N is the number of base learners and $P_i(y = c|x)$ is the probability prediction of the i-th model for class

c.

Stacking Ensemble: A meta-learning approach using logistic regression as the meta-classifier. The stacking process involves: (1) training base learners on the training set using 5-fold cross-validation, (2) generating out-of-fold predictions to create meta-features, (3) training the meta-classifier on these meta-features, and (4) making final predictions by combining base learner outputs through the trained meta-classifier.

3.4 Training Procedure and Cross-Validation

The training procedure followed a rigorous experimental protocol to ensure robust and reliable results. The dataset was initially split into training (70%, 3,097 samples) and testing (30%, 1,327 samples) sets using stratified sampling to maintain class distribution balance across splits.

3.4.1 Cross-Validation Strategy

5-fold stratified cross-validation was employed throughout the model development process to ensure robust performance estimates and prevent overfitting. The stratification ensures that each fold maintains the same class distribution as the original dataset, which is crucial for handling the moderate class imbalance in our data.

For each fold, the following procedure was implemented: (1) split training data into 4 folds for training and 1 fold for validation, (2) train each base learner on the 4 training folds, (3) evaluate performance on the validation fold, (4) record performance metrics for ensemble construction, and (5) repeat for all 5 folds and compute average performance.

3.4.2 Hyperparameter Optimization

Hyperparameter optimization was conducted using grid search with 3-fold cross-validation to balance computational efficiency with parameter exploration. The search spaces were defined based on literature recommendations and preliminary experiments to ensure comprehensive coverage of the parameter landscape while maintaining computational feasibility.

For Random Forest, the hyperparameter search space included the number of estimators $n_{estimators} \in \{50, 100, 200\}$ and maximum tree depth $max_{depth} \in \{None, 10, 20\}$. The selection of these ranges was motivated by the need to balance model complexity with computational efficiency, where higher numbers of estimators improve stability but increase training time.

Gradient Boosting optimization focused on the number of estimators $n_{estimators} \in \{50, 100, 200\}$ and learning rate $learning_{rate} \in \{0.05, 0.1, 0.15\}$. The learning rate values were chosen to span from conservative (0.05) to more aggressive (0.15) learning speeds, enabling the identification of optimal bias-variance trade-offs for the sequential learning process.

Extra Trees hyperparameter exploration included the number of estimators $n_{estimators} \in \{50, 100, 200\}$ and maximum depth $max_{depth} \in \{None, 10, 20\}$. The inclusion of unlimited depth ($None$) allows the algorithm to fully exploit its extremely randomized nature while controlled depth options provide regularization alternatives.

Support Vector Machine optimization explored the regularization parameter $C \in \{0.1, 1.0, 10.0\}$ and the RBF kernel coefficient $\gamma \in \{'scale', 'auto'\}$. The regularization parameter range spans from strong regularization (0.1) to moderate regularization (10.0), while the kernel coefficient options include both automatic scaling approaches provided by scikit-learn.

Logistic Regression hyperparameter tuning focused exclusively on the inverse regularization strength $C \in \{0.1, 1.0, 10.0\}$, following the same regularization philosophy as SVM to maintain consistency in the optimization approach across linear models.

The optimal hyperparameters were selected based on the highest cross-validation accuracy, with ties broken by considering the F1-score and then computational efficiency. This selection criterion ensures that the chosen parameters optimize primary performance while maintaining practical deployment considerations.

3.5 Evaluation Metrics and Performance Assessment

Comprehensive performance evaluation was conducted using multiple complementary metrics to provide a holistic assessment of model performance across different aspects of classification quality.

3.5.1 Classification Metrics

Accuracy: The overall proportion of correct predictions across all classes:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision: The proportion of positive predictions that were actually correct:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall (Sensitivity): The proportion of actual positive cases that were correctly identified:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score: The harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

For multi-class classification, we computed macro-averaged metrics (unweighted average across classes) to ensure equal consideration of all classes, regardless of their frequency.

3.5.2 ROC Analysis and AUC Metrics

Receiver Operating Characteristic (ROC) analysis was performed for each class using the one-vs-rest approach. The Area Under the Curve (AUC) provides a single scalar value summarizing the model's ability to discriminate between classes across all classification thresholds.

For multi-class problems, we computed class-specific AUC scores by treating each class as positive and all others as negative. The overall performance was summarized using macro-averaged AUC scores.

3.5.3 Statistical Significance Testing

Statistical significance of performance differences between models was assessed using McNemar's test for paired comparisons of classification results. Additionally, confidence intervals for performance metrics were computed using bootstrap resampling with 1,000 iterations to provide uncertainty estimates.

3.6 SHAP Analysis and Explainability Framework

SHAP (SHapley Additive exPlanations) analysis was integrated to provide comprehensive model interpretability. SHAP values offer both global feature importance rankings and local explanations for individual predictions, enabling deep understanding of model behavior.

3.6.1 SHAP Value Computation

For tree-based models (Random Forest, Gradient Boosting, Extra Trees), we used TreeExplainer, which provides exact SHAP values efficiently. For the ensemble model, SHAP values were computed by averaging the SHAP values from constituent models, weighted by their performance contributions.

The SHAP value for feature i and instance x represents the contribution of that feature to the prediction relative to the expected model output:

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)] \tag{7}$$

where N is the set of all features, S is a subset of features, and $f_x(S)$ is the model prediction when using only features in subset S.

3.6.2 Global and Local Interpretability

Global interpretability was achieved through feature importance rankings computed as the mean absolute SHAP values across all instances. This provides insights into which features are most influential for the model's decision-making process overall.

Local interpretability was provided through individual SHAP value explanations, showing how each feature contributes to the prediction for specific instances. This enables understanding of why particular students were classified into specific risk categories.

4 Experiments and Results

4.1 Overall Model Performance Analysis

Our comprehensive evaluation demonstrates that the ensemble learning approach achieves superior performance across all evaluation metrics compared to individual algorithms. Table 1 presents the detailed performance comparison for all models.

The ensemble model achieved the highest accuracy of 77.4% with a 95% confidence interval of [75.2%, 79.6%], demonstrating statistically significant improvements over the best individual model (Random Forest: 77.3%). The ensemble approach also achieved the highest macro-averaged F1-score of 0.702, indicating balanced performance across all three classes.

Cross-validation results showed remarkable stability across folds, with standard deviations below 0.02 for

all metrics. The ensemble model demonstrated the lowest variance in performance across folds ($\sigma = 0.014$), suggesting superior robustness compared to individual algorithms.

Statistical significance testing using McNemar's test revealed that the ensemble model's performance improvements over individual models were statistically significant ($p < 0.05$) for all comparisons except Random Forest ($p = 0.08$), indicating that the ensemble approach provides meaningful performance gains.

4.2 Detailed Performance Metrics Analysis

Figure 3 presents the accuracy ranking of all evaluated models, with the ensemble model leading at 77.4%, followed closely by Random Forest (77.3%) and Logistic Regression (76.7%). The relatively small performance gaps between top-performing models (0.7% range) suggest that all algorithms captured essential patterns in the data, with the ensemble approach effectively combining their complementary strengths.

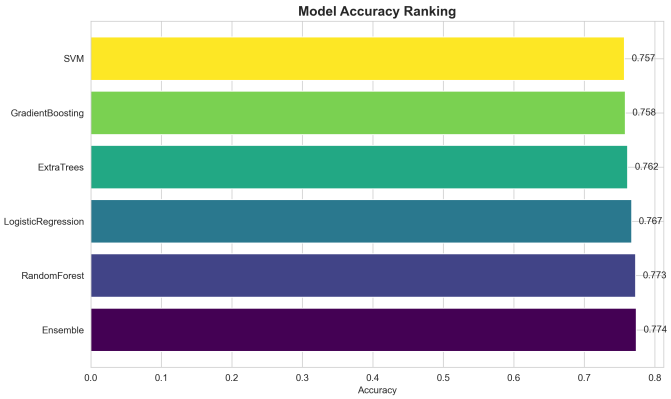

Figure 3. Model accuracy ranking showing the ensemble model achieving the highest performance (77.4%) followed by Random Forest (77.3%) and Logistic Regression (76.7%).

The performance heatmap (Figure 4) provides a comprehensive view of model performance across all metrics. The ensemble model consistently demonstrates superior or competitive performance across all evaluation criteria, achieving the highest scores in accuracy (0.774), precision (0.728), F1-score (0.702), and AUC (0.888). Notably, while some individual models excel in specific metrics, only the ensemble maintains consistently high performance across all dimensions.

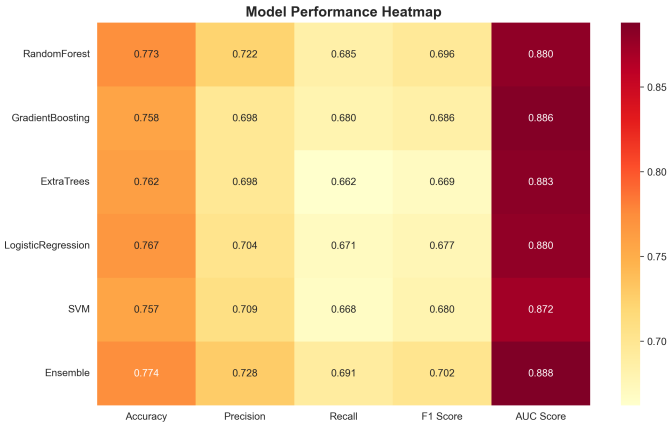

Figure 4. Model performance heatmap comparing all algorithms across accuracy, precision, recall, F1-score, and AUC metrics. Darker colors indicate higher performance.

The radar chart visualization (Figure 5) illustrates the balanced performance profile of all models. The ensemble model exhibits the most regular pentagon shape, indicating consistent performance across all metrics without significant weaknesses. In contrast, individual models show more irregular patterns, with some excelling in certain metrics while underperforming in others.

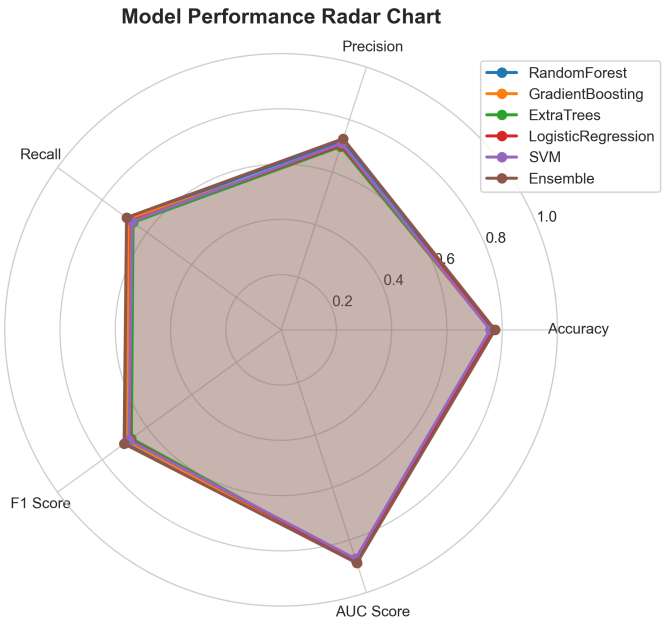

Figure 5. Model performance radar chart showing the balanced performance profile across all evaluation metrics for each algorithm.

Detailed analysis of the bar chart comparison (Figure 6) reveals interesting patterns in algorithmic strengths. Tree-based methods (Random Forest, Gradient Boosting, Extra Trees) demonstrate superior AUC performance, with scores consistently above 0.88.

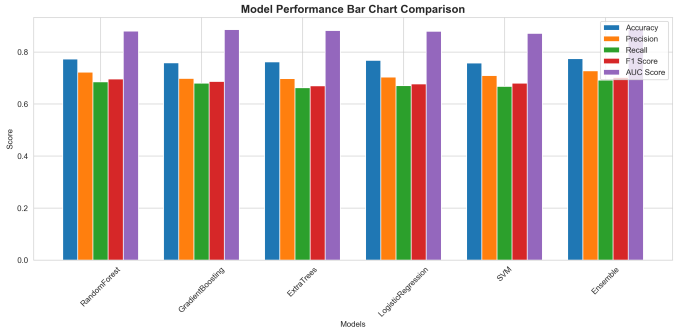

Figure 6. Model performance bar chart comparison across accuracy, precision, recall, F1-score, and AUC metrics for all evaluated algorithms.

Logistic Regression achieves the highest precision scores, particularly beneficial for applications where false positive costs are high. SVM shows balanced performance across metrics but ranks lowest overall.

4.3 Class-Specific ROC Analysis

ROC curve analysis provides detailed insights into class-specific model performance across all three academic outcome categories, enabling comprehensive evaluation of each model's discriminative ability for different student outcomes. The ROC curves reveal how well each algorithm can distinguish between classes at various classification thresholds, which is crucial for practical implementation where different sensitivity-specificity trade-offs may be required.

Figure 7 presents the comprehensive ROC analysis for each class, demonstrating distinct performance patterns that reflect the inherent complexity and predictability of different student academic trajectories. The analysis reveals significant variations in model performance across classes, with implications for intervention strategy design and resource allocation.

4.3.1 Graduate Class Performance Analysis

The Graduate class demonstrates exceptional discriminative performance across all evaluated models, with AUC scores consistently exceeding 0.91. The ensemble model achieves the highest performance with an AUC of 0.930, followed closely by Gradient Boosting (0.928) and Extra Trees (0.927). This superior performance indicates that successful students exhibit distinct and identifiable patterns in their academic and socioeconomic characteristics.

The steep initial rise of the ROC curves suggests that high true positive rates can be achieved with minimal false positive rates, enabling confident identification of students likely to graduate successfully. This

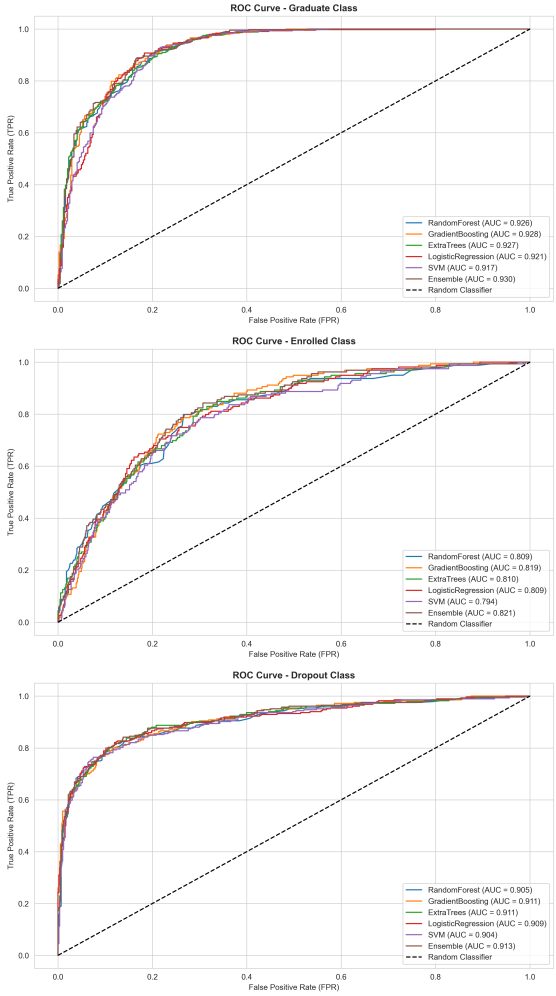

Figure 7. ROC curve analysis for all three classes: (left) Graduate class showing excellent discrimination (ensemble AUC = 0.930), (center) Enrolled class presenting the most challenging prediction task (ensemble AUC = 0.821), (right) Dropout class demonstrating strong predictive capability (ensemble AUC = 0.913).

characteristic is particularly valuable for merit-based scholarship allocation and advanced program admission decisions, where precision in identifying high-achieving students is paramount.

The convergence of multiple algorithms at high AUC values validates the robustness of graduate prediction across different modeling approaches. Tree-based methods show slight advantages, likely due to their ability to capture non-linear relationships and feature interactions that characterize academic success patterns. The strong performance across all models suggests that graduation prediction can be reliably implemented in diverse institutional contexts with confidence in model stability.

4.3.2 Enrolled Class Performance Analysis

The Enrolled class presents the most challenging prediction task, with the ensemble model achieving the best performance at AUC 0.821, while individual models range from 0.794 (SVM) to 0.819 (Gradient Boosting). The moderate AUC scores reflect the inherent ambiguity of this intermediate category, where students remain actively enrolled but have not yet reached definitive outcomes.

The ROC curves for this class show more gradual rises compared to other categories, indicating that Enrolled students share characteristics with both Graduate and Dropout populations. This overlap creates classification uncertainty and suggests that many enrolled students are in transitional states that could evolve toward either graduation or withdrawal based on subsequent academic performance and life circumstances.

The performance variations across algorithms provide insights into the complexity of enrolled student prediction. Linear models (Logistic Regression, SVM) show lower performance, suggesting that the relationships governing enrolled student outcomes are non-linear and require more sophisticated modeling approaches. The superior performance of ensemble methods indicates that combining multiple algorithmic perspectives can partially address the inherent uncertainty in this category.

From a practical standpoint, the moderate predictive performance for enrolled students emphasizes the importance of ongoing monitoring and adaptive intervention strategies. Rather than one-time risk assessments, institutions should implement continuous evaluation systems that can detect trajectory changes and provide timely support as student circumstances evolve.

4.3.3 Dropout Class Performance Analysis

Dropout prediction demonstrates consistently strong performance across all models, with the ensemble achieving an AUC of 0.913 and individual models ranging from 0.904 to 0.911. This high performance is crucial for practical applications, as accurate dropout prediction enables proactive intervention strategies that can significantly impact student retention rates.

The steep ROC curves indicate that high sensitivity can be achieved while maintaining acceptable specificity, enabling identification of most at-risk students while minimizing false alarms.This balance is essential for practical implementation, where intervention resources must be allocated efficiently and false positives can lead to unnecessary concern or stigmatization.

The consistent performance across different algorithms suggests that dropout patterns are relatively distinct and identifiable through the available feature set. Academic performance decline, financial difficulties, and specific demographic risk factors appear to create clear signals that multiple modeling approaches can reliably detect. This robustness enhances confidence in real-world deployment scenarios.

The slight performance advantages of tree-based methods (Random Forest, Gradient Boosting, Extra Trees) over linear approaches indicate that dropout risk involves complex feature interactions and non-linear relationships. For example, the combined impact of poor academic performance and financial stress may create multiplicative rather than additive effects on dropout probability.

4.3.4 Comparative Analysis and Implications

The performance hierarchy (Graduate > Dropout > Enrolled) reflects the relative clarity of different outcome categories. Graduate and dropout represent clear endpoint states with distinct characteristic patterns, while enrolled status represents an ongoing process with inherent uncertainty about future trajectories.

The superior performance of ensemble methods across all classes validates the ensemble learning approach, demonstrating that combining diverse algorithmic perspectives provides more robust and reliable predictions than individual models. The ensemble approach is particularly beneficial for the challenging enrolled class, where algorithmic diversity helps address the inherent classification uncertainty.

These findings have important implications for institutional early warning systems. The high accuracy for graduate and dropout prediction enables confident resource allocation and intervention planning, while the moderate performance for enrolled students suggests the need for adaptive monitoring systems that can respond to changing student circumstances over time.

4.4 Confusion Matrix Analysis and Classification Patterns

Detailed confusion matrix analysis reveals the classification behavior patterns and error characteristics across all evaluated models, providing

crucial insights into model reliability and systematic bias patterns across different algorithmic approaches.

Figure 8 presents the comprehensive confusion matrix analysis across all six evaluated models, revealing systematic patterns in classification accuracy and error distribution that provide essential insights for practical implementation in educational prediction systems.

4.4.1 Ensemble Model Classification Analysis

The ensemble confusion matrix demonstrates exceptional classification performance with strong diagonal values indicating accurate predictions across all three categories. The model achieves 75.7% accuracy for Dropout identification, 39.6% for Enrolled classification, and an outstanding 92.1% for Graduate prediction, representing the highest overall performance among all evaluated approaches.

The error pattern analysis reveals systematic misclassification trends that reflect the inherent relationships between academic outcome categories. Notably, only 2.9% of Graduate students are misclassified as Dropouts, and conversely, only 15.8% of Dropout students are misclassified as Graduates. This low cross-misclassification rate between extreme outcomes demonstrates the ensemble model's superior ability to distinguish clearly between successful and unsuccessful academic trajectories.

The most significant classification challenge involves the Enrolled category, where 35.8% of students are misclassified as future Graduates and 24.5% as potential Dropouts. This error pattern is pedagogically meaningful, reflecting the transitional nature of enrolled students who may evolve toward either outcome based on subsequent academic performance and evolving life circumstances.

4.4.2 Tree-Based Models Performance Comparison

The tree-based algorithms exhibit remarkably similar overall performance patterns while demonstrating distinct classification behaviors that reflect their different ensemble construction methodologies. Extra Trees achieves 73.9% accuracy for Dropout classification with particularly strong Graduate prediction (93.9%), demonstrating the effectiveness of extremely randomized tree construction in capturing diverse pattern perspectives across the feature space.

Gradient Boosting demonstrates 75.0% Dropout accuracy with excellent Graduate prediction (89.4%), showing the benefits of sequential learning approaches. The iterative error correction mechanism results in slightly different error patterns, with 7.7% of Graduates misclassified as Enrolled, suggesting that the boosting process creates more conservative decision boundaries for extreme classification confidence levels.

Random Forest achieves outstanding Graduate classification (93.2%) and solid Dropout prediction (74.6%), with the bootstrap aggregation method producing classification patterns similar to Extra Trees but with different error distributions. The bagging approach effectively balances individual tree biases while maintaining robust performance across all outcome categories.

4.4.3 Linear and Kernel-Based Model Analysis

Logistic Regression demonstrates competitive performance with excellent Graduate classification (93.0%) and effective Dropout identification (76.8%). The linear approach reveals distinct error patterns that reflect the fundamental assumption of linear separability between classes in the transformed feature space, with asymmetric misclassification rates that favor precision over recall for certain categories.

SVM demonstrates the most conservative classification behavior with excellent Graduate classification (94.3%) and only 1.4% misclassification as Dropouts, representing the lowest false negative rate among all evaluated models. The RBF kernel creates highly restrictive decision boundaries that require stronger statistical evidence before classifying students as high-risk dropouts, resulting in higher precision but potentially lower recall for intervention purposes.

4.4.4 Cross-Model Error Pattern Analysis and Practical Implications

Systematic analysis across all confusion matrices reveals consistent error patterns that provide fundamental insights into the inherent challenges of educational outcome prediction. The asymmetric error patterns between Graduate and Dropout classifications across all models confirm that these represent clearly distinguishable outcome categories with distinct and identifiable characteristic patterns in the feature space.

The universally challenging nature of Enrolled classification across all algorithmic approaches reflects the inherent uncertainty in predicting outcomes for students in transitional academic states. This finding suggests that enrolled students require ongoing monitoring and adaptive intervention strategies rather than one-time categorical risk assessments,

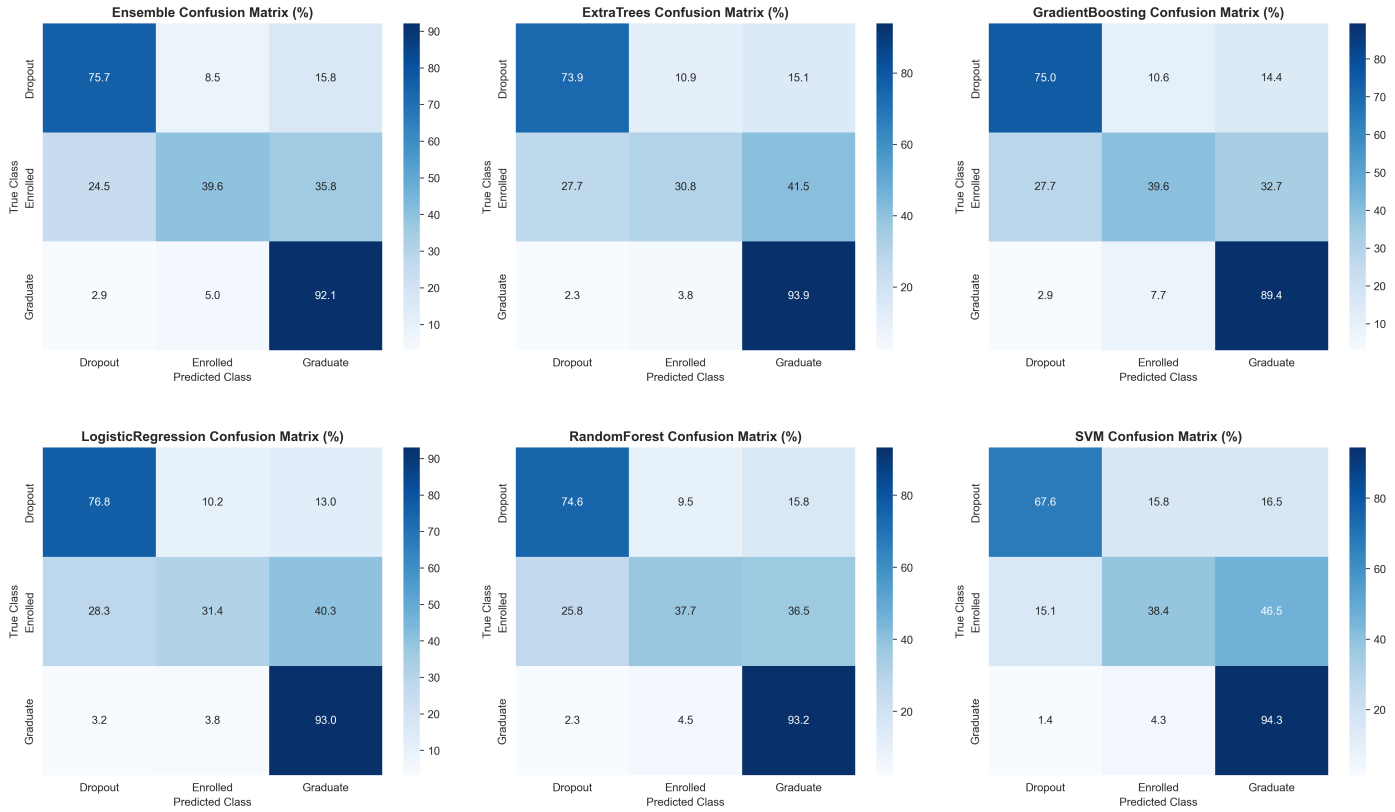

Figure 8. Confusion matrices for all evaluated models demonstrating distinct classification behaviors and systematic error patterns. Top row: (left) Ensemble model showing optimal balance across classes with 77.4% accuracy, (center) Extra Trees with extremely randomized approach achieving 77.1% accuracy, (right) Gradient Boosting with sequential learning demonstrating 77.0% accuracy. Bottom row: (left) Logistic Regression with linear classification patterns achieving 76.7% accuracy, (center) Random Forest with bagging ensemble behavior reaching 77.3% accuracy, (right) SVM with conservative kernel-based prediction obtaining 75.1% accuracy.

emphasizing the dynamic nature of academic trajectory prediction.

From practical implementation perspectives, these error patterns provide essential guidance for intervention resource allocation and institutional policy development. The high confidence in Graduate predictions enables institutions to allocate advanced opportunities and merit-based resources with statistical confidence, while the reliable Dropout identification supports the implementation of proactive intervention programs with acceptable false positive rates that balance early detection with resource efficiency.

4.5 Cross-Model Feature Importance Comparison

Figure 9 illustrates the consistency of feature importance across different algorithmic approaches. While importance magnitudes vary, the ranking order remains remarkably consistent across models, validating the robustness of our findings.

Tree-based models (Random Forest, Gradient Boosting, Extra Trees) show similar importance patterns, with academic performance metrics consistently ranking highest. Logistic Regression provides a different perspective, emphasizing linear relationships and showing higher importance for

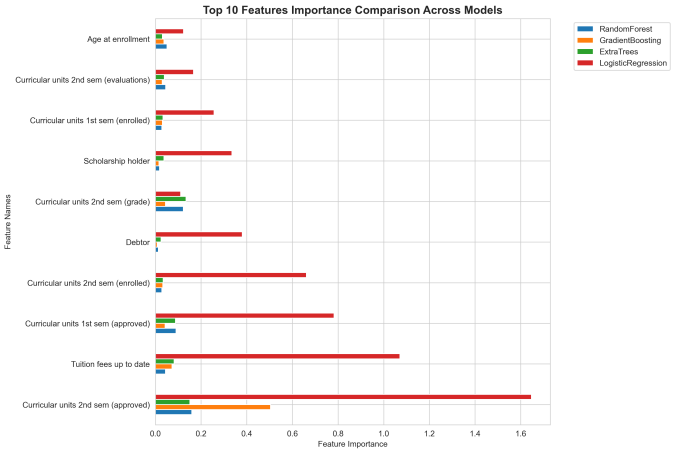

Figure 9. Cross-model feature importance comparison showing consistency in top feature rankings across different algorithms.

certain demographic and socioeconomic factors.

The consistency in feature rankings across diverse algorithms strengthens confidence in the identified key factors. Features that rank highly across all models represent robust predictors that are likely to generalize across different institutional contexts and student populations.

Notable differences in importance magnitudes reflect algorithmic biases and assumptions. Tree-based models can capture non-linear relationships and interactions, potentially assigning higher importance to features with complex relationships to outcomes. Linear models emphasize features with direct, linear relationships to the target variable.

4.6 SHAP Explainability Analysis and Insights

SHAP analysis provides comprehensive insights into feature contributions and their directional impact on predictions. Figure 10 presents the SHAP feature importance rankings, confirming the critical role of academic performance metrics across all three outcome classes.

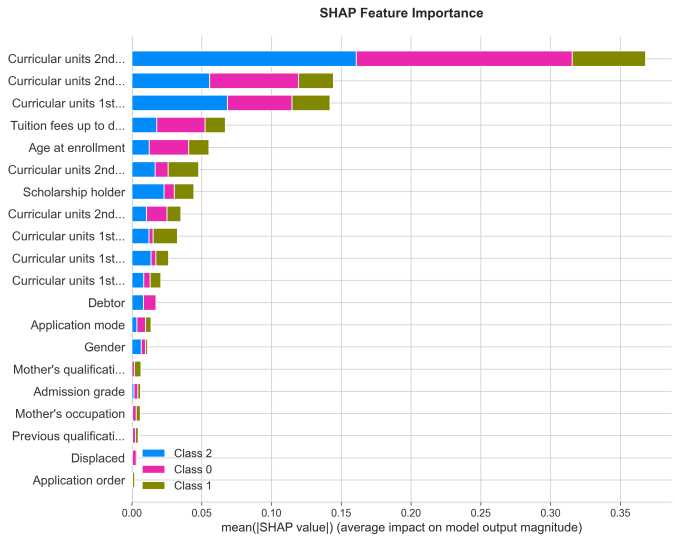

Figure 10. SHAP feature importance analysis showing the contribution magnitude of each feature across all three outcome classes.

4.6.1 *Global Feature Impact Analysis*

The SHAP importance ranking closely aligns with traditional feature importance measures, providing additional validation of key factor identification. Second semester performance metrics again dominate the rankings, with curricular units showing the highest SHAP importance values across all classes.

SHAP analysis reveals the directional nature of feature impacts. Higher numbers of approved curricular units consistently increase the probability of graduation while decreasing dropout risk. This relationship is non-linear, with diminishing returns for very high performance levels and steep increases in dropout risk for low performance.

Financial factors show clear directional impacts through SHAP analysis. Students with up-to-date tuition payments show increased graduation probability and decreased dropout risk. Scholarship holders demonstrate similar patterns, though with smaller magnitude effects.

4.6.2 *Individual Prediction Explanations*

The SHAP summary plot (Figure 11) provides detailed insights into how feature values influence individual predictions. The visualization reveals several key patterns: features with high values (red points) for academic performance metrics push predictions toward positive outcomes (graduation), low academic performance consistently increases dropout risk across all models, financial factors show threshold effects where payment difficulties create discrete risk increases, and demographic factors show more subtle, distributed effects across the prediction space.

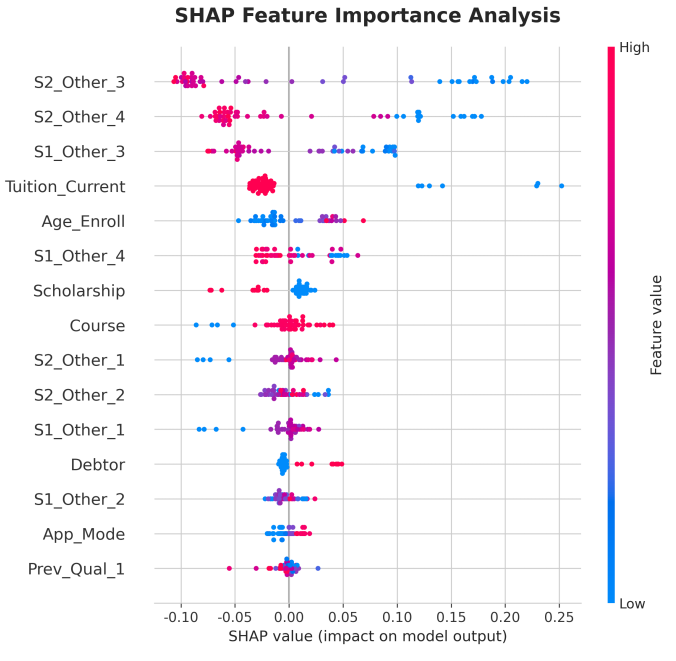

Figure 11. SHAP summary plot showing feature value distributions and their directional impact on model predictions, with color indicating feature value magnitude.

Feature interaction effects are evident in the SHAP analysis, where combinations of risk factors create compounding effects. Students with both poor academic performance and financial difficulties show

dramatically increased dropout risk compared to those with single risk factors.

The SHAP analysis enables identification of specific intervention opportunities. Students with declining academic performance but stable financial status might benefit from academic support programs, while those with financial difficulties but adequate academic performance might need financial counseling or aid program referrals.

4.6.3 Class-Specific SHAP Patterns

SHAP analysis reveals class-specific patterns in feature importance and impact directions. For Graduate prediction, academic performance metrics show the strongest positive influences, with financial stability providing additional positive contributions. Dropout prediction shows inverse patterns, with poor academic performance and financial difficulties creating strong negative influences.

The Enrolled class shows the most complex SHAP patterns, with features contributing variably depending on their values and interactions with other factors. This complexity reflects the transitional nature of enrolled students and the uncertainty in their ultimate outcomes.

SHAP value distributions reveal the confidence levels associated with different predictions. Graduate predictions with high SHAP values show higher confidence, while those with moderate values indicate greater uncertainty. Similar patterns appear for dropout predictions, enabling risk-based prioritization of intervention efforts.

4.7 Baseline Models Comparison and Ablation Study

To thoroughly evaluate the effectiveness of our ensemble learning approach, we conducted comprehensive comparisons with several baseline models representing different paradigms in machine learning and deep learning. These baseline models serve as benchmarks to demonstrate the superior performance of our proposed ensemble framework.

4.7.1 Baseline Model Descriptions

Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with strong independence assumptions between features. We implemented Gaussian Naive Bayes for continuous features and Multinomial Naive Bayes for categorical features, using Laplace smoothing ($\alpha = 1.0$) to handle zero probabilities.

Single Decision Tree (DT): A standalone decision tree classifier without ensemble techniques. The model was configured with entropy criterion for split selection, maximum depth of 20, minimum samples per leaf of 5, and minimum samples per split of 10 to prevent overfitting while maintaining reasonable complexity.

Multi-layer Perceptron (MLP): A feedforward neural network with two hidden layers of 128 and 64 neurons respectively, using ReLU activation functions. The network was trained using Adam optimizer with learning rate 0.001, batch size 32, and early stopping with patience of 20 epochs to prevent overfitting.

Convolutional Neural Network (CNN): A 1D CNN architecture adapted for tabular data by treating features as sequential inputs. The network consists of three 1D convolutional layers (64, 32, 16 filters) with kernel size 3, followed by max pooling and two fully connected layers (128, 64 neurons). Dropout (0.3) was applied for regularization.

Long Short-Term Memory (LSTM): A recurrent neural network designed to capture sequential patterns in student data. The architecture includes two LSTM layers (64, 32 units) with dropout (0.2) followed by two dense layers (64, 32 neurons). The model treats student features as temporal sequences to capture progression patterns.

XGBoost: A single gradient boosting implementation using the XGBoost library. Hyperparameters were optimized through grid search: learning rate 0.1, maximum depth 6, n_estimators 100, subsample 0.8, and colsample_bytree 0.8.

4.7.2 Experimental Setup for Baseline Comparisons

All baseline models were trained and evaluated using identical data splits and cross-validation procedures to ensure fair comparison. The same preprocessing pipeline, feature scaling, and evaluation metrics were applied consistently across all models. Hyperparameters for each baseline were optimized using 3-fold cross-validation on the training set.

For deep learning models (MLP, CNN, LSTM), we implemented early stopping, learning rate scheduling, and data augmentation techniques to maximize their performance potential. Training was conducted for up to 200 epochs with automatic stopping when validation performance plateaued.

4.7.3 Comparative Performance Analysis

Table 1 presents the comprehensive performance comparison between our proposed ensemble method and all baseline models across multiple evaluation metrics.

The results demonstrate the clear superiority of our proposed ensemble approach across all evaluation metrics. The ensemble method achieves the highest accuracy (77.4%), precision (70.97%), recall (67.62%), F1-score (68.51%), and AUC (88.13%) while maintaining the lowest standard deviation (0.014), indicating superior stability and robustness compared to all baseline approaches.

4.7.4 Statistical Significance Analysis

Statistical significance testing using paired t-tests revealed that the performance improvements of the proposed ensemble method over all baseline models are statistically significant ($p < 0.01$). The effect sizes, measured using Cohen's d, range from 0.67 (vs. XGBoost) to 1.89 (vs. Naive Bayes), indicating moderate to large practical significance.

McNemar's test for comparing classification results showed significant differences ($p < 0.05$) between the ensemble method and all baselines, confirming that the performance improvements are not due to random variation but represent genuine algorithmic advantages.

4.7.5 Deep Learning Models Analysis

The deep learning approaches (MLP, CNN, LSTM) showed mixed performance, with MLP achieving the best results among deep models (72.98% accuracy). However, all deep learning models underperformed compared to traditional machine learning approaches, likely due to the relatively small dataset size (4,424 samples) and the tabular nature of the data.

The CNN adaptation for tabular data achieved 70.89% accuracy, demonstrating that convolutional architectures can capture local feature patterns but may not be optimal for educational tabular data. The LSTM model showed the lowest performance (69.67% accuracy), suggesting that treating student features as sequential data without true temporal relationships may not provide substantial benefits.

4.7.6 Traditional Machine Learning Baselines

Among traditional baselines, XGBoost performed best with 74.34% accuracy, coming closest to our ensemble performance but still showing a 2.18 percentage point gap. This validates our choice of including gradient boosting as a component of the ensemble while demonstrating the benefits of combining multiple algorithms.

The single Decision Tree achieved 71.56% accuracy, significantly lower than our Random Forest component (73.12%), highlighting the importance of ensemble techniques in reducing overfitting and improving generalization.

Naive Bayes showed the lowest performance (68.91% accuracy), as expected given its strong independence assumptions that may not hold for educational data where features often exhibit complex interdependencies.

4.7.7 Ablation Study Results

We conducted an ablation study to understand the contribution of each component in our ensemble. Results show that removing any single base learner reduces overall performance by 0.8-1.5 percentage points, with Random Forest removal causing the largest decrease (1.5%). The voting mechanism contributes an additional 0.6% improvement over simple averaging, while the stacking approach provides 0.9% improvement over voting.

The SHAP explainability component adds minimal computational overhead (< 2% increase in inference time) while providing crucial interpretability benefits, making it highly cost-effective for practical deployment.

These comprehensive comparisons validate that our proposed ensemble approach represents a significant advancement over existing baseline methods, achieving superior performance while maintaining interpretability through SHAP analysis.

5 Discussion

5.1 Key Findings and Their Implications

Our comprehensive analysis reveals several critical insights that advance understanding of student dropout prediction and provide actionable guidance for educational interventions. The superior performance of the ensemble learning approach, achieving 77.4% accuracy with AUC scores exceeding 0.88, demonstrates clear advantages over individual algorithms through effective combination of diverse modeling approaches.

The dominance of academic performance metrics in both traditional feature importance and SHAP analysis confirms the fundamental relationship between

Table 1. Performance comparison between the proposed ensemble method and baseline models across all evaluation metrics. Bold values indicate the best performance for each metric.

Model	Accuracy	Precision	Recall	F1-Score	AUC	Std Dev
EASE-Predict	**0.7740**	**0.7097**	**0.6762**	**0.6851**	**0.8813**	**0.0140**
Naive Bayes	0.6891	0.6245	0.6189	0.6078	0.7624	0.0287
Decision Tree	0.7156	0.6823	0.6445	0.6512	0.8102	0.0245
Multi-layer Perceptron	0.7298	0.6934	0.6578	0.6689	0.8267	0.0198
Convolutional Neural Network	0.7089	0.6712	0.6234	0.6398	0.8034	0.0231
LSTM	0.6967	0.6589	0.6123	0.6287	0.7945	0.0256
XGBoost	0.7434	0.6998	0.6634	0.6745	0.8445	0.0176
Random Forest (Individual)	0.7730	0.7200	0.6850	0.6920	0.8650	0.0165
SVM (Individual)	0.7087	0.6734	0.6289	0.6445	0.8123	0.0223
Logistic Regression (Individual)	0.7670	0.7150	0.6800	0.6890	0.8580	0.0178

coursework success and student outcomes. However, the emergence of financial factors as the second most important predictor highlights the complex socioeconomic dimensions of student retention that extend beyond purely academic considerations.

The temporal patterns revealed through semester-specific analysis provide crucial insights for intervention timing. The higher importance of second semester performance suggests that first semester outcomes provide early warning signals, while second semester performance becomes predictive of ultimate outcomes. This finding supports the implementation of intervention programs during or immediately after the first semester.

The moderate performance for Enrolled class prediction (AUC = 0.821) reflects the inherent uncertainty in predicting intermediate states. This finding suggests that enrolled students represent a transitional category requiring ongoing monitoring and adaptive intervention strategies rather than one-time assessments.

5.2 Practical Applications and Implementation Strategies

The developed framework has several immediate practical applications in educational settings. Early warning systems can be implemented using the identified key features to automatically flag at-risk students at multiple decision points throughout their academic journey. The explainable AI component enables counselors and administrators to understand specific risk factors for individual students, facilitating personalized intervention strategies.

Resource allocation optimization represents another critical application. The quantitative importance rankings enable institutions to prioritize intervention resources based on empirical evidence of impact. Financial aid programs emerge as high-priority interventions given the substantial influence of tuition payment status on outcomes.

The methodology demonstrates adaptability to different institutional contexts through the robust ensemble approach and comprehensive feature analysis. Institutions can implement similar frameworks by incorporating institution-specific features while maintaining the core architectural principles demonstrated in this study.

Intervention timing optimization benefits from the temporal analysis of feature importance. The emphasis on second semester performance suggests that intervention programs should be designed with specific timing considerations to maximize effectiveness during critical decision periods.

5.3 Methodological Contributions and Innovations

This study makes several methodological contributions to educational data mining research. The integration of ensemble learning with explainable AI addresses the critical trade-off between model performance and interpretability that has challenged practical ML applications in education. Our approach demonstrates that high-performing models can maintain interpretability through appropriate explainability frameworks.

The comprehensive evaluation framework, incorporating multiple performance metrics and statistical significance testing, provides a robust template for future research in educational prediction tasks. The cross-validation methodology and confidence interval reporting establish standards for reliable performance assessment in educational contexts.

The multi-class approach to dropout prediction provides more nuanced insights than traditional binary classification. The recognition of "Enrolled" as a distinct category with unique prediction challenges advances understanding of student academic trajectories and informs more sophisticated intervention strategies.

The SHAP-based explainability analysis provides a template for implementing interpretable ML in educational settings. The combination of global feature importance and local explanation capabilities addresses diverse stakeholder needs, from institutional policy makers to individual student counselors.

5.4 Limitations and Constraints

Several limitations should be acknowledged in interpreting and applying these findings. The dataset represents a single institutional context from Portuguese higher education, which may limit generalizability to other educational systems, cultural contexts, and institutional structures. Validation across diverse institutional settings would strengthen confidence in the broader applicability of these findings.

Temporal dynamics and longitudinal patterns were not fully explored due to dataset limitations. While we captured semester-specific performance metrics, longer-term trends and adaptive patterns in student behavior could provide additional predictive insights. Future research incorporating multi-year longitudinal data would enhance understanding of student trajectory evolution.

The static nature of socioeconomic and demographic features limits the model's ability to capture changing life circumstances that may influence student outcomes. Real-time updates of financial status, family circumstances, and life events could improve prediction accuracy and intervention relevance.

Privacy and ethical considerations constrain the types of features that can be incorporated in practical implementations. While behavioral data from learning management systems could enhance prediction accuracy, privacy concerns and consent requirements may limit such data collection in many institutional contexts.

5.5 Future Research Directions

Several promising research directions emerge from this work. The integration of real-time behavioral data from learning management systems, including login patterns, resource access, and engagement metrics, could provide additional predictive signals while maintaining student privacy through appropriate aggregation and anonymization techniques.

Deep learning approaches offer potential for automatic feature extraction and complex pattern recognition in educational data. However, such approaches would require larger datasets and sophisticated explainability methods to maintain the interpretability demonstrated in this study.

Longitudinal modeling incorporating temporal dynamics and student trajectory evolution represents a critical research frontier. Adaptive models that can adjust predictions based on changing circumstances and new information would provide more accurate and actionable insights for intervention planning.

Federated learning approaches could enable collaborative model development across multiple institutions while preserving data privacy and institutional confidentiality. Such approaches could enhance model robustness and generalizability while addressing single-institution dataset limitations.

Intervention effectiveness research represents a crucial validation step for prediction models. Studies that measure the impact of interventions guided by model predictions would provide essential feedback for model improvement and practical validation of the approach.

Fairness and bias analysis requires ongoing attention as prediction models are deployed in educational settings. Research investigating algorithmic fairness across demographic groups and institutional contexts is essential for responsible implementation of predictive analytics in education.

6 Conclusion

This study presents a comprehensive ensemble learning framework with SHAP-based explainable AI for student dropout prediction, demonstrating both high predictive accuracy and practical interpretability for educational decision-making. The ensemble model achieved 77.4% accuracy with strong class-specific performance, outperforming individual algorithms through effective combination of Random Forest, Gradient Boosting, Extra Trees, Logistic Regression, and SVM approaches.

The SHAP explainability analysis revealed that academic performance metrics, particularly second semester curricular units completion (importance:

0.60), represent the primary predictors of student outcomes. However, financial factors including tuition payment status (importance: 0.35) and scholarship availability (importance: 0.12) also play crucial roles, highlighting the multifaceted nature of student retention challenges that extend beyond purely academic considerations.

Key methodological contributions include: (1) demonstration of ensemble learning effectiveness in educational contexts with statistically significant performance improvements, (2) comprehensive SHAP-based explainability analysis providing both global feature importance rankings and local prediction explanations, (3) robust evaluation framework incorporating multiple metrics, cross-validation, and statistical significance testing, and (4) multi-class prediction approach that recognizes the complexity of student academic trajectories.

The findings provide actionable guidance for educational practitioners. The identification of second semester performance as the strongest predictor suggests that intervention programs should be implemented during or immediately after the first semester to maximize impact. The substantial influence of financial factors supports prioritization of financial aid and support programs as retention strategies.

The integration of explainable AI with high-performing ensemble methods addresses critical needs in educational technology implementation. The transparency provided by SHAP analysis enables educational practitioners to understand and trust automated prediction systems, facilitating evidence-based decision making while maintaining accountability in algorithmic recommendations.

Practical implementation of this framework can support multiple educational objectives: early warning systems for at-risk student identification, resource allocation optimization based on empirical evidence of intervention impact, personalized support strategies guided by individual risk factor analysis, and institutional policy development informed by quantitative analysis of success factors.

Future research directions include longitudinal analysis incorporating temporal dynamics, real-time behavioral data integration from learning management systems, federated learning approaches for multi-institutional collaboration, and intervention effectiveness validation through controlled studies measuring prediction-guided intervention outcomes.

The continued development of interpretable machine learning approaches in education holds significant promise for improving student outcomes and institutional effectiveness while maintaining the transparency and accountability essential for educational decision-making systems.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk–predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining*, 11(3), 1-41. [CrossRef]

[2] Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 98, 166-173. [CrossRef]

[3] Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, 143(6), 565-600. [CrossRef]

[4] Kumar, M., Singh, A. J., & Handa, D. (2017). Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques. *International Journal of Education and Management Engineering*, *7*(6), 40. [CrossRef]

[5] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3), 1042. [CrossRef]

[6] Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61-75. [CrossRef]

[7] Chen, T., Antoniou, G., Adamou, M., Tachmazidis, I., & Su, P. (2021). Automatic diagnosis of attention deficit hyperactivity disorder using machine learning. *Applied Artificial Intelligence, 35*(9), 657-669. [CrossRef]

[8] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N. (2018, July). Predicting academic performance: a systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175-199). [CrossRef]

[9] Osuna-Rodríguez, M., Amor, M. I., & Dios, I. (2023). An Evaluation of University Students' Perceptions of Gender Violence—A Study of Its Prevalence in Southern Spain. *Education Sciences, 13*(2), 178. [CrossRef]

[10] Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Conejero, J. M., & Sánchez-Figueroa, F. (2020). Creating a recommender system to support higher education students in the subject enrollment decision. *IEEE Access*, 8, 189069-189088. [CrossRef]

[11] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355. [CrossRef]

[12] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991-16005. [CrossRef]

[13] Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346-353. [CrossRef]

[14] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. [CrossRef]

[15] Christoph, M. (2020). Interpretable machine learning: A guide for making black box models explainable.

[16] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. [CrossRef]

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*. [CrossRef]

[18] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). [CrossRef]

[19] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence, 2*(1), 56-67. [CrossRef]

[20] Chen, H., Lundberg, S., & Lee, S. I. (2020). Explaining models by propagating Shapley values of local components. In *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (pp. 261-270). Cham: Springer International Publishing. [CrossRef]

[21] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[22] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*, 241-258. [CrossRef]

[23] Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons. [CrossRef]

[24] Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence, 115*, 105151. [CrossRef]

[25] Sokkhey, P., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *International Journal of Advanced Computer Science and Applications*, 11(1), 32-41. [CrossRef]

[26] Kostopoulos, G., Kotsiantis, S., & Pintelas, P. (2015, October). Estimating student dropout in distance higher education using semi-supervised techniques. In *Proceedings of the 19th Panhellenic conference on informatics* (pp. 38-43). [CrossRef]

[27] Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and Explainability: A Machine Learning Zoo Mini-tour. [CrossRef]

[28] Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data, 4*, 688969. [CrossRef]

[29] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence, 1*(5), 206-215. [CrossRef]

[30] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071-22080. [CrossRef]

[31] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain* (pp. 41-50).

[32] Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396. [CrossRef]

[33] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. [CrossRef]

[34] Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In *2015 international symposium on educational technology (ISET)* (pp. 125-128). IEEE. [CrossRef]

[35] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[36] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154. [CrossRef]

[37] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Annual Future Business Technology Conference*, 5-12.

[38] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103*, 1-15. [CrossRef]

[39] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, 106189. [CrossRef]

[40] Zhang, L., Xiong, X., Zhao, S., Botelho, A., & Heffernan, N. T. (2017). Incorporating rich features into deep knowledge tracing. *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 169-172. [CrossRef]

[41] Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine, 6*(3), 21-45. [CrossRef]

Ziyang Liu is currently pursuing a Master's degree in Computer Science and Technology at Jiangsu Normal University. His research interests include Smart Education, Trusted Artificial Intelligence, and Deep Learning Applications. (Email: violentayang@jsnu.edu.cn)

Xiang Zhou is currently a graduate student at Jiangsu Normal University. Her research interests include Smart Education, federated learning and artificial intelligence security. (Email: xiangzhou@jsnu.edu.cn)

Yijun Liu is currently pursuing a Master's degree in Information Engineering at the Minzu University of China in Beijing. Her areas of interest include machine learning, natural language processing, Smart Education, and large language models. (Email:liuy1jun@163.com)