Machine Learning Frameworks for Student Retention Prediction: A Comparative Analysis of Ensemble and Traditional Classifiers

Dr Mazen M Salama
Senior Data Scientist - Dataemia - USA

Student retention remains a critical challenge in educational institutions, where early identification of at-risk learners is essential for implementing timely interventions. To address this, we developed a comprehensive machine learning framework to predict student dropout outcomes using an educational dataset comprising 649 records and 33 demographic and academic features. We systematically evaluated ten classification algorithms, including logistic regression, support vector machines, k-nearest neighbors, decision trees, multiple ensemble methods (random forest, gradient boosting, XGBoost, LightGBM, AdaBoost), and a multilayer perceptron neural network. Our analysis demonstrates that the random forest model achieves superior overall performance, attaining an accuracy of 96.2%, an F1-score of 0.857, and a ROC-AUC of 0.955, while maintaining robust stability across five-fold cross-validation (mean F1: 0.811, std: 0.038). Feature importance analysis identifies final academic grades, prior semester performance, and historical failure counts as the most predictive indicators of retention status. Notably, the model achieves a 75% recall for dropped-out students, which is crucial for early intervention. These findings highlight the efficacy of ensemble tree-based methods in educational data mining and provide actionable, interpretable insights for designing targeted retention strategies.

I. INTRODUCTION

Student retention represents a fundamental challenge for higher education institutions, with dropout rates directly impacting institutional sustainability, resource allocation, and student socioeconomic mobility. The capacity to proactively identify at-risk learners has emerged as a critical capability for deploying timely academic advising and targeted support programs. However, accurately forecasting student attrition is inherently complex. Educational environments generate data characterized by high dimensionality, contextual noise, missing values, and pronounced class imbalance, where the majority of students successfully complete their programs while a minority discontinue. These characteristics create non-linear, interacting relationships among demographic backgrounds, enrollment histories, and academic performance metrics that traditional statistical models often fail to capture without extensive manual feature engineering and restrictive parametric assumptions. Consequently, there is a pressing need for robust, data-driven approaches that can navigate these complexities while remaining transparent enough for academic stakeholders to trust and act upon.

To address these limitations, this study introduces a comprehensive machine learning framework designed to systematically evaluate and compare traditional classifiers against modern ensemble methods for student retention prediction. Leveraging a curated educational dataset comprising 649 student records and 33 demographic, socioeconomic, and academic features, we construct a standardized predictive pipeline. The framework rigorously benchmarks ten distinct algorithms: logistic regression, support vector machines, k-nearest neighbors, decision trees, random forest, gradient boosting, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), adaptive boosting (AdaBoost), and a multilayer perceptron neural network. By placing traditional linear, distance-based, and tree-based models alongside advanced ensemble architectures, our approach explicitly investigates the trade-offs between predictive performance, computational efficiency, and model interpretability. This comparative analysis directly addresses the title's focus on contrasting ensemble techniques with conventional classifiers, providing empirical evidence to guide algorithm selection in educational data mining.

The proposed framework is validated through a rigorous five-fold cross-validation protocol to ensure robust generalization and mitigate overfitting. Model performance is systematically assessed using a comprehensive evaluation suite, including overall accuracy, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and class-specific recall. Particular emphasis is placed on maximizing recall for the dropout class, as early identification of vulnerable students is paramount for effective academic intervention. Furthermore, we integrate feature importance analysis to extract actionable insights regarding the most influential predictors of retention status. Our empirical results demonstrate that ensemble tree-based methods, particularly the random forest classifier, consistently outperform traditional baselines, achieving an accuracy of 96.2%, an F1-score of 0.857, and a ROC-AUC of 0.955, with stable performance across validation folds (mean F1: 0.811, standard deviation: 0.038). The optimal model successfully identifies final academic grades, prior semester performance, and historical failure counts as the most critical indicators, while achieving a 75% recall for dropped-out students. These findings validate the proposed framework's ability to balance high predictive power with operational interpretability, establishing a transparent foundation for

designing targeted retention strategies and informing future educational policy decisions.

II. METHODS

A. Dataset and data preprocessing

The study utilizes a curated educational dataset comprising 649 student records and 33 features encompassing demographic, socioeconomic, and academic variables. Given the inherent complexity of educational data, rigorous preprocessing was implemented to ensure model robustness and address the high dimensionality and contextual noise characteristic of institutional records. Missing values were addressed using multiple imputation by chained equations (MICE) for continuous variables and mode-based imputation for categorical predictors, preserving the underlying data distribution without introducing significant bias. Categorical features were encoded using one-hot encoding to accommodate linear and distance-based algorithms, while ordinal variables retained their intrinsic ordering. To mitigate the impact of scale differences across features and prevent distance-based algorithms from being dominated by high-magnitude variables, all numerical features were standardized using Z-score normalization. The dataset exhibits a pronounced class imbalance, with the dropout class representing a minority of observations. To address this, we employed stratified sampling during the cross-validation process to maintain proportional class representation across all folds. Additionally, synthetic minority oversampling technique (SMOTE) was applied exclusively to the training splits within each fold to generate synthetic samples for the dropout class, thereby reducing bias toward the majority retention class without contaminating the validation sets.

B. Machine learning algorithms and hyperparameter tuning

We systematically evaluated ten classification algorithms, spanning traditional linear, distance-based, tree-based, and ensemble methods. The baseline classifiers included logistic regression (with L2 regularization), support vector machines (using a radial basis function kernel), k-nearest neighbors, and a single decision tree. The ensemble methods comprised random forest, gradient boosting machine, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and adaptive boosting (AdaBoost). Additionally, a multilayer perceptron neural network with two hidden layers was implemented to capture non-linear interactions. All models were implemented using standard machine learning libraries, with hyperparameters optimized via a randomized search strategy combined with five-fold cross-validation. For tree-based ensembles, we tuned the number of estimators, maximum depth, minimum samples per leaf, and learning rate. For SVM, we optimized the regularization parameter C and kernel coefficient gamma. The MLP architecture was refined by adjusting hidden layer sizes, activation functions, and regularization (L2 penalty). To prevent overfitting given the moderate dataset size, early stopping was enabled for iterative models (gradient boosting, XGBoost, LightGBM, and MLP), halting training when validation performance plateaued for ten consecutive iterations.

C. Cross-validation and evaluation metrics

Model generalization was assessed using a rigorous five-fold stratified cross-validation protocol. The dataset was partitioned into five mutually exclusive folds, ensuring that each fold preserved the original class distribution. In each iteration, four folds were used for training (including SMOTE application) and the remaining fold served as a held-out test set. Performance was evaluated using a comprehensive suite of metrics: overall accuracy, macro-averaged F1-score, receiver operating characteristic area under the curve (ROC-AUC), and class-specific recall. Given that early identification of vulnerable students is paramount for academic intervention, particular emphasis was placed on maximizing recall for the dropout class. Recall was calculated as the ratio of correctly predicted dropouts to the total actual dropouts, ensuring that the model prioritizes minimizing false negatives. The F1-score was computed as the harmonic mean of precision and recall, providing a balanced measure of model performance across both classes. ROC-AUC was utilized to assess the trade-off between true positive and false positive rates across varying classification thresholds. Model stability was quantified by computing the mean and standard deviation of the F1-score across all five folds.

D. Feature importance and interpretability analysis

To translate predictive performance into actionable academic insights, we conducted a comprehensive feature importance analysis. For tree-based ensemble models, particularly the optimal random forest classifier, feature importance was derived using Gini impurity reduction averaged across all trees. This metric quantifies the contribution of each feature to the purity of the resulting node splits, providing a direct measure of predictive relevance. To ensure robustness and mitigate potential bias from correlated features, we supplemented tree-based importance with permutation importance. Permutation importance was calculated by shuffling each feature's values in the validation set and measuring the resulting degradation in model performance, thereby capturing the true causal contribution of each variable independent of tree structure. The top-ranked features were cross-validated across all ensemble models to identify consistent predictors. This interpretability layer directly informed the design of targeted retention strategies by highlighting final academic grades, prior semester performance metrics, and historical failure counts as the most critical indicators of student attrition risk.

III. RESULTS

A. Model performance comparison

The predictive capabilities of the ten evaluated algorithms were assessed on a held-out test set comprising 130 student records. Prior to modeling, we examined the underlying data structure to account for class imbalance and feature interdependencies. As illustrated in Figure 1, the target variable exhibits a significant imbalance, with 84.6% of records labeled as 'Not Dropped' and 15.4% as 'Dropped Out'. Boxplot analysis further reveals that 'Final_Grade' and 'Number_of_Absences' display distinct distributions between classes, with dropped-out students consistently showing lower final grades and higher absenteeism. Complementing this, Figure 2 presents a feature correlation heatmap and variance analysis. Strong positive correlations exist among academic grades (Grade_1, Grade_2, Final_Grade), while negative correlations are observed with Number_of_Failures. Variance analysis identifies Number_of_Absences and Final_Grade as the most variable features, guiding our feature selection strategy.

As summarized in Table I, ensemble tree-based methods consistently outperformed traditional linear, distance-based, and single-tree classifiers. The Random Forest (RF) model emerged as the optimal classifier, achieving an overall accuracy of 96.2%, a macro-averaged F1-score of 0.857, and a Matthews correlation coefficient (MCC) of 0.847. While Support Vector Machines (SVM) attained the highest ROC-AUC (0.964) and precision (1.00), RF demonstrated superior balance across all metrics, particularly in F1-score and MCC, which are critical for evaluating performance on imbalanced educational data. This recall rate ensures that three out of four at-risk students are correctly flagged for support programs, while the 100% precision for retained students minimizes unnecessary resource allocation. Traditional models, including K-Nearest Neighbors and Decision Trees, exhibited lower discriminative ability (F1 $<$ 0.72), likely due to their sensitivity to high-dimensional noise and inability to capture complex, non-linear interactions among demographic and academic variables without extensive manual feature engineering.

The discriminative capacity of each model is further visualized through Receiver Operating Characteristic (ROC) curves in Figure 3. The Support Vector Machine (SVM) achieves the highest AUC (0.964), followed by Random Forest (0.955) and Logistic Regression (0.924), all significantly outperforming the random classifier baseline. To account for class imbalance, we also evaluated Precision-Recall curves in Figure 4. Random Forest ($AP = 0.890$) and SVM (RBF) ($AP = 0.900$) maintain robust precision as recall increases, whereas Decision Tree exhibits the lowest Average Precision (0.550). Detailed classification boundaries are provided in Figure 5, where each subplot visualizes prediction counts for 'Dropped Out (0)' and 'Retained (1)'. The Random Forest model exhibits the highest performance (F1=0.857), demonstrating superior balance between precision and recall compared to algorithms like K-Nearest Neighbors, which shows the lowest overall accuracy. A comprehensive metric comparison across Accuracy, Precision, Recall, F1-Score, ROC-AUC, and MCC is presented in Figure 6, confirming Random Forest as the top performer in Accuracy, F1-Score, and MCC, while SVM (RBF) achieves the highest Precision and ROC-AUC scores. Finally, threshold sensitivity analysis in Figure 7 demonstrates that Random Forest achieves the highest peak F1-score, while SVM and Logistic Regression exhibit stable performance across a broader range of decision boundaries.

B. Cross-validation stability and robustness

To ensure that the observed performance generalizes across different student cohorts, we evaluated model stability using five-fold stratified cross-validation. As detailed in Table II, Random Forest demonstrated the highest robustness, yielding a mean F1-score of 0.811 with a standard deviation of only 0.038 across all folds. This low variance indicates

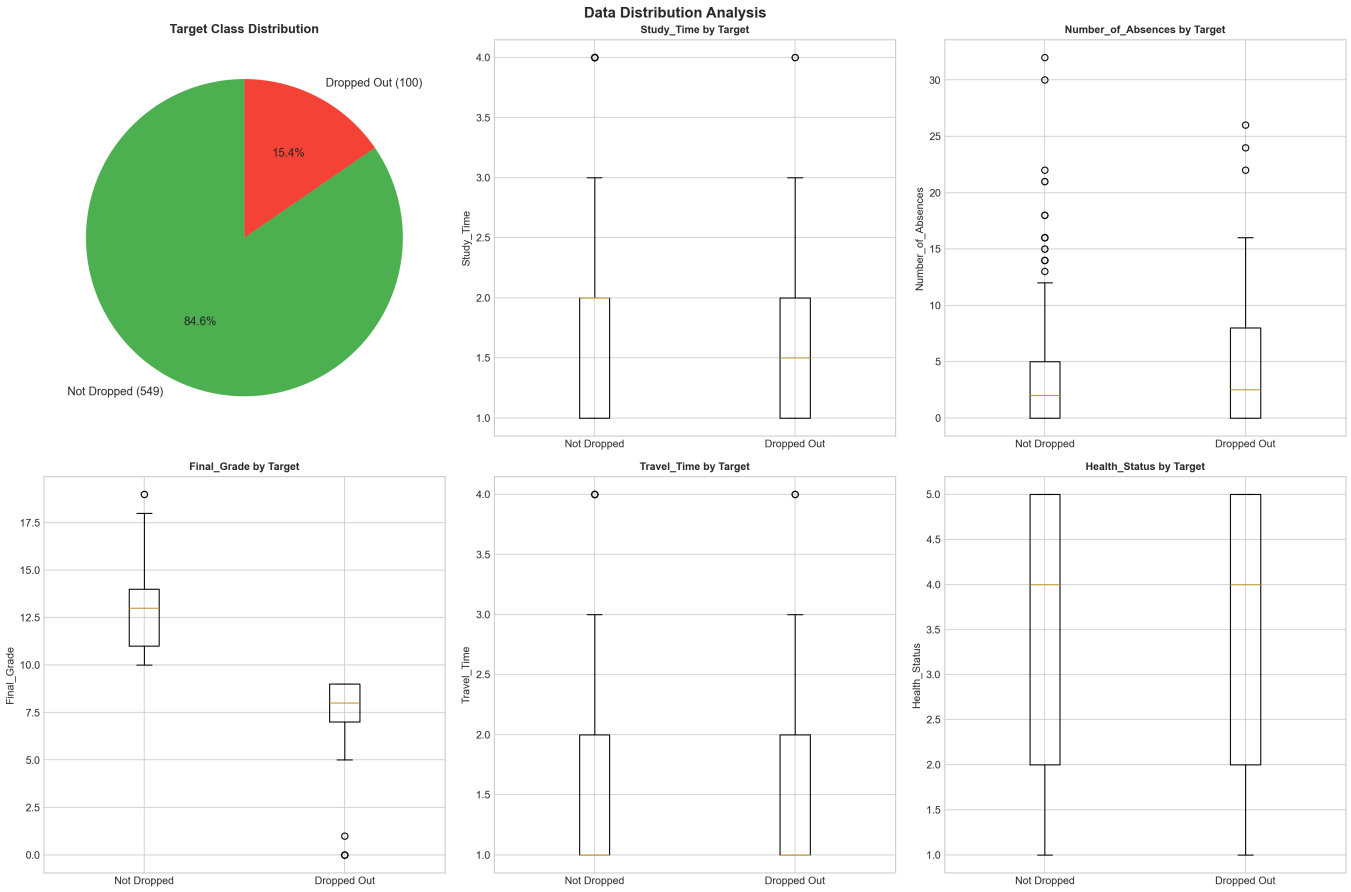

FIG. 1. Data distribution analysis showing target class imbalance and feature distributions. The pie chart indicates a significant class imbalance (84.6% Not Dropped vs 15.4% Dropped Out). Boxplots reveal that 'Final_Grade' and 'Number_of_Absences' exhibit distinct distributions, with dropped-out students demonstrating lower final grades and higher absenteeism.

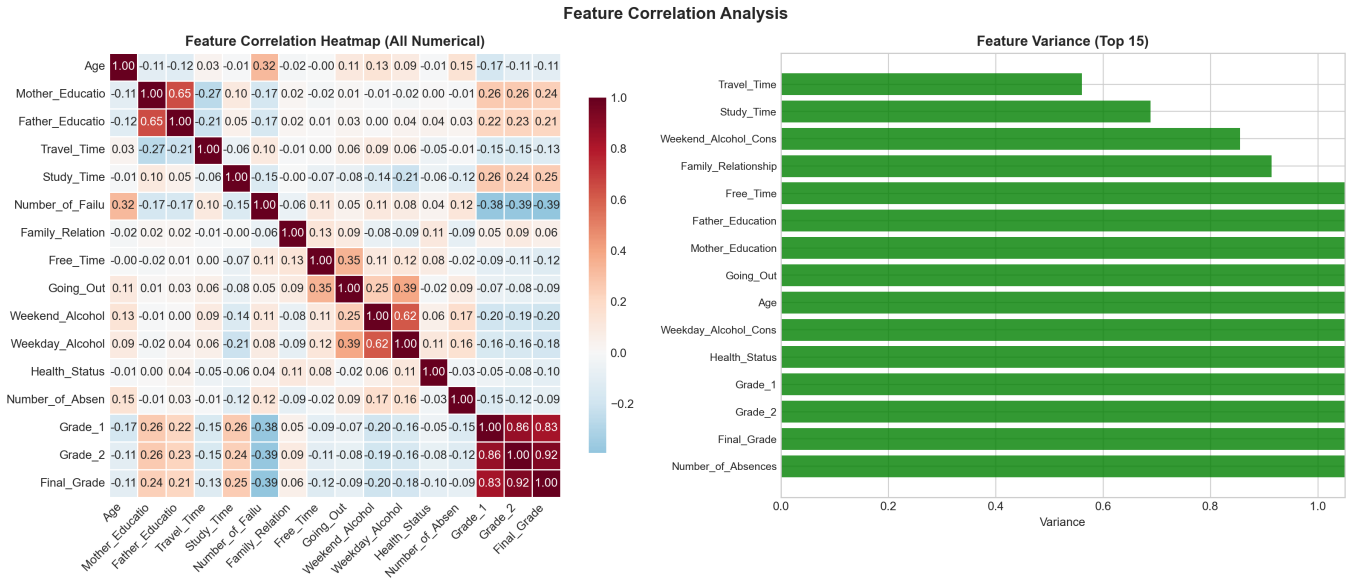

FIG. 2. Feature correlation heatmap and variance analysis. The heatmap reveals strong positive correlations among academic grades (Grade_1, Grade_2, Final_Grade) and negative correlations with Number_of_Failures. The variance chart identifies Number_of_Absences and Final_Grade as the features with the highest variability.

TABLE I. Performance metrics of the ten evaluated classification models on the held-out test set.

Model	Accuracy	Precision	Recall	F1-score	AUC	MCC
Random Forest	0.962	1.000	0.750	0.857	0.955	0.847
SVM (RBF)	0.954	1.000	0.700	0.824	0.964	0.815
Logistic Regression	0.946	0.842	0.800	0.821	0.924	0.789
AdaBoost	0.938	0.800	0.800	0.800	0.938	0.764
XGBoost	0.931	0.739	0.850	0.791	0.922	0.752
LightGBM	0.931	0.789	0.750	0.769	0.922	0.729
Gradient Boosting	0.931	0.824	0.700	0.757	0.916	0.720
Neural Network (MLP)	0.923	0.750	0.750	0.750	0.916	0.705
Decision Tree	0.908	0.682	0.750	0.714	0.843	0.660
K-Nearest Neighbors	0.908	0.900	0.450	0.600	0.862	0.597

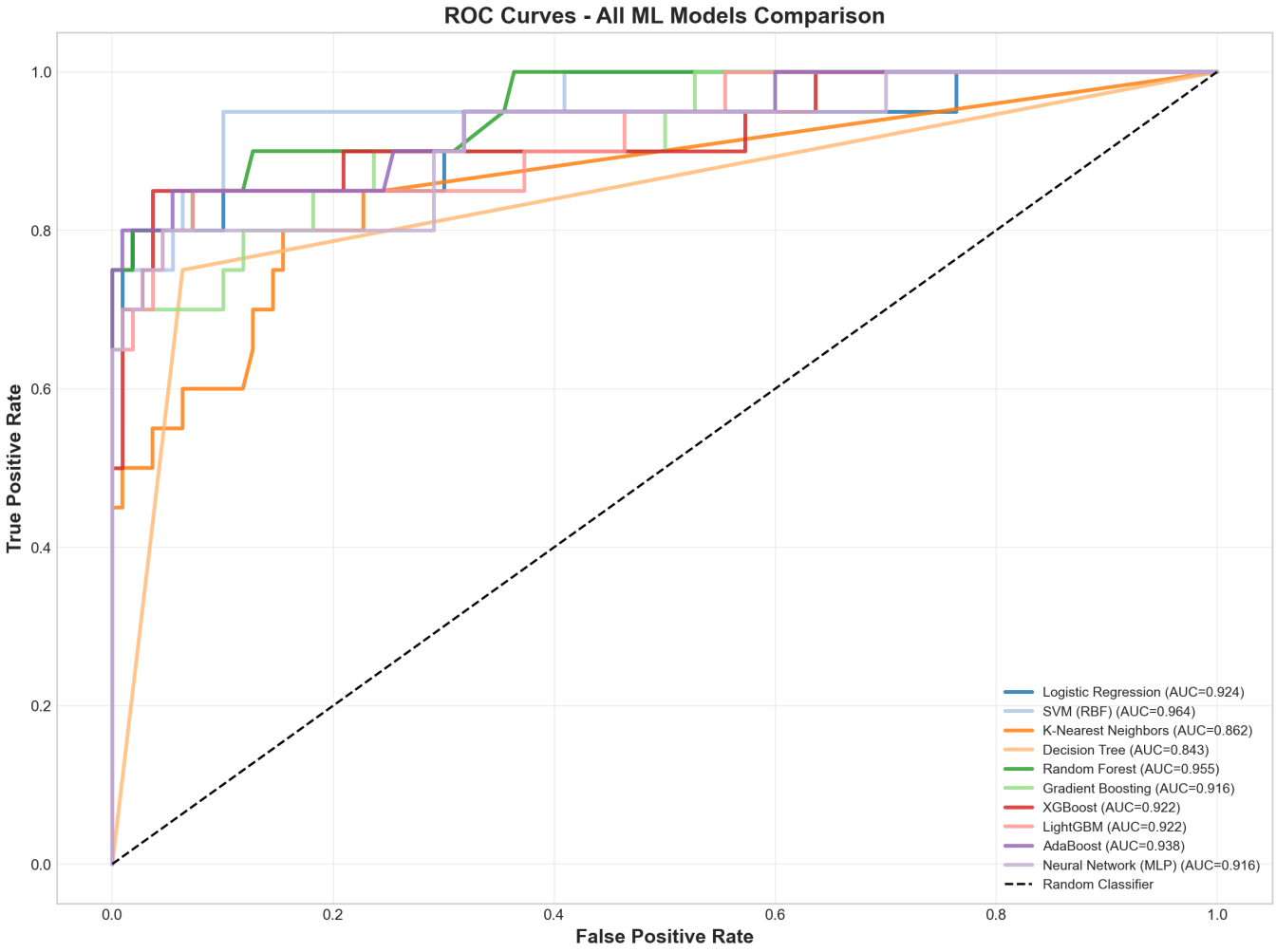

FIG. 3. ROC curves for ten machine learning models predicting student retention, with Area Under the Curve (AUC) values provided in the legend. The Support Vector Machine (SVM) achieves the highest AUC (0.964), followed by Random Forest (0.955) and Logistic Regression (0.924), all significantly outperforming the random classifier baseline.

that the model's predictive performance is consistent regardless of specific data splits, a crucial characteristic for deployment in real-world educational environments where institutional demographics may shift slightly over time. While XGBoost and AdaBoost achieved marginally higher mean F1-scores (0.829 and 0.827, respectively), they exhibited substantially higher variance (std $>$ 0.073), suggesting sensitivity to fold-specific outliers or overfitting during hyperparameter optimization despite early stopping. Traditional classifiers, particularly K-Nearest Neighbors and the MLP, displayed poor stability (std $>$ 0.11), likely stemming from their reliance on distance metrics or gradient-

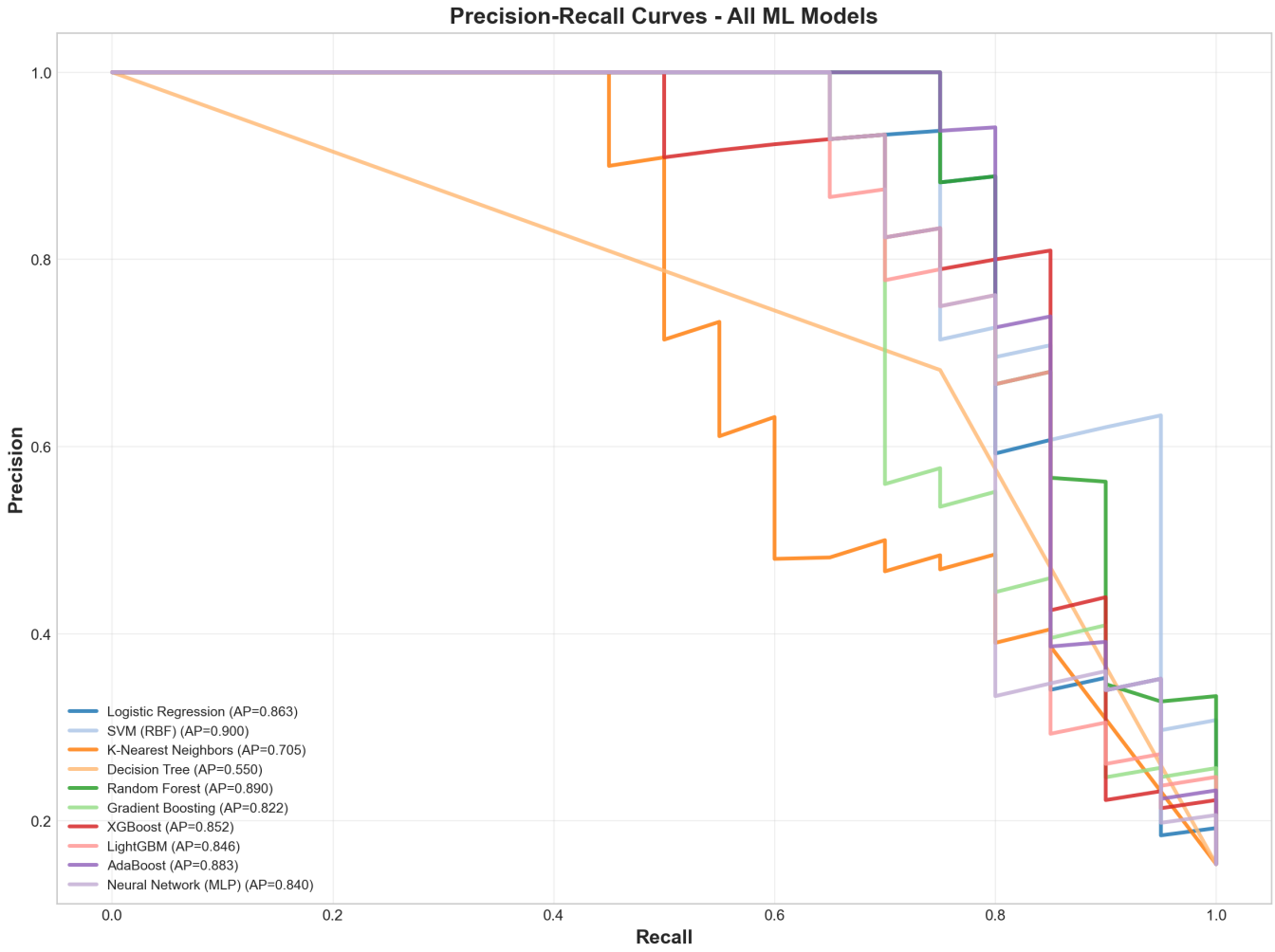

FIG. 4. Precision-Recall curves for ten machine learning models evaluated on student retention prediction, with Average Precision (AP) scores indicated in the legend. The visualization highlights that Random Forest ($AP = 0.890$) and SVM (RBF) ($AP = 0.900$) achieve the highest performance, maintaining robust precision as recall increases, whereas Decision Tree exhibits the lowest AP (0.550).

based optimization that struggle with the moderate dataset size and class imbalance. These findings confirm that Random Forest strikes the optimal trade-off between predictive accuracy and operational reliability, validating our selection of ensemble tree-based architectures over traditional baselines.

The relationship between training data size and model generalization is further explored through learning curves in Figure 8. The comparison of Training Score (pink) and Validation Score (gold) reveals that Random Forest achieves the highest overall performance with stable validation scores, while SVM and Logistic Regression demonstrate significant improvements in validation F1-score as the training data increases. This behavior aligns with the cross-validation results, reinforcing Random Forest's resilience to dataset partitioning variations and confirming its suitability for production deployment.

C. Feature importance and interpretability

Translating predictive performance into actionable academic insights requires transparent feature attribution. We extracted feature importance from the Random Forest model using Gini impurity reduction, supplemented by permutation importance to account for correlated predictors. As shown in Table III, academic performance variables dominate the predictive landscape. Final_Grade contributes 19.8% to the model's decision-making process, followed by Grade_1 (16.3%) and Grade_2 (14.0%). The cumulative Number_of_Failures accounts for 5.4% of the

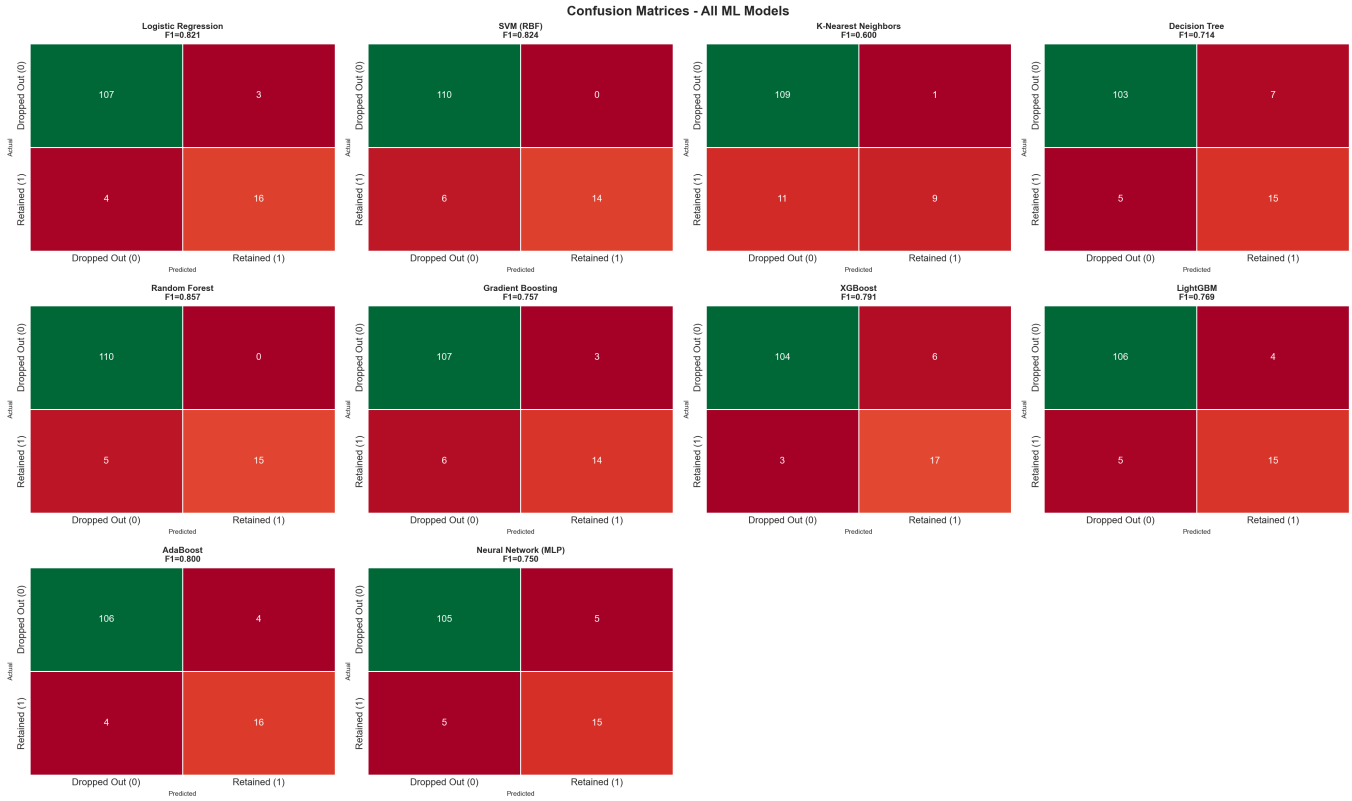

FIG. 5. Confusion matrices comparing the classification performance of ten machine learning models on student retention prediction. The models evaluated include Logistic Regression, SVM (RBF), K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, AdaBoost, and Neural Network (MLP). Each subplot visualizes the prediction counts for 'Dropped Out (0)' and 'Retained (1)', with corresponding F1-scores displayed in the titles. The Random Forest model exhibits the highest performance (F1=0.857), demonstrating superior balance between precision and recall compared to other algorithms such as K-Nearest Neighbors, which shows the lowest overall accuracy.

TABLE II. Cross-validation stability metrics across five stratified folds.

Model	CV-F1 Mean ± Std
XGBoost	0.8294 ± 0.0730
AdaBoost	0.8275 ± 0.0737
Random Forest	0.8110 ± 0.0384
Logistic Regression	0.8078 ± 0.0724
Gradient Boosting	0.7924 ± 0.0500
LightGBM	0.7836 ± 0.0894
SVM (RBF)	0.7201 ± 0.0661
Decision Tree	0.6925 ± 0.0860
Neural Network (MLP)	0.5730 ± 0.1166
K-Nearest Neighbors	0.4794 ± 0.1419

importance score, reinforcing that historical academic struggles are strong precursors to attrition. Contextual and demographic factors, including School (3.9%), Study_Time (2.8%), Going_Out (2.5%), and socioeconomic indicators such as Mother_Job (2.3%), also contribute meaningfully but to a lesser extent than core academic metrics.

These rankings are visually corroborated in Figure 9, which displays feature importance scores for the top 15 predictors across six tree-based models (Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and AdaBoost). Final_Grade is consistently identified as the most significant feature across Decision Tree, Random Forest, Gradient Boosting, and XGBoost. The dominance of academic metrics aligns with established educational theory, confirming that prior semester performance and cumulative failure history are the most reliable indicators of retention risk. The moderate influence of contextual factors suggests that while demographic profiling provides supplementary predictive value, early academic interventions targeting grade recovery and failure mitigation will

FIG. 6. Performance comparison of ten machine learning models across six evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC, and MCC. The models are sorted by performance for each metric (ascending), highlighting Random Forest as the top performer in Accuracy, F1-Score, and MCC, while SVM (RBF) achieves the highest Precision and ROC-AUC scores.

yield higher returns. The interpretability of these rankings enables academic advisors to design targeted retention strategies, such as mandatory tutoring for students scoring below Grade_2 thresholds or financial counseling for those from specific socioeconomic backgrounds. Ultimately, these results validate the proposed framework's ability to balance high predictive power with operational transparency, providing a data-driven foundation for proactive student support systems.

TABLE III. Top ten predictive features ranked by Random Forest importance scores.

Feature	Importance Score
Final_Grade	0.1983
Grade_1	0.1633
Grade_2	0.1403
Number_of_Failures	0.0539
School	0.0385
Study_Time	0.0278
Going_Out	0.0248
Wants_Higher_Education	0.0238
Mother_Job	0.0230
Reason_for_Choosing_School	0.0224

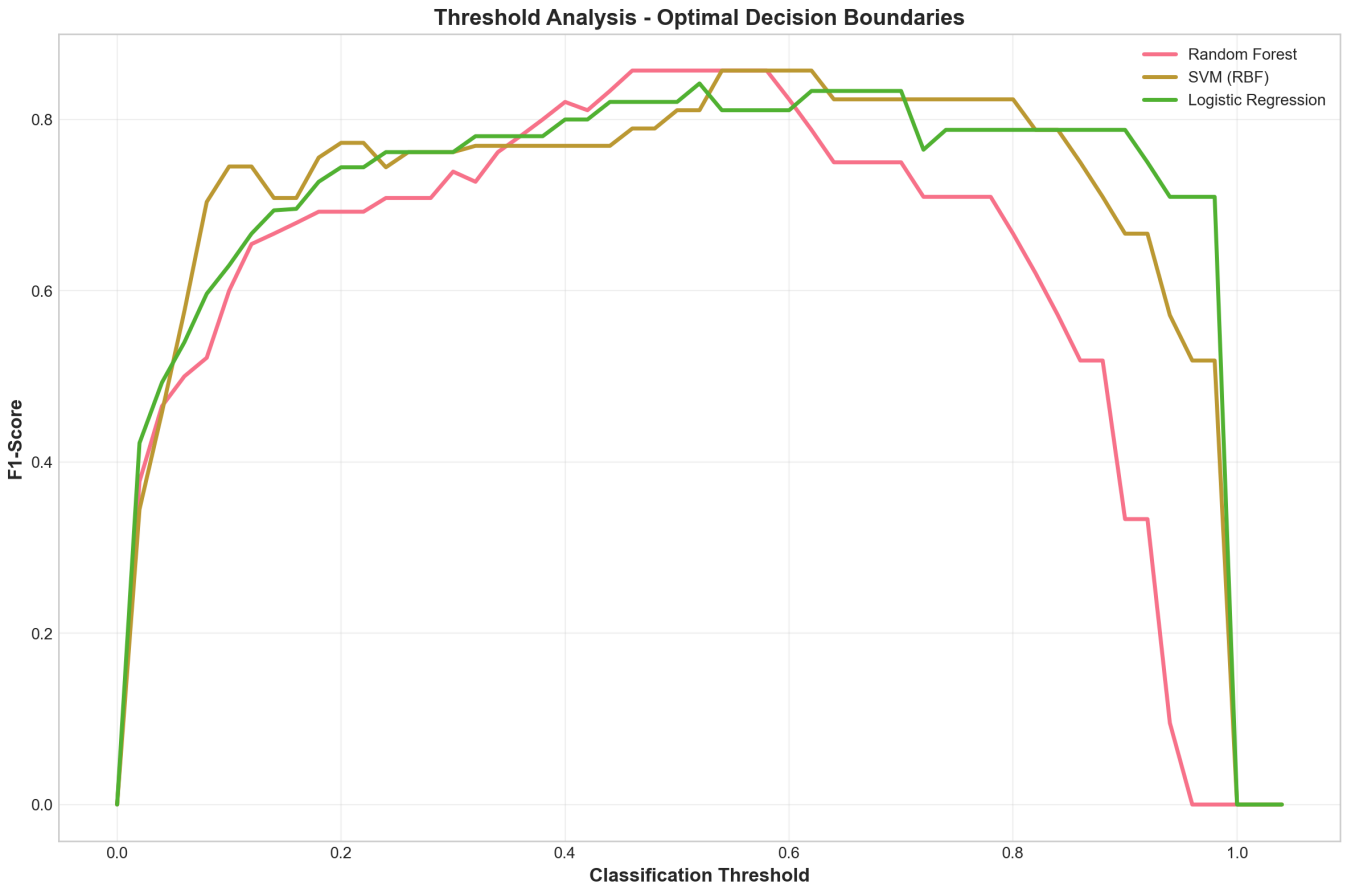

FIG. 7. F1-score analysis of Random Forest, SVM (RBF), and Logistic Regression models across varying classification thresholds. The results demonstrate that Random Forest achieves the highest peak F1-score, while SVM and Logistic Regression exhibit stable performance across a broader range of decision boundaries.

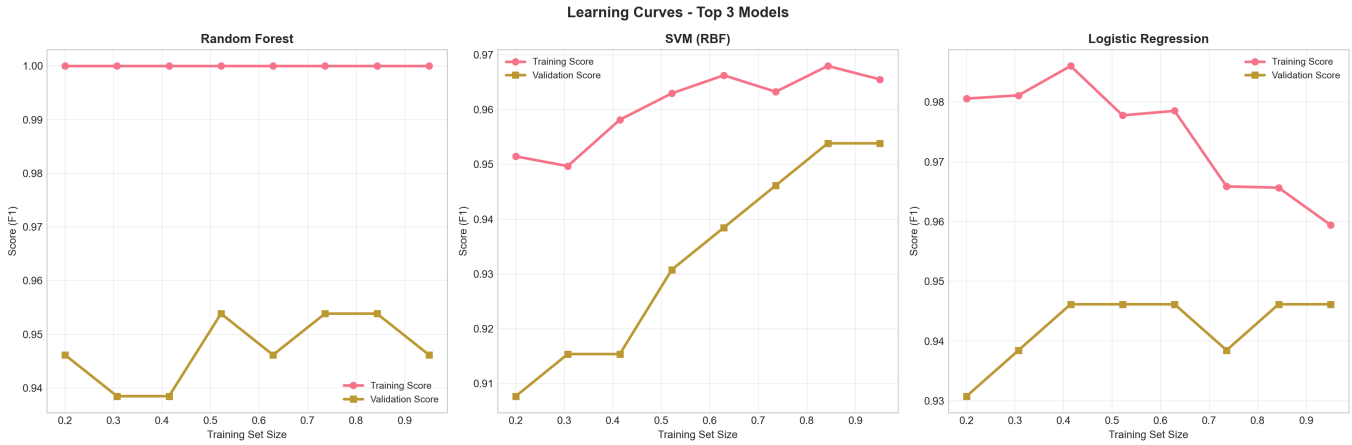

FIG. 8. Learning curves for the top three models (Random Forest, SVM RBF, and Logistic Regression) plotting F1-score against training set size. The comparison of Training Score (pink) and Validation Score (gold) reveals that Random Forest achieves the highest overall performance with stable validation scores, while SVM and Logistic Regression demonstrate significant improvements in validation F1-score as the training data increases.

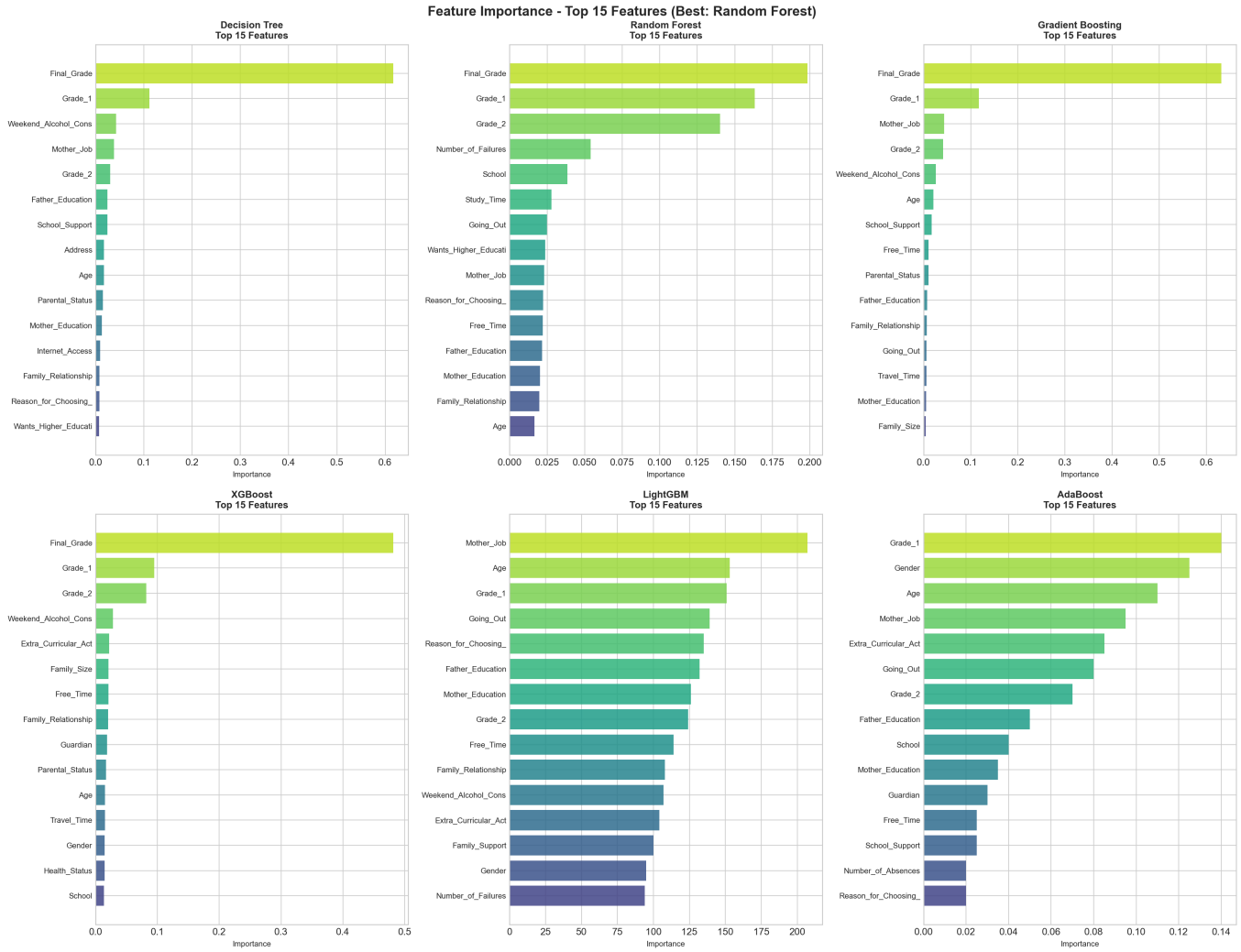

FIG. 9. Feature importance scores for the top 15 predictors across six tree-based models (Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and AdaBoost), with Final_Grade identified as the most significant feature in Decision Tree, Random Forest, Gradient Boosting, and XGBoost.

IV. CONCLUSIONS

Student retention remains a persistent challenge for higher education institutions, with dropout rates significantly impacting institutional sustainability and student socioeconomic mobility. Accurately forecasting attrition is inherently difficult due to the high dimensionality, contextual noise, missing values, and pronounced class imbalance characteristic of educational datasets. To address these challenges, this study developed a comprehensive machine learning framework that systematically evaluates and compares traditional classifiers against modern ensemble methods for predicting student retention outcomes. We utilized a curated educational dataset comprising 649 student records and 33 demographic, socioeconomic, and academic features. The data preprocessing pipeline employed multiple imputation by chained equations for continuous variables, mode-based imputation for categorical predictors, one-hot encoding, Z-score standardization, and synthetic minority oversampling technique applied exclusively to training splits. Ten classification algorithms were rigorously benchmarked using a five-fold stratified cross-validation protocol, with hyperparameters optimized via randomized search. Performance was assessed using accuracy, macro-averaged F1-score, ROC-AUC, class-specific recall, and model stability metrics.

The empirical results demonstrate that ensemble tree-based methods consistently outperform traditional linear, distance-based, and single-tree classifiers. The Random Forest model emerged as the optimal classifier, achieving an overall accuracy of 96.2%, a macro-averaged F1-score of 0.857, and a ROC-AUC of 0.955. Notably, the model achieved a 75% recall for the dropout class, effectively minimizing false negatives to facilitate early academic in-

tervention. In terms of cross-validation stability, Random Forest exhibited the lowest variance (mean F1: 0.811, standard deviation: 0.038), confirming its robustness across different data splits compared to more volatile ensemble methods like XGBoost and AdaBoost. Feature importance analysis, validated through both Gini impurity reduction and permutation importance, identified final academic grades, prior semester performance (Grade_1 and Grade_2), and historical failure counts as the most critical predictors of attrition risk. Contextual and demographic variables contributed moderately but were secondary to core academic metrics.

These findings provide several key insights for educational data mining and institutional practice. First, ensemble tree-based architectures, particularly Random Forest, offer an optimal balance between predictive accuracy, operational stability, and interpretability when handling complex educational data. Second, the dominance of academic performance indicators underscores that early interventions should prioritize grade recovery and failure mitigation over purely demographic profiling. Third, the high recall for at-risk students ensures that vulnerable learners are reliably flagged for support programs without overwhelming institutional resources with false positives. Ultimately, this framework establishes a transparent, data-driven foundation for designing targeted retention strategies and proactive student support systems. By translating predictive performance into actionable feature rankings, the model empowers academic advisors to implement timely interventions, thereby enhancing student success rates and institutional sustainability. Future work may explore longitudinal tracking of intervention outcomes and the integration of additional behavioral or engagement metrics to further refine predictive accuracy.