EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems

EAI.EU Research Article

Artificial Intelligence-Driven Early Prediction of Student Dropout and Academic Outcomes in Higher Education: A Comparative Study of Advanced Machine Learning Approaches

Nghia Trong Vo[1], Quang Nhat Le[1], Hang Le[2,*]

[1]Memorial University, St. John's, NL A1B3X5, Canada
[2]Duy Tan University, Da Nang, Vietnam

Abstract

Student dropout in higher education remains a critical challenge with significant academic, social, and economic implications. Early identification of students at risk of dropout enables institutions to design timely and targeted interventions that support academic success and improve retention rates. This study proposes a machine learning (ML)–driven framework for the early prediction of student dropout and academic outcomes in higher education using a comprehensive, real-world dataset collected from a higher education institution. The prediction task is formulated as a multiclass classification problem with three outcomes: dropout, enrolled, and graduate. To evaluate the effectiveness of different modeling approaches, we conduct a comparative analysis of widely used ML algorithms, including Logistic Regression, Naïve Bayes, k-Nearest Neighbors, Support Vector Machine, Decision Trees, Random Forest (RF), AdaBoost, XGBoost, LightGBM, and CatBoost. Results indicate that ensemble models achieve the best performance. RF attains the highest test accuracy (0.7797) and ROC-AUC (OvR) (0.8919), while LightGBM yields the best Macro-F1 (0.7082). Feature importance analysis shows that early academic progress indicators (approved units and semester grades) are the strongest predictors, followed by selected administrative/contextual factors such as tuition-fee status and course. Overall, this study provides empirical evidence supporting the use of ML techniques as effective decision-support tools for higher education institutions. The proposed framework offers actionable insights for administrators and policymakers seeking to develop data-driven strategies aimed at reducing dropout rates, improving academic success, and promoting equitable access to educational opportunities.

Received on 29 January 2026; accepted on 17 February 2026; published on 03 March 2026

Keywords: Dropout prediction, machine learning, student performance prediction

doi:10.4108/eetinis.131.11758

1. Introduction

The digital transformation of higher education has led to the widespread adoption of learning management systems, online assessment platforms, and virtual learning environments. These systems continuously generate large volumes of educational data, including attendance records, assessment results, and detailed logs of student interactions. Leveraging such data to better understand and support student learning has become a central focus of educational data mining and learning analytics [1]. In particular, the ability to accurately predict students' academic performance at an early stage is increasingly recognized as a critical tool for improving educational outcomes.

Early identification of students at risk of poor academic performance enables instructors and institutions to provide timely interventions such as personalized feedback, academic advising, or targeted support programs [2]. Traditional performance evaluation methods,

[*]Corresponding author. E-mail address: hangle@duytan.edu.vn

which rely heavily on final examinations or cumulative grades, often detect learning difficulties too late to allow effectively corrective and timely actions. As a result, predictive models that can forecast student performance during the early phases of a course have gained significant attention in recent years [3].

Machine learning (ML) techniques have demonstrated strong potential for modeling the complex and nonlinear relationships inherent in educational data [4, 5]. By analyzing historical academic records together with behavioral indicators extracted from online learning activities, ML-based approaches can uncover hidden patterns that are difficult to capture using conventional statistical methods [6]. In particular, fine-grained online activity logs such as login frequency, content access behavior, assignment submission patterns, and interaction intensity provide valuable insights into students' engagement and learning habits, which are closely linked to academic success [7].

Despite the promising results reported in existing studies, accurately predicting student performance remains a challenging task [8]. Educational datasets are often heterogeneous, noisy, imbalanced, and student learning behavior may vary significantly across courses, institutions, and time periods. Single predictive models may therefore suffer from limited generalization capability or sensitivity to specific data characteristics. To address these challenges, ensemble and multi-model learning strategies have been increasingly explored. By combining the strengths of multiple base learners, ensemble approaches can improve prediction robustness, reduce variance, and enhance overall accuracy compared to individual models [9–11].

Motivated by these considerations, recent research has focused on integrating early-stage academic indicators, online behavioral features, and ensemble ML techniques to develop more reliable student performance prediction systems [12]. Such approaches aim not only to improve predictive accuracy but also to provide practical decision-support tools for educators and academic administrators. By enabling early and data-driven identification of at-risk students, these systems have the potential to contribute to improved retention rates, better learning experiences, and more effective educational strategies in primary [13], secondary [14], and higher education [15–20].

1.1. Related Works

Building on the growing interest in early prediction of student academic performance, a substantial body of research has explored the use of educational data mining and learning analytics to model students' learning behaviors and outcomes. The work in [9] developed an ensemble ML model that leverages multiple classifiers to recommend potential students by accurately predicting their academic suitability and performance. Numerical results demonstrated that the ensemble approach outperforms individual models, which provides more reliable and effective decision support for student recommendation and academic planning. In [12], the authors proposed an early prediction framework for student academic performance in higher education by combining admission criteria and first-year course results with ML and t-SNE–based dimensionality reduction. Experimental results showed that integrating early academic features significantly improves GPA prediction accuracy, thus enabling institutions to identify at-risk students and intervene at early stages. In [20], ML models were used to predict students' academic outcomes, i.e., on-time graduation, delayed completion, or dropout, at different phases of the first academic year using demographic, socioeconomic, macroeconomic, and academic data. Results showed that random forest (RF)–based models perform the best, i.e., with the most accurate predictions achieved by the end of the first semester. This enables earlier identification of at-risk students for targeted interventions. In [21], a comprehensive evaluation of ML techniques for estimating student academic performance using diverse educational data was presented. The study compared multiple models and metrics to identify effective approaches that can support early prediction and data-driven decision-making in education. A web-based ML system was proposed in [22] to predict university students' academic performance at early stages using academic and demographic data. By comparing multiple algorithms, the study showed that academic factors, especially midterm exam scores, have the strongest impact on performance, and the proposed system can support early identification of at-risk students and informed academic intervention. In [23], the authors devised a ML–based academic advising framework that uses real-world engineering student data to predict academic performance and identify at-risk students early. Experimental results showed that the proposed models achieve strong predictive accuracy and can support advisors in making timely and personalized interventions to improve student success and retention. The paper in [24] investigated different ML algorithms, i.e., Support Vector Machine (SVM), Decision Tree (DT), RF, and k-Nearest Neighbors (kNN), to predict and compare students' academic performance in online and offline learning environments using real datasets. The results showed that tree-based models, particularly DT and RF, achieve high predictive accuracy. The authors also revealed that students' habits, e.g., study time, sleep, and screen exposure, are key factors that influence academic success in both learning modes. In [25], a fuzzy propositional model (FPM) was proposed to integrate fuzzy set theory with propositional logic to reason and predict students' academic performance under uncertainty, thus addressing limitations of traditional deterministic and statistical models. Experiments on real university datasets showed that FPM achieves more accurate and interpretable predictions than linear regression, particularly in scenarios with weak or imprecise relationships such as absenteeism and exam performance.

1.2. Problem Statement and Research Objectives

Despite the advances of learning analytics and educational data mining, building reliable predictive systems for student outcomes remains challenging due to heterogeneous student profiles, class imbalance, and the complex interplay between socioeconomic context and academic performance [8]. In this work, we study student outcome prediction using a real-world higher education dataset that integrates information available at enrollment, e.g., demographic, application, and socioeconomic attributes, together with aggregated academic performance indicators from the first and second semesters, and regional macroeconomic indicators.

We address the task of predicting each student's academic status at the end of the normal course duration as a *three-class classification* problem with outcome categories *Dropout*, *Enrolled*, and *Graduate*. In contrast to approaches that rely primarily on fine-grained learning management system (LMS) interaction logs, our focus is on institutional variables that are commonly available in administrative and academic information systems, making the proposed analysis relevant for a wide range of higher education settings.

The main research objectives of this study are as follows:

- **O1: Performance benchmarking.** Evaluate and compare widely used ML classifiers for predicting student outcomes in a consistent experimental setting.
- **O2: Robust multi-class assessment.** Report performance using metrics appropriate for imbalanced multi-class prediction, e.g., Macro-F1 in addition to overall accuracy, and analyze typical misclassification patterns.
- **O3: Insight into predictive factors.** Identify the most influential predictors of dropout and academic success to support early risk identification and data-driven student support strategies.

1.3. Contributions and Paper Organization

This paper makes the following contributions:

- We present a comprehensive empirical study of student outcome prediction using a higher education dataset that combines demographic, socioeconomic/administrative, academic performance, and macroeconomic variables.
- We benchmark a diverse set of common classification models, including Logistic Regression (LR), Naïve Bayes (NB), kNN, DT, RF, AdaBoost, SVM, and gradient boosting methods, i.e., XGBoost, LightGBM, and CatBoost, and compare them using consistent evaluation metrics.
- We provide an interpretability-oriented analysis based on feature importance to highlight key factors associated with student dropout and academic success, offering insights that may inform institutional retention strategies.

The remainder of the paper is organized as follows. Section 2 describes the dataset and provides the formal problem formulation. Section 3 presents the ML models and preprocessing pipeline. Section 4 details the experimental setup and evaluation metrics. Section 5 reports and discusses the results, including model comparison and feature importance analysis. Finally, Section 6 concludes the paper.

2. Dataset and Problem Formulation

2.1. Dataset Description

We use the publicly available *Predict students' dropout and academic success* dataset released on Zenodo [26]. The dataset was constructed from a higher education institution by integrating information from multiple disjoint administrative databases. It includes variables available at the time of enrollment, e.g., demographic and socioeconomic characteristics, application-related information, and students' academic performance summaries at the end of the first and second semesters. The goal is to support the development of predictive models for student outcomes at the end of the normal duration of the degree program.

After loading the dataset, we obtain $N = 4424$ student records, each described by $d = 34$ predictor variables and one categorical outcome label. No missing values are observed in the provided data.

2.2. Features

The dataset predictors capture complementary aspects of the student profile and learning trajectory. For clarity, we group the features into four categories:

- **Demographic and enrollment context:** e.g., marital status, gender, age at enrollment, nationality, course, daytime/evening attendance, application mode and order.

- **Socioeconomic and administrative factors:** e.g., scholarship holder status, debtor status, tuition fees up to date, displaced status, parents' qualifications and occupations, and educational special needs.
- **Early academic performance indicators:** aggregated curricular-unit information for the first and second semesters, including the number of units credited/enrolled/evaluated/approved and the corresponding grades.
- **Macroeconomic indicators:** unemployment rate, inflation rate, and GDP of the region in the relevant period.

Most categorical features are provided as integer-coded categories, consistent with the dataset documentation. Continuous features include semester grades and macroeconomic indicators. In our experiments, we apply feature scaling to ensure comparability across models that are sensitive to feature magnitudes, e.g., LR, kNN, and SVM.

2.3. Outcome Definition and Class Distribution

The target variable (*Target*) is defined with three outcome categories: *Dropout*, *Enrolled*, and *Graduate*. In our dataset, the class distribution is: *Graduate* (2209; 49.93%), *Dropout* (1421; 32.12%), and *Enrolled* (794; 17.95%), indicating a moderately imbalanced multi-class setting.

2.4. Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector describing student i, and $y_i \in \mathcal{Y}$ is the corresponding outcome label. The label space is $\mathcal{Y} = \{Dropout, Enrolled, Graduate\}$. Our objective is to learn a classifier $f : \mathbb{R}^d \rightarrow \mathcal{Y}$ that predicts the outcome of each student based on enrollment, socioeconomic, academic performance, and macroeconomic indicators:

$$\hat{y}_i = f(\mathbf{x}_i), \quad y_i \in \mathcal{Y}. \tag{1}$$

This formulation defines a three-class supervised classification problem. In subsequent sections, we evaluate multiple ML algorithms under a unified experimental protocol and report performance using metrics appropriate for imbalanced multi-class classification, e.g., Macro-F1 in addition to overall accuracy.

3. Methods

This section describes the proposed ML pipeline for predicting student outcomes. Figure 1 summarizes the end-to-end workflow, including preprocessing, model selection, training, classification, and evaluation.

3.1. Workflow Overview

As shown in Figure 1, the proposed system starts from the curated student dataset (Section 2) and applies a uniform preprocessing pipeline. We then train a diverse set of supervised classifiers, each producing a predicted class label in $\{Dropout, Enrolled, Graduate\}$. Finally, we evaluate predictive performance using complementary metrics and diagnostic plots, e.g., confusion matrices and ROC curves, and we analyze feature relevance using model-based importance scores.

3.2. Data Preprocessing

Let $\mathbf{x}_i$ denote the input features for student i and y_i denote the corresponding outcome label. Prior to model training, the following preprocessing steps are applied:

- **Target encoding:** The categorical target labels (*Dropout*, *Enrolled*, *Graduate*) are mapped to numeric class indices for model training.
- **Feature preparation:** Predictor variables include demographic and enrollment attributes, socioeconomic and administrative indicators, academic performance summaries (first and second semesters), and regional macroeconomic indicators. Many categorical variables are provided as integer-coded categories in the dataset; we use these codes as provided to maintain consistency across all compared models.
- **Feature scaling:** We standardize input variables using z-score normalization (StandardScaler), i.e., each feature is transformed to have approximately zero mean and unit variance based on the training data. This step is important for magnitude-sensitive models such as LR, kNN, and SVM, and it enables a consistent feature representation across all classifiers.

3.3. Classification Models

To provide a comprehensive benchmark, we evaluate a set of widely used classification algorithms commonly adopted in educational data mining and applied ML. The model set, as shown in Figure 1, includes:

- **Linear model:** LR, used as a strong and interpretable baseline.
- **Probabilistic baseline:** NB, representing a simple generative approach.
- **Distance-based method:** kNN, capturing local similarity in feature space.
- **Kernel method:** SVM with an RBF kernel, enabling nonlinear decision boundaries.

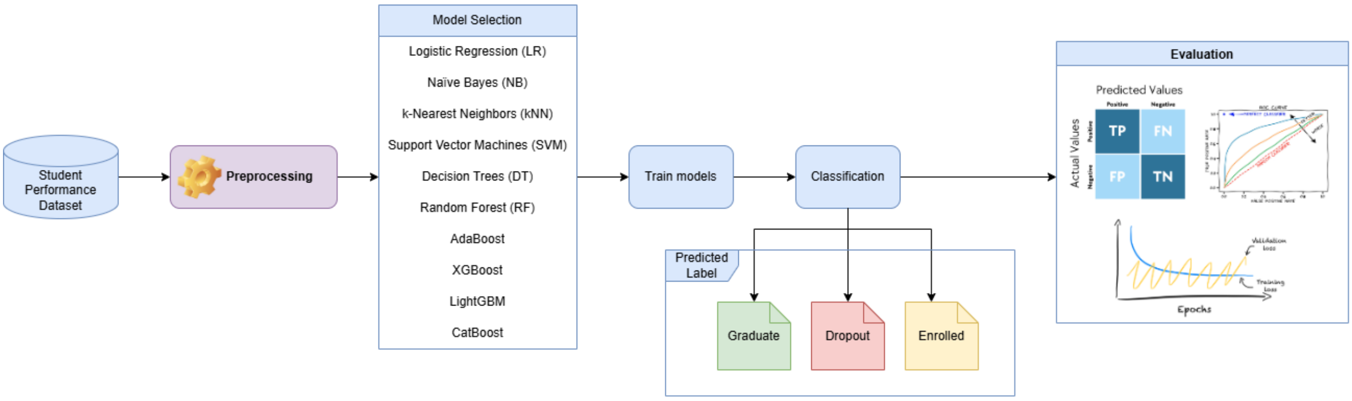

Figure 1. Overview of the proposed model-training workflow: the student dataset is preprocessed, multiple classification models are trained, predictions are produced for the three outcome classes (*Dropout*, *Enrolled*, *Graduate*), and performance is evaluated using standard classification metrics

- **Tree-based models:** DT and RF, which handle nonlinear interactions and provide model-based feature importance.
- **Boosting-based ensembles:** AdaBoost, XGBoost, LightGBM, and CatBoost, which often achieve strong predictive performance by combining multiple weak learners into an ensemble.

This selection spans multiple learning paradigms (linear, probabilistic, distance-based, kernel-based, tree-based, and boosting-based), ensuring that performance comparisons are not biased toward a single model family.

4. Experimental Setup and Evaluation Metrics

This section describes the experimental protocol used to compare all models and the evaluation metrics reported in the results.

4.1. Experimental Setup

All experiments are implemented in Python using standard ML libraries. Specifically, classical models and the preprocessing pipeline are implemented with scikit-learn, while gradient boosting models are trained using their corresponding toolkits, i.e., XGBoost, LightGBM, and CatBoost. To ensure a fair comparison, all classifiers are trained and evaluated on the same preprocessed feature matrix.

For evaluation, we adopt a stratified hold-out split, where 80% of the instances are used for training and the remaining 20% are reserved for testing, preserving the class proportions across *Dropout*, *Enrolled*, and *Graduate*. In addition, to assess robustness to data partitioning, we report stratified k-fold cross-validation (CV) results with $k = 5$, summarizing performance using the mean and standard deviation of accuracy across folds.

Reproducibility is ensured by using a fixed random seed for data splitting and for stochastic learning algorithms where applicable. Unless otherwise stated, models are trained with standard settings to provide a consistent baseline comparison, e.g., ensemble models use a fixed number of estimators. For metrics that require probabilistic outputs, we use predicted class probabilities produced by each classifier; for SVM, probability estimation is enabled to support ROC-AUC computation.

4.2. Evaluation Metrics

Let $C = 3$ be the number of classes, and let M denote the number of instances in the test set. Let y_i and $\hat{y}_i$ denote the true and predicted labels of instance i, respectively.

Accuracy.

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(\hat{y}_i = y_i), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Macro-averaged Precision, Recall, and F1-score. To account for class imbalance, we report macro-averaged metrics, which weight each class equally. For class c, we define TP_c, FP_c, and FN_c as the number of true positives, false positives, and false negatives, respectively:

$$\text{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad \text{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{3}$$

$$\mathrm{F1}_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{4}$$

Macro-averaged scores are computed as

$$\text{Precision}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{Precision}_c, \tag{5}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{Recall}_c, \tag{6}$$

$$\text{F1}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c. \tag{7}$$

ROC-AUC (One-vs-Rest). We report multiclass ROC-AUC using the One-vs-Rest (OvR) strategy. For each class c, we treat c as the positive class and all other classes as negative, we compute AUC_c using predicted probabilities, and then macro-average across classes:

$$\text{ROC-AUC}_{\text{OvR}} = \frac{1}{C}\sum_{c=1}^{C}\text{AUC}_c. \tag{8}$$

Confusion matrix. We analyze class-wise errors using confusion matrices, where entry (r, c) counts the number of instances with true class r predicted as class c.

Reporting. For each model, we report hold-out test performance (Accuracy, Macro Precision/Recall/F1, and ROC-AUC OvR) and CV accuracy (mean ± standard deviation).

5. Results

This section reports the experimental results obtained using the protocol and metrics defined in Section 4. We first compare overall predictive performance across all models, then analyze error patterns and discriminative ability, and finally examine the most influential predictors.

5.1. Overall Model Performance

Table 1 summarizes the test-set performance (Accuracy, Macro Precision/Recall/F1, and ROC-AUC OvR) together with CV accuracy. Overall, ensemble methods achieve the strongest performance. RF yields the best test accuracy (0.7797) and the highest ROC-AUC (0.8919). Gradient boosting methods (LightGBM and CatBoost) perform comparably, with LightGBM achieving the best Macro-F1 (0.7082) among the evaluated models. SVM and XGBoost also show competitive performance, while simpler baselines such as NB and a single DT achieve lower scores.

Beyond the ranking itself, two observations are noteworthy. First, the difference between the top models is relatively small, e.g., RF vs. LightGBM/CatBoost, suggesting that the dataset contains strong predictive signals that can be captured by multiple high-capacity learners. Second, the gap between test accuracy and CV accuracy is modest for most models, and the CV standard deviations are generally low, indicating that performance is reasonably stable under different data partitions. This is important in educational settings, where predictive systems should not be overly sensitive to small variations in training data.

The comparison between accuracy and Macro-F1 also provides insight into class-wise behavior. While RF achieves the highest overall accuracy, LightGBM attains the best Macro-F1, which indicates a more balanced performance across classes (Macro-F1 weights each class equally). This distinction matters because the outcome classes are not equally represented, and a model can achieve high accuracy by favoring majority classes. The stronger Macro-F1 scores of the top ensemble models therefore suggest improved recognition of underrepresented or more ambiguous outcomes.

Model family characteristics help explain these trends. Single DT can capture nonlinear relationships but often suffer from high variance, whereas RF reduces variance through bagging and feature subsampling. Boosting methods (AdaBoost, XGBoost, LightGBM, and CatBoost) further improve performance by iteratively correcting errors and modeling complex interactions, which is well-suited to heterogeneous educational features. In contrast, NB relies on conditional independence assumptions that are unlikely to hold in this dataset, e.g., strong dependencies among academic performance variables, which can limit its discriminative ability. The competitive results of LR suggest that part of the predictive structure may be captured by relatively simple decision boundaries once informative early academic indicators are included, even though nonlinear ensemble models still provide the best overall performance.

5.2. Error Analysis via Confusion Matrices

To better understand misclassification patterns, we visualize confusion matrices for all models in Figure 2. This analysis complements scalar metrics by revealing which outcomes are most frequently confused. In general, higher-performing ensemble models exhibit fewer confusions between *Dropout*, *Enrolled*, and *Graduate*, whereas simpler models tend to misclassify a larger fraction of minority or ambiguous cases.

A qualitative inspection of Figure 2 can be used to identify whether errors are *symmetric* (mutual confusion between two classes) or *asymmetric* (systematic bias toward a particular outcome). In multi-class educational prediction tasks, ambiguous students often fall near the boundary between *Enrolled* and the other two outcomes, since continued enrollment beyond the normal duration may reflect heterogeneous situations (academic delay, part-time enrollment, or temporary interruption). Models that better separate these borderline cases tend to achieve stronger Macro-F1, as Macro-F1 penalizes poor performance on any single class.

Table 1. Model comparison on the hold-out test split and 5-fold stratified CV

Model	Acc.	Prec. (M)	Rec. (M)	F1 (M)	ROC-AUC (OvR)	CV Acc.	CV Std.
RF	0.7797	0.7347	0.6916	0.7036	0.8919	0.7703	0.0073
LightGBM	0.7706	0.7188	0.7013	0.7082	0.8901	0.7780	0.0095
CatBoost	0.7695	0.7155	0.6990	0.7055	0.8804	0.7697	0.0126
SVM	0.7593	0.7088	0.6702	0.6814	0.8689	0.7658	0.0089
XGBoost	0.7582	0.7009	0.6872	0.6927	0.8831	0.7710	0.0019
LR	0.7571	0.6949	0.6644	0.6717	0.8797	0.7642	0.0108
AdaBoost	0.7458	0.6720	0.6546	0.6578	0.8517	0.7529	0.0140
kNN	0.6915	0.6100	0.5939	0.5952	0.7893	0.7009	0.0095
DT	0.6825	0.6241	0.6247	0.6239	0.7309	0.6704	0.0153
NB	0.6667	0.5809	0.5686	0.5706	0.7909	0.6851	0.0100

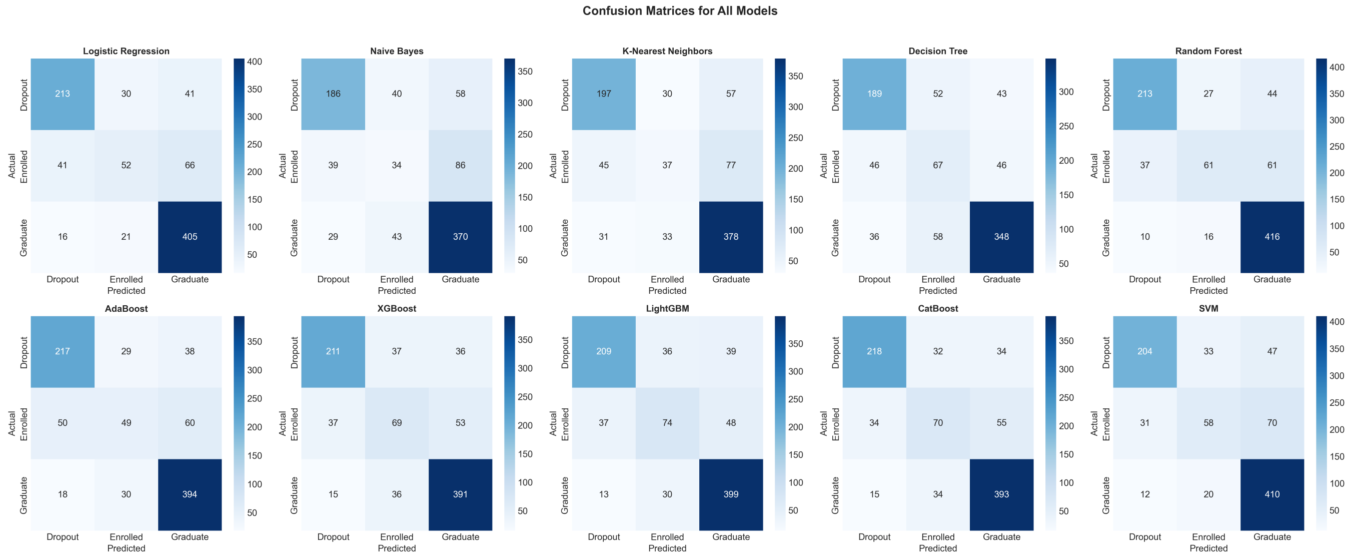

Figure 2. Confusion matrices for all evaluated models on the hold-out test spli

From an intervention perspective, confusion matrices are also useful because different error types have different operational consequences. For example, predicting *Graduate* for a student who eventually *Drops out* may reduce the likelihood of timely support, while predicting *Dropout* for a student who ultimately *Graduates* may lead to unnecessary allocation of resources. Therefore, the confusion matrices provide an actionable view of model behavior beyond aggregate scores.

5.3. Discriminative Ability via ROC Curves

Figure 3 presents One-vs-Rest ROC curves for selected top-performing models. Overall, the best-performing models maintain strong separability across the three classes, consistent with the high ROC-AUC values reported in Table 1. ROC curves provide a threshold-independent view of performance and are especially informative when class proportions are imbalanced.

While Accuracy and Macro-F1 summarize performance at a specific decision rule (the model's default argmax prediction), ROC-AUC evaluates the *ranking quality* of predicted probabilities. This matters for real-world early-warning systems, where institutions may prioritize a subset of students for intervention based on risk scores rather than relying on hard class labels alone. The consistently high OvR ROC-AUC values of the top ensemble models indicate that their probability estimates retain useful separability even when the final predicted class may be uncertain.

It is also common for ROC curves to differ across classes, reflecting that some outcomes are more distinctive given the available features. In practice, classes associated with clear academic signals, e.g., strong early performance, often exhibit stronger separability, whereas intermediate outcomes may be harder to distinguish due to overlapping characteristics. These class-specific ROC curves therefore complement confusion matrices by revealing which outcomes are intrinsically more separable under the chosen feature set.

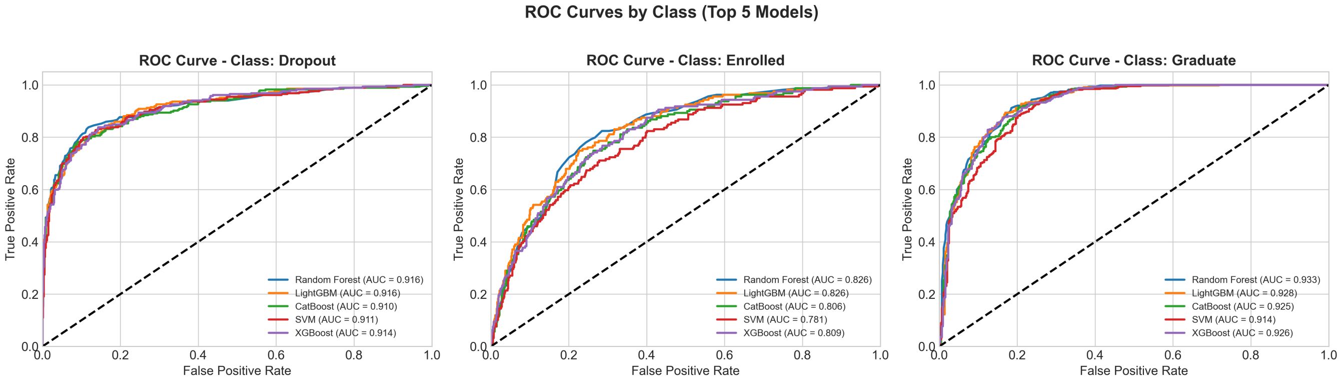

Figure 3. One-vs-Rest ROC curves for the top-performing models across the three outcome classes

5.4. Feature Importance Analysis

To provide interpretability and highlight key predictors, we compute model-based feature importance scores using the RF model. Table 2 lists the top predictors, and Figure 4 visualizes the top-15 ranked features. The most informative variables are dominated by early academic performance indicators, e.g., approved curricular units and grades in the first and second semesters, followed by selected administrative and contextual factors, e.g., tuition fees up to date, course, and macroeconomic indicators. This suggests that students' early academic trajectory provides the strongest signal for final outcomes, while socioeconomic and contextual features offer additional explanatory value.

The prominence of *approved curricular units* and *semester grades* is consistent with the intuition that early academic progress captures both cognitive achievement and behavioral persistence, e.g., completing evaluations and passing courses. These variables likely serve as high-level proxies for attendance, engagement, and effective study habits, and they may therefore be particularly useful for identifying at-risk students during the first year. The importance of *age at enrollment* may reflect differences in student trajectories and external responsibilities, which are often associated with persistence patterns in higher education.

Administrative factors such as *tuition fees up to date* and *scholarship holder* status also contribute to prediction, highlighting the role of financial stability and institutional support in academic outcomes. Moreover, the presence of course-related and macroeconomic indicators suggests that program context and broader economic conditions may moderate dropout risk and completion probabilities. Taken together, these findings support the view that effective retention strategies should combine academic monitoring with targeted administrative and financial support mechanisms.

Finally, it is important to interpret feature importance as *predictive association* rather than causation. Importance scores indicate which variables the model uses to improve prediction under the observed data distribution, but they do not by themselves establish that manipulating a variable would change outcomes. Nevertheless, these results provide a useful evidence base for prioritizing signals when designing decision-support tools for early identification and intervention.

Table 2. Top-10 most important features (RF)

Feature	Importance
Curricular units 2nd sem (approved)	0.1469
Curricular units 2nd sem (grade)	0.1039
Curricular units 1st sem (approved)	0.0919
Curricular units 1st sem (grade)	0.0729
Age at enrollment	0.0461
Curricular units 2nd sem (evaluations)	0.0437
Tuition fees up to date	0.0383
Course	0.0377
Curricular units 1st sem (evaluations)	0.0363
Father's occupation	0.0341

6. Conclusion

This paper studied student outcome prediction in higher education as a three-class classification task (*Dropout*, *Enrolled*, and *Graduate*) using demographic/enrollment, socioeconomic/administrative, early academic performance, and macroeconomic variables. We compared a range of widely used ML classifiers under a unified preprocessing and evaluation protocol.

Overall, ensemble models achieved the best performance, with RF and gradient boosting methods, e.g., LightGBM and CatBoost, providing strong and stable results. Feature importance analysis indicated that early academic progress indicators (approved units and semester grades) are the most influential predictors, followed by selected administrative and contextual factors. These findings support the use of ensemble-based

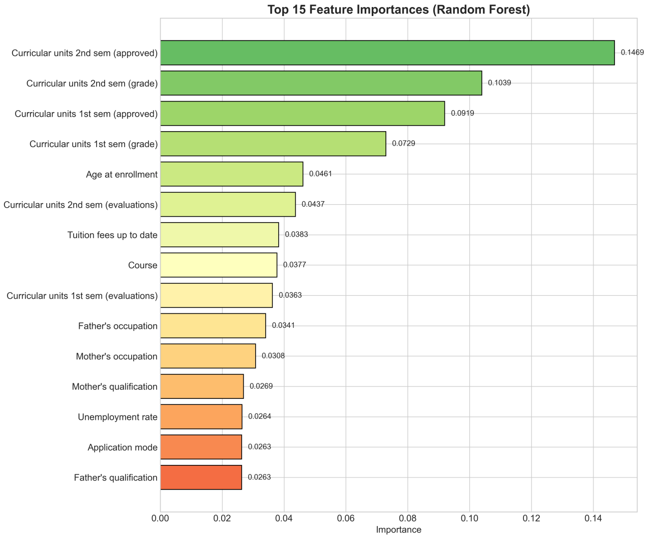

Figure 4. Top-15 feature importances computed from the RF model.

models and early academic signals for building practical, data-driven student support and retention systems.

References

[1] J. Rajni and D. B. Malaya, "Predictive analytics in a higher education context," *IT Prof.*, vol. 17, no. 4, pp. 24–33, Jul. 2015.

[2] Y. Wang, F. You, and Q. Li, "Machine learning algorithms for fostering innovative education for university students," *Electronics*, vol. 13, no. 8, pp. 1506–1519, Apr. 2024.

[3] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Appl. Sci.*, vol. 10, no. 3, pp. 1042–1057, Feb. 2020.

[4] K. Ahmad, W. Iqbal, A. El-Hassan, J. Qadir, D. Benhaddou, M. Ayyash, and A. Al-Fuqaha, "Data-driven artificial intelligence in education: A comprehensive review," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 12–31, Sep. 2023.

[5] T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, and L. Galligan, "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," *IEEE Access*, vol. 10, pp. 56 720–56 739, May 2022.

[6] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: A literature review," *Educ. Sci.*, vol. 13, no. 12, pp. 1216–1242, Dec. 2023.

[7] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of students' performance prediction using machine learning techniques," *Educ. Sci.*, vol. 11, no. 9, pp. 552–578, Sep. 2021.

[8] B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, and J. B. Idoko, "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies," *Appl. Sci.*, vol. 11, no. 22, pp. 10 907–10 929, Nov. 2021.

[9] L. Yan and Y. Liu, "An ensemble prediction model for potential student recommendation using machine learning," *Symmetry*, vol. 12, no. 5, pp. 728–744, May 2020.

[10] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb, "Intelligent decision support system for predicting students' e-learning performance using ensemble machine learning," *Mathematics*, vol. 9, no. 17, pp. 2078–2099, Aug. 2021.

[11] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance prediction of students in higher education using multi-model ensemble approach," *IEEE Access*, vol. 11, pp. 136 091–136 108, Dec. 2023.

[12] E. Alhazmi and A. Sheneamer, "Early predicting of students performance in higher education," *IEEE Access*, vol. 11, pp. 27 579–27 589, Mar. 2023.

[13] B. Pardamean, T. Suparyanto, T. W. Cenggoro, D. Sudigyo, and A. Anugrahana, "AI-based learning style prediction in online learning for primary education," *IEEE Access*, vol. 10, pp. 35 725–35 735, Apr. 2022.

[14] M. Zafari, A. Sadeghi-Niaraki, S.-M. Choi, and A. Esmaeily, "A practical model for the evaluation of high school student performance based on machine learning," *Appl. Sci.*, vol. 11, no. 23, pp. 11 534–11 550, Dec. 2021.

[15] D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq, "Feature selection for UK disabled students' engagement post higher education: A machine learning approach for a predictive employment model," *IEEE Access*, vol. 8, pp. 159 530–159 541, Sep. 2020.

[16] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour, "An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19," *IEEE Access*, vol. 10, pp. 6286–6303, Jan. 2022.

[17] M. Nafuri, A. F. Sani, N. S. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering analysis for classifying student academic performance in higher education," *Appl. Sci.*, vol. 12, no. 19, pp. 9467–9488, Sep. 2022.

[18] G. Latif, S. E. Abdelhamid, K. S. Fawagreh, G. B. Brahim, and R. Alghazo, "Machine learning in higher education: Students' performance assessment considering online activity logs," *IEEE Access*, vol. 11, pp. 69 586–69 600, Jul. 2023.

[19] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, "Identification of student behavioral patterns in higher education using k-means clustering and support vector machine," *Appl. Sci.*, vol. 13, no. 5, pp. 3267–3280, Mar. 2023.

[20] M. V. Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-class phased prediction of academic performance and dropout in higher education," *Appl. Sci.*, vol. 13, no. 8, pp. 4702–4716, Apr. 2023.

[21] A. S. Mohammad, M. T. S. Al-Kaltakchi, J. Alshehabi Al-Ani, and J. A. Chambers, "Comprehensive evaluations of student performance estimation via machine learning," *Mathematics*, vol. 11, no. 14, pp. 3153–3168, Jul. 2023.

[22] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a web-based prediction system for students' academic performance," *Data*, vol. 7, no. 2, pp. 21–39, Jan. 2022.

[23] M. Maphosa, W. Doorsamy, and B. Paul, "Improving academic advising in engineering education with machine learning using a real-world dataset," *Algorithms*, vol. 17, no. 2, pp. 85–107, Feb. 2024.

[24] B. Holicza and A. Kiss, "Predicting and comparing students' online and offline academic performance using machine learning algorithms," *Behav. Sci.*, vol. 13, no. 4, pp. 289–309, Mar. 2023.

[25] M. O. Hegazi, B. Almaslukh, and K. Siddig, "A fuzzy model for reasoning and predicting student's academic performance," *Appl. Sci.*, vol. 13, no. 8, pp. 5140–5163, Apr. 2023.

[26] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predict students' dropout and academic success," Dec. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5777340