

Quantifying Predictive Uncertainty in Student Academic Outcomes Using Conformal Prediction

Kazi Sabbir Ahmad Nahin

sabbir.sta@iubat.edu

University of Kentucky

Research Article

Keywords: student dropout prediction, academic success, machine learning, conformal prediction, uncertainty quantification, explainable AI, SHAP, educational data mining, learning analytics, early warning systems, higher education policy

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11037494/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantifying Predictive Uncertainty in Student Academic

Outcomes Using Conformal Prediction

Authors:

Kazi Sabbir Ahmad Nahin ^{1,2*}

Affiliations:

¹ Dr. Bing Zhang Department of Statistics, College of Arts and Sciences, University of Kentucky, Lexington, Kentucky, USA

² Department of Quantitative Sciences, International University of Business Agriculture and Technology, Dhaka, Bangladesh

***Correspondence:**

Kazi Sabbir Ahmad Nahin

Email: sabbirahmadnahin@gmail.com

sabbir.sta@iubat.edu

Abstract

Early identification of students at risk of dropout can support timely and targeted interventions in higher education. However, the extensive literature on machine learning-based dropout prediction has emphasized predictive accuracy and, more recently, explainability, while rarely communicating how reliable an individual prediction is. This study developed an uncertainty-aware and interpretable framework for predicting student academic outcomes by integrating machine learning, split conformal prediction, and SHapley Additive exPlanations (SHAP). Using 4,424 student records from a Portuguese higher education institution, six models were evaluated for binary Graduate-Dropout and three-class Graduate-Enrolled-Dropout classification. In the binary task, LightGBM achieved the highest accuracy (0.906) and macro-F1 (0.900), while XGBoost achieved the highest ROC-AUC (0.942); these values are within the range reported by previous studies of the same dataset. The three-class task was more challenging because of the transitional Enrolled category; Random Forest achieved the highest macro-F1 (0.750) and ROC-AUC (0.925). Conformal prediction sets achieved empirical coverage close to nominal levels: at 95% coverage in the binary task, LightGBM and XGBoost reached 0.951 and 0.952 with mean set sizes of 1.120 and 1.121. At least 9–12% of binary cases and 38–43% of three-class cases received multi-outcome sets, explicitly flagging students whose trajectories remained undetermined. SHAP identified academic progression, grades, admission performance, tuition-payment status, and financial vulnerability as key predictors. Compared with prior work, conformal prediction adds a distribution-free, per-student reliability signal that can support tiered, human-in-the-loop early-warning policies.

Keywords: student dropout prediction; academic success; machine learning; conformal prediction; uncertainty quantification; explainable AI; SHAP; educational data mining; learning analytics; early warning systems; higher education policy.

Background

Student dropout in higher education is a major challenge for universities worldwide, with significant consequences for students, institutions, and society [1, 2]. Academic non-completion affects economic growth, workforce competitiveness, and social equity [3, 4]. The Organization for Economic Co-operation and Development (OECD) reported that the higher education dropout rate was approximately 20% in 2019 [5], while the undergraduate dropout rate in the United States reached 40% in 2020 [6]. Students who leave college without completing a degree face poorer employment outcome: they are 20% more likely to be unemployed than degree holders [7] and earn about 66% less than those who complete a bachelor's degree [8]. Institutions also experience substantial losses, with universities forfeiting an estimated \$16.5 billion each year in tuition revenue because of attrition [9], while public investments in grants and subsidies are diminished when students do not complete their programs [8].

Student dropout is shaped by a complex combination of academic, demographic, socioeconomic, and institutional factors [10, 11]. Prior research consistently identifies academic performance; especially grade point average, credit accumulation, and course completion as the strongest predictor of student persistence [12]. Financial features, including tuition payment status, scholarship receipt, and debt burden, are also important indicators of dropout risk [13, 14]. Demographic factors such as age at enrollment, gender, and socioeconomic background further influence student outcomes [15]. Dropout timing is equally important, as most attrition occurs

during the first year of enrollment, highlighting the need for early intervention strategies to improve retention [16, 17].

To address these challenges, researchers and institutions increasingly use machine learning (ML) to identify students at risk of dropout before they leave their programs [18, 19]. Existing studies have examined a range of ML methods, with tree-based ensemble models repeatedly showing strong performance [20–22]. Algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost often achieve high predictive accuracy, with F1-scores above 0.85 and AUC values near 0.95 [23–25]. Comparative studies indicate that gradient boosting methods generally outperform traditional approaches such as logistic regression and support vector machines, particularly in heterogeneous and moderately imbalanced feature spaces [26–28]. Feature selection methods, including Boruta and LASSO, have further improved model interpretability and performance by identifying the most influential predictors [29, 30]. These developments, together with the theoretical, ethical, and methodological debates that surround them, are reviewed in detail in the following Literature Review section.

Despite this progress, most dropout prediction studies share an important limitation: they rarely quantify predictive uncertainty. Current models typically provide point estimates i.e. single class labels or probabilities, without indicating how reliable those predictions are. This is especially problematic in high-stakes educational settings, where misclassification can have serious consequences [31]. A false positive—labeling a student as at risk when they would have graduated—may lead to unnecessary intervention, inefficient use of institutional resources, and possible student stigmatization. A false negative—failing to identify a student who is truly at risk—can prevent timely support and contribute to avoidable attrition. Without uncertainty quantification, predictive models are less trustworthy and less useful for decision-making [32–35].

Conformal prediction (CP) provides a principled way to address this gap by producing valid predictive inferences without strong distributional assumptions [36–38]. Rather than generating only point predictions, CP creates prediction sets that include the true outcome with a user-specified probability, known as the coverage guarantee [39, 40]. This allows educational stakeholders to assess both the predicted outcome and the uncertainty surrounding it, supporting more informed and risk-aware decisions. Split, or inductive, conformal prediction is especially practical because it is computationally efficient and model agnostic, making it easy to integrate into existing ML pipelines [36, 41, 42].

This study addresses this gap by developing and evaluating a framework that combines conformal prediction with widely used machine learning models for student outcome prediction. We apply split conformal prediction to both binary and multiclass classification tasks using the Predict Students' Dropout and Academic Success dataset, which includes 4,424 student records from a Portuguese higher education institution over ten academic years [2]. The framework is applied to six models, namely Logistic Regression, Support Vector Machine, Neural Network, Random Forest, XGBoost, and LightGBM, and is evaluated at multiple confidence levels (99%, 95%, and 90%). In addition, the study incorporates explainable AI (XAI) through SHAP-based interpretation and model-specific feature importance, helping to address the need for transparent and actionable prediction in higher education. The study makes three contributions. First, it situates uncertainty quantification within the established literature on dropout theory, educational data mining, and early-warning systems, and benchmarks its predictive results against previous studies that used the same dataset. Second, it provides a systematic comparison of conformal prediction across six model families, two outcome formulations, and three coverage levels, showing how prediction sets make the ambiguity of the Enrolled trajectory explicit. Third, it translates conformal outputs into

a tiered decision framework and discusses the resulting implications for institutional policy, equity, and resource allocation. The remainder of the paper is organized as follows: the next section reviews related literature and states the research questions; subsequent sections describe the data and methods, report the results, and discuss them in relation to prior work and their implications for policy and practice.

Literature Review

Theoretical perspectives on student dropout

Research on student departure has a long theoretical tradition that predates computational prediction. Tinto's student integration model conceptualizes dropout as the outcome of a longitudinal process in which students' pre-entry attributes, goals, and institutional commitments are continually modified by the degree of their academic and social integration into the institution [43, 44]. Bean and Metzner's attrition model, developed for nontraditional students, places greater weight on environmental factors such as finances, employment, and outside encouragement, which can outweigh social integration in shaping persistence [45]. Cabrera et al. showed that the two frameworks are complementary rather than competing, with financial factors and the ability to pay affecting persistence both directly and indirectly through academic integration [46]. Contemporary reviews broadly confirm this multidimensional view, identifying prior and current academic performance, financial constraints, socioeconomic background, motivation, and institutional characteristics as the main determinants of non-completion [1, 5, 10, 11]. A recent synthesis of 95 studies concluded that grade point average and obtained credits are the most consistent predictors

of academic performance and dropout, and that their effects are mediated by student-level and throughput factors such as finances, work, and academic engagement [47].

These theories matter for predictive modeling in two ways. First, they explain why administrative variables routinely recorded by institutions, such as curricular-unit progression, grades, tuition-payment status, and scholarship holding, carry predictive signal: they are observable traces of academic integration and financial capacity. Second, they caution that dropout is a process rather than a single event, and that students' trajectories can remain undecided for extended periods [3, 17]. The Enrolled category in the dataset analyzed here is a direct manifestation of such an unresolved process. A modeling framework that can express when a student's trajectory is still undetermined, rather than forcing an early and possibly unreliable label, is therefore theoretically well motivated.

Machine learning for dropout and academic-success prediction

Educational data mining and learning analytics have produced a large body of predictive studies. Systematic reviews document a progression from statistical models and single decision trees toward ensembles and deep learning, with academic records, demographic characteristics, and online learning activity as the most frequently used inputs [18, 19, 48–50]. Prenkaj et al. reviewed dropout prediction in online courses and showed that design choices, such as the definition of dropout, the prediction horizon, and the evaluation metric, strongly influence reported performance [51]. Early institutional studies, such as Delen's comparison of neural networks, decision trees, support vector machines, and logistic regression for freshman retention, already found that ensemble models performed best and that accuracy was markedly lower for the minority class [52]. Subsequent work using administrative records from European universities reached

similar conclusions: Kemper et al. showed that dropout could be predicted accurately from examination data after only a few semesters [53], and Berens et al. developed an early-detection system for German universities whose accuracy increased as semester performance data accumulated [54].

More recent studies extend these findings to diverse institutional and national contexts, including Korean, Somaliland, Brazilian, and Moroccan settings, and repeatedly report strong performance for gradient boosting, random forests, and stacked ensembles [20, 21, 24, 55–57]. Other work has incorporated behavioral and engagement data, for example from smartphone applications or learning management systems, to improve retention prediction [12, 58]. A quality-weighted systematic review of 75 studies published between 2019 and 2024 concluded that ensemble methods, particularly Random Forest and XGBoost, dominate the field, typically achieving accuracies of approximately 86–88%; however, most studies relied on single-institution data, did not account for temporal change, and rarely evaluated the fairness or impact of the interventions triggered by predictions [59]. Comparative reviews also emphasize that model rankings depend heavily on the evaluation design, so that small differences between algorithms should be interpreted cautiously [26]. Bertolini et al. further showed that bootstrapped estimates of discrimination were lower and more variable than single point estimates, highlighting the risk of relying on point metrics alone when building educational prediction pipelines [60].

Prior studies using the Portuguese higher education dataset

The dataset used in the present study was released by Realinho et al. as part of a learning analytics tool that provides the tutoring team at the Polytechnic Institute of Portalegre with estimates of dropout and failure risk [2]. Because it is public, well documented, and contains demographic,

socioeconomic, macroeconomic, and academic variables, it has become a common benchmark for comparing modeling approaches. Martins et al. used data from the same institution for early prediction and showed that strong overall performance can coexist with weak classification of the smaller and less distinct classes [15]. In a later multi-class phased analysis, the same group found that Random Forest models with imbalance-handling strategies performed best and that the most useful prediction moment was the end of the first semester, when early academic performance becomes available [61]. Table 1 summarizes selected subsequent studies. Villar and de Andrade compared several supervised algorithms with hyperparameter optimization and SHAP, reporting the strongest results for CatBoost and LightGBM [22]; Ridwan et al. reported an accuracy of approximately 0.87 for XGBoost in a dropout versus academic-success formulation [25]; Romero and Liao, who used the same Graduate versus Dropout formulation as the present study, reported F1-scores of approximately 0.90 for XGBoost and Random Forest and combined prediction with propensity-score analysis of scholarship effects [14]; Liu et al. obtained a ROC-AUC of 0.957 for Random Forest in a dropout versus non-dropout formulation and studied privacy-preserving synthetic data with SHAP explanations [27]; and Bettahi et al. replicated an explainable pipeline on the same public data, reporting an F1-score of 0.792 and a ROC-AUC of 0.922 for XGBoost [56]. Studies that retained all three outcomes consistently reported lower performance, particularly for the Enrolled class, with macro-F1 values of approximately 0.70–0.72 when only enrollment-time information was used [23, 28].

Table 1. Selected prior studies that used the Predict Students' Dropout and Academic Success dataset, compared with the present study. Reported values are taken from the original publications; because outcome definitions, preprocessing, validation schemes, and metrics differ, values are indicative rather than directly comparable.

Study	Outcome formulation	Best model(s)	Reported performance	Explainability	Uncertainty quantification
Martins et al. [61]	Three-class, three phases of first year	Random Forest with imbalance handling	Best results at end of first semester	Feature relevance	None
Villar & de Andrade [22]	Three-class	CatBoost, LightGBM	Dropout-class F1 $\approx$ 0.87–0.88	SHAP	None
Ridwan et al. [25]	Dropout vs academic success	XGBoost	Accuracy $\approx$ 0.87; F1 $\approx$ 0.81	None	None
Romero & Liao [14]	Graduate vs Dropout	XGBoost; Random Forest; GAM	XGBoost accuracy 0.882, F1 0.907; RF ROC-AUC 0.935	Feature importance; GAM; propensity scores	None
Liu et al. [27]	Dropout vs non-dropout	Random Forest	ROC-AUC 0.957; accuracy 0.924 (5-fold CV)	SHAP	None
Bettahi et al. [56]	Replication on public data	XGBoost	F1 0.792; ROC-AUC 0.922	SHAP	None
Alameri [23]	Binary and three-class	Tuned XGBoost and ensembles	Binary balanced accuracy $>$ 0.91, ROC-AUC $>$ 0.95; lower multiclass performance	SHAP	None
Morsu [28]	Three-class, enrollment-time data	XGBoost	Macro-F1 $\approx$ 0.70–0.72; macro ROC-AUC $\approx$ 0.88–0.90	Feature importance	Seed-to-seed variability only
Present study	Graduate vs Dropout; three-class	LightGBM, XGBoost (binary); Random Forest, LightGBM (three-class)	Binary: accuracy 0.906, ROC-AUC 0.942; three-class: macro-F1 0.750, ROC-AUC 0.925	SHAP; model-specific importance	Split conformal prediction sets at 90%, 95%, and 99% coverage

Three regularities emerge from this body of work. First, binary formulations are consistently easier than three-class formulations, with reported accuracies of roughly 0.87–0.92 and ROC-AUC values of roughly 0.92–0.96. Second, the intermediate Enrolled category is repeatedly identified as the main source of error. Third, second-semester academic progression, grades, and tuition-payment status are the most influential predictors across modeling approaches. Notably, every study in Table 1 evaluates performance with aggregate point metrics, and none reports how reliable

an individual student's prediction is. As a result, two students who receive the same predicted label are treated as equally certain, even when the evidence supporting the two predictions differs substantially.

Explainable artificial intelligence in dropout prediction

As predictive models have become more complex, explainability has become a central requirement for their educational use. SHAP provides a unified, game-theoretic framework for attributing a model's output to individual features at both global and local levels [62]. Nagy and Molontay demonstrated how permutation importance, partial-dependence plots, LIME, and SHAP can explain dropout predictions from pre-enrollment data and reported a user study in which higher education stakeholders evaluated these explanations [63]. Krüger et al. combined dataset enrichment with model-explanation techniques to justify dropout predictions across educational stages [55], and Bettahi et al. proposed a governance-oriented SHAP layer that surfaces risk and protective factors, as well as ambiguous cases, for human review [56]. Studies of the Portuguese dataset have also used SHAP or interpretable models to show that curricular-unit progression and financial variables dominate predictions [14, 22, 27]. At the same time, Rudin argued that post hoc explanations of black-box models can be unreliable in high-stakes settings [64], and explanation methods describe associations learned by a model rather than causal effects. More fundamentally, an explanation answers the question of why a model made a prediction, but not how much confidence that prediction deserves; a detailed explanation of an uncertain prediction can therefore convey false precision.

Early-warning systems, uncertainty, and responsible use

Predictive models in higher education are typically deployed within early-warning or early-alert systems. Pioneering implementations such as Course Signals at Purdue University and the Open Academic Analytics Initiative demonstrated that predictive alerts could be integrated into institutional workflows [65, 66], and later systems have been developed for online and blended higher education [58]. However, evidence regarding the effects of such systems on student outcomes is mixed. Observational studies have linked early-alert systems to improved retention [67], whereas a field experiment in the Netherlands found no significant reduction in dropout when counselors were provided with early-warning information [68]. The literature on learning analytics ethics adds that labeling students as at risk can stigmatize them, create self-fulfilling expectations, and shift institutional attention toward students who are easier to predict [69–72].

A related line of research examines algorithmic fairness. Reviews show that predictive models in education can encode and amplify existing inequities [73, 74]. Bird et al. showed that algorithmic bias in course- and degree-completion models would lead to fewer marginal Black students receiving additional support [75], and studies of whether protected attributes should be included in dropout models have found only modest effects on performance and fairness, with consequences that depend on local context [76, 77]. These concerns are directly connected to uncertainty: when a model is systematically less informative for some groups, point predictions conceal this difference, whereas an explicit measure of reliability can reveal it.

Despite these concerns, uncertainty quantification remains rare in educational prediction. Hu and Rangwala used Bayesian deep learning to attach uncertainty estimates to grade predictions and argued that such estimates are essential for reliable early-warning systems [78]. Bertolini et al.

used bootstrap resampling to characterize the variability of performance estimates [60], and Kwon proposed probability calibration combined with SHAP to make dropout-risk estimates more reliable and transportable [79]. These approaches are valuable but limited: Bayesian methods are model-specific and depend on prior and likelihood assumptions, bootstrap analyses describe the variability of aggregate metrics rather than the reliability of individual predictions, and calibrated probabilities do not provide a finite-sample guarantee that the true outcome is contained in a reported set.

Conformal prediction and its emerging use in education

Conformal prediction offers a distribution-free alternative. Built on the assumption of exchangeability, it wraps any fitted model and produces prediction sets that contain the true outcome with a user-specified probability [31, 33–35]. In classification, set-valued predictors make ambiguity explicit by returning more than one label when the evidence does not support a single outcome [32, 80, 81]. Methodological extensions address class- or group-conditional validity through Mondrian and conditional conformal procedures [82, 83], fairness through equalized coverage across protected groups [84], and distribution shift through weighted and adaptive procedures [39, 40, 85].

Evidence from human-AI decision making suggests that prediction sets can be useful decision aids. Human-subject experiments, including a pre-registered randomized controlled trial, have found that conformal prediction sets improved human accuracy compared with top-1 or fixed-size alternatives [86, 87], and decision-support systems that ask experts to choose from conformal sets can improve expert predictions [88]. The benefit, however, depends on task difficulty and set size [89], and sets that satisfy equalized coverage may still produce disparate downstream decisions

[90]. In education, applications remain scarce. Morsomme and Smirnov used inductive conformal prediction to construct valid prediction intervals for course grades in a recommender system [91], and Elmasry et al. recently integrated calibrated support vector classification, split conformal prediction, SHAP, and graph-based refinement into an uncertainty-aware early-warning framework for dropout in engineering education [92]. To the best of our knowledge, no previous study has systematically compared conformal prediction across multiple model families, both binary and three-class outcome formulations, and several coverage levels on this widely used benchmark dataset, nor examined how prediction sets express the ambiguity of the Enrolled trajectory and how they could be translated into institutional policy.

Research gap and research questions

The literature reviewed above reveals four gaps. First, dropout prediction studies, including all those that used the present dataset, report aggregate point metrics without quantifying the reliability of individual predictions. Second, the three-class formulation is usually treated as a harder classification problem, rather than as a setting in which some trajectories are inherently undetermined. Third, explainability has been studied separately from uncertainty, even though both are needed for responsible decisions. Fourth, few studies translate model outputs into concrete decision rules that institutions and policymakers can apply. Accordingly, this study addresses three research questions. RQ1: How do six model families perform on binary and three-class outcome prediction, and how do these results compare with previous studies of the same dataset? RQ2: Does split conformal prediction achieve near-nominal coverage with informative prediction sets across models and coverage levels, and how does set size differ between the binary and three-class formulations? RQ3: Which predictors drive the models, and how can prediction sets and SHAP explanations be combined into an uncertainty-aware decision-support workflow for institutions?

Data Description and Preprocessing

Dataset

This study used the “Predict Students' Dropout and Academic Success” dataset from the UCI Machine Learning Repository, originally developed by Realinho and colleagues at the Polytechnic Institute of Portalegre in Portugal. The dataset contains records for 4,424 students enrolled between the 2008/2009 and 2018/2019 academic years, covering a decade of undergraduate education following the implementation of the Bologna Process [2]. Students in the dataset represent 17 academic programs spanning fields such as agronomy, design, education, nursing, journalism, management, social services, and technology.

Each observation corresponds to a single student and includes 36 predictor features (supplementary table 1) along with one outcome feature. The dataset contains no missing values. The predictors capture information across five domains: demographic characteristics, socioeconomic background, macroeconomic conditions, academic information at admission, and academic performance during the first two semesters. Together, these features provide a broad view of factors that may influence student success and persistence in higher education.

The outcome feature, *Target*, categorizes students into one of three mutually exclusive groups: Graduate, Dropout, or Enrolled. Graduates are students who completed their degree within the expected timeframe, while dropouts are students who left the institution before graduation, including those who transferred to another institution or program. The enrolled category includes students who remain registered at the end of the observation period. The dataset is moderately imbalanced, with 2,209 graduates (49.9%), 1,421 dropouts (32.1%), and 794 enrolled students

(17.9%). Graduation is therefore the most common outcome, followed by dropout, with ongoing enrollment representing the smallest group.

Preprocessing Strategies

Feature Engineering

To capture additional information beyond the original features, several derived features were created based on findings from previous research [14] [23]. The `avg_grade` feature was calculated as the mean of first- and second-semester grades, providing a single measure of overall academic achievement. Likewise, `avg_approved` represented the average number of approved curricular units across the two semesters. An `approval_rate` feature was generated to reflect academic progression, while `financial_stress` was constructed by combining debtor status and tuition payment status into a score ranging from 0 to 2. Higher values indicated greater financial vulnerability. These engineered features were retained for all subsequent analyses and modeling procedures.

Feature Encoding

Several categorical features were first recoded to reduce dimensionality and improve interpretability, following the approach proposed by Romero and Liao [14]. The Previous qualification, originally represented by 17 numerical codes, was grouped into Secondary Education, Advanced/Specialized, and Unknown. Parental education was consolidated into Basic Education, Secondary+ Education, and Unknown, while mother's and father's occupations were grouped into Other/Student, White Collar/Admin, and Blue Collar/Labor. Binary features, including displaced status, educational special needs, debtor status, tuition-payment status, scholarship status, and international student status, were converted from numeric values to

descriptive Yes/No labels; gender was similarly recoded as Male/Female. These transformations made the categorical features easier to interpret during exploratory analysis and model explanation while preparing them for model-specific encoding.

After recoding, features were processed through model-specific pipelines. Logistic Regression, SVM, and Neural Network used a scaled pipeline in which numeric features were median-imputed and standardized, binary features were mode-imputed, and categorical features were mode-imputed and one-hot encoded to avoid multicollinearity [93] and unintended ordinal assumptions [94, 95].

For Random Forest and XGBoost, numeric features were median-imputed, categorical features were mode-imputed and ordinally encoded, and no scaling was applied because decision-tree methods are not sensitive to feature scale [96, 97].

For LightGBM, numeric features were median-imputed, while categorical features were mode-imputed, converted to categorical data types, and retained in their original form so the model could use its native categorical feature support [98].

Target Encoding

The target feature was encoded numerically before model training. For multiclass classification, the labels were assigned as Dropout = 0, Enrolled = 1, and Graduate = 2. For binary classification, students in the Enrolled category were excluded, and the remaining classes were encoded as Dropout = 0 and Graduate = 1. This binary formulation aligns with prior studies focused on distinguishing student completion from attrition [14, 22, 25]. To address class imbalance, class weights were calculated using the balanced weighting scheme and incorporated into model

training. This increased the penalty for misclassifying minority classes and helped improve model performance across outcome categories.

Handling Multicollinearity

Correlation analysis revealed substantial overlap among several academic performance features. Strong correlations were observed between first- and second-semester measures of credited units, enrolled units, approved units, grades, and evaluations, with correlation coefficients ranging from 0.80 to 0.96. Nationality and international status also showed a very high correlation ($r = 0.91$).

To reduce redundancy and limit the effects of multicollinearity, highly correlated first-semester features were removed, while their second-semester counterparts were retained. This decision was supported by previous research suggesting that later academic performance is more informative for predicting final student outcomes [12, 15]. The nationality feature was also excluded because of its strong overlap with international status.

Feature Selection

Feature selection was performed using two complementary methods: Boruta and LASSO [29, 30, 99]. Boruta identifies relevant features by comparing the importance of original features with randomized shadow features [99], whereas LASSO performs feature selection through L1-regularized logistic regression [29, 100]. Both methods were applied separately to the binary and multiclass classification tasks.

For the binary classification task, Boruta selected 14 features and LASSO selected 31, producing a combined set of 33 predictors. For the multiclass task, Boruta selected 11 features and LASSO selected 38, yielding a combined set of 39 predictors. Across both methods, academic performance

indicators, financial features, demographic characteristics, and macroeconomic measures consistently emerged as important predictors. The engineered features avg_grade, avg_approved, approval_rate, and financial_stress were selected in all settings, supporting their usefulness for predicting student outcomes.

Data Splitting for Modeling and Conformal Prediction

For the primary analysis, the dataset was partitioned into mutually exclusive training, calibration, and test sets using a stratified 70/15/15 split in each repetition. The training set was used for preprocessing, model fitting, probability calibration, and hyperparameter selection; the calibration set was used only to estimate nonconformity scores and determine the conformal threshold; and the test set was reserved for final evaluation of predictive performance and conformal prediction sets. Stratification preserved class proportions across the subsets. Because random splitting does not evaluate performance under cohort or institutional change, these results are interpreted as internal validation rather than evidence of external transportability.

The full workflow is summarized in Figure 1, which illustrates the study framework for split conformal prediction: data splitting, training of the candidate machine learning models, calibration-based estimation of nonconformity scores, and construction of prediction sets for new observations. The entire procedure was repeated across five random seeds, with the train-calibration-test partition reconstructed for each seed to assess model stability and reduce dependence on a single random split.

To prevent information leakage, all preprocessing operations, including imputation, scaling, encoding, feature selection, and probability calibration, were fitted using the training data only and then applied unchanged to the calibration and test sets. No information from the calibration or test

observations was used during hyperparameter tuning or model selection, thereby preserving the validity of the conformal calibration step and ensuring a fair assessment of out-of-sample performance.

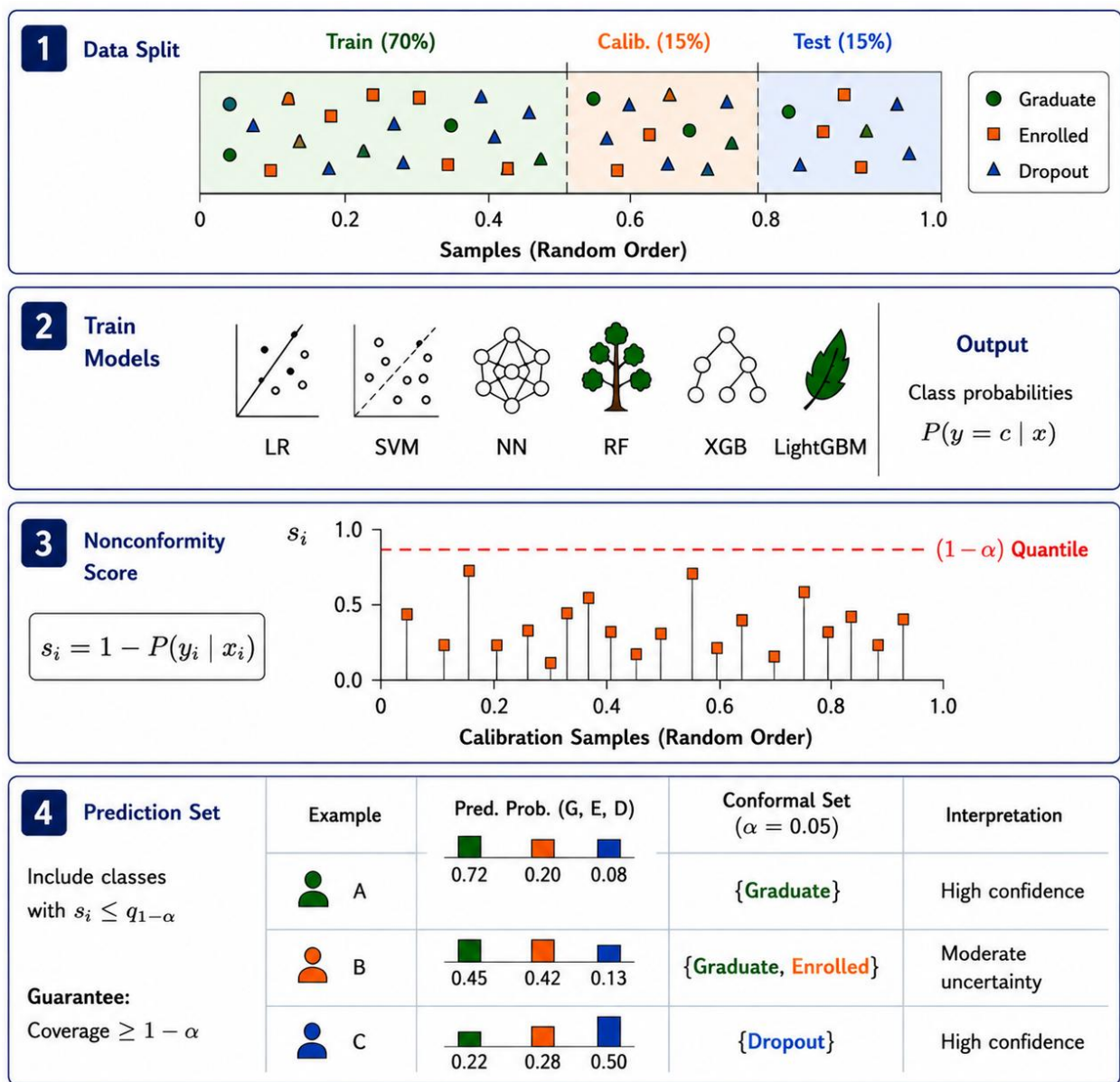


Figure 1. Overview of the split conformal prediction framework used for student outcome classification. The workflow shows the stratified division of the dataset into training, calibration, and test subsets; training of machine learning models to estimate class probabilities; computation of nonconformity scores on the calibration set; and construction of prediction sets with finite-sample

398 *coverage guarantees for new observations. It was created with assistance from ChatGPT to visually summarize the study workflow.*
399 *The authors reviewed and finalized the figure for accuracy.*

400 Methodology

401 Predictive models

402 Six supervised classification algorithms were evaluated to represent complementary model
403 families: multinomial logistic regression, support vector machine (SVM), artificial neural network,
404 random forest, XGBoost, and LightGBM. This model set spans linear probabilistic classification,
405 maximum-margin kernel methods, flexible nonlinear function approximation, bagged decision-
406 tree ensembles, and gradient-boosted trees. Using heterogeneous model classes allows predictive
407 discrimination, probabilistic performance, and conformal-set efficiency to be compared under a
408 common data-splitting and evaluation framework.

409 Logistic regression

410 For a K -class outcome, multinomial logistic regression models the conditional probability of class
411 k as

$$412 \quad P(Y = k \mid \mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x} + b_k)}{\sum_{j=1}^K \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)}, \quad k = 1, \dots, K.$$

413 The parameters are estimated by minimizing the multinomial cross-entropy loss with L_2
414 regularization. In common implementations, the tuning parameter C is the inverse regularization
415 strength, so smaller values impose stronger shrinkage. Logistic regression provides an
416 interpretable linear baseline and directly yields class-probability estimates [101].

417 Support vector machine

418 For binary classification with training observations $\{(x_i, y_i)\}_{i=1}^n$ and $y_i \in \{-1, +1\}$, the soft-
419 margin SVM solves

$$420 \quad \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

421 subject to

$$422 \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

423 Nonlinear decision boundaries were modeled with the radial basis function (RBF) kernel

$$424 \quad K(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\}.$$

425 The multiclass problem was handled through the multiclass strategy implemented by the selected
426 software. Because the conformal procedure below requires class probabilities rather than only
427 margin scores, SVM decision values must be converted to probability estimates using a calibration
428 procedure fitted exclusively within the training data, thereby avoiding information leakage from
429 the calibration and test sets. The maximum-margin formulation and kernel construction follow
430 standard SVM methodology [102, 103].

431 Artificial neural network

432 A neural network was used to model nonlinear relationships among predictors. For an illustrative
433 hidden layer,

$$434 \quad h = g(W_1 x + b_1),$$

$$435 \quad z = W_2 h + b_2,$$

436 and the output probabilities are

437
$$P(Y = k | \mathbf{x}) = \frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)}.$$

438 Here, $g(\cdot)$ denotes a nonlinear activation function, W_1 and W_2 are weight matrices, and b_1 and b_2
439 are bias vectors. Network parameters were estimated by minimizing categorical cross-entropy
440 through backpropagation with L_2 regularization. Neural networks provide flexible nonlinear
441 function approximation but generally require careful regularization and tuning to control
442 overfitting [101, 104].

443 Random forest

444 Random forest constructs an ensemble of T decision trees from bootstrap samples while
445 considering a random subset of predictors at each split [96, 105]. For classification, the class-
446 probability estimate can be written as the average of tree-level class indicators or probabilities,

447
$$\hat{p}_k(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T \hat{p}_{tk}(\mathbf{x}), \quad k = 1, \dots, K,$$

448 with point prediction

449
$$\hat{y} = \operatorname{argmax}_k \hat{p}_k(\mathbf{x}).$$

450 Bootstrap aggregation and randomized feature selection reduce correlation among trees and
451 generally improve stability relative to a single decision tree.

452 XGBoost

453 XGBoost is a regularized gradient-boosted decision-tree method in which new trees are added
454 sequentially to minimize an additive objective [97]. It combines regularization with computational
455 strategies for scalable tree boosting and sparse data. At boosting iteration t , the objective may be
456 expressed as

457
$$\mathcal{L}^{(t)} = \sum_{i=1}^n \ell \left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) \right) + \Omega(f_t),$$

458 where

459
$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

460 penalizes tree complexity, T is the number of terminal leaves, and w denotes leaf weights.

461 LightGBM

462 LightGBM is a gradient-boosted decision-tree framework designed for computational efficiency
463 and scalability. It uses histogram-based split finding and a leaf-wise tree-growth strategy. The
464 original method also introduced Gradient-based One-Side Sampling (GOSS) and Exclusive
465 Feature Bundling (EFB) to reduce computational cost for large or high-dimensional data [98]. Its
466 predictive objective is based on gradient boosting, while its main methodological distinction lies
467 in the efficient construction of trees.

468 *Table 2. Selected optimal hyperparameters for each model and target formulation, identified from the prespecified search grid*
469 *using stratified five-fold cross-validation on the training data only. C denotes regularization strength; γ denotes the kernel*
470 *coefficient; α denotes the L2 regularization penalty; max_depth denotes maximum tree depth; $n_estimators$ denotes the number*
471 *of trees; $learning_rate$ denotes the boosting learning rate; and num_leaves denotes the maximum number of leaves in LightGBM.*

Target	Model	Selected hyperparameters
Binary	Logistic Regression	$C = 1$
	SVM	$C = 1$; $\gamma = \text{auto}$
	Neural Network	$\alpha = 0.0001$
	Random Forest	$\text{max_depth} = \text{None}$; $n_estimators = 200$
	XGBoost	$\text{learning_rate} = 0.2$; $\text{max_depth} = 5$; $n_estimators = 50$
	LightGBM	$\text{learning_rate} = 0.1$; $n_estimators = 50$; $\text{num_leaves} = 31$
Multiclass	Logistic Regression	$C = 10$
	SVM	$C = 1$; $\gamma = \text{scale}$
	Neural Network	$\alpha = 0.001$
	Random Forest	$\text{max_depth} = 10$; $n_estimators = 200$
	XGBoost	$\text{learning_rate} = 0.2$; $\text{max_depth} = 5$; $n_estimators = 50$
	LightGBM	$\text{learning_rate} = 0.01$; $n_estimators = 200$; $\text{num_leaves} = 31$

472

473 Hyperparameter tuning and class imbalance

474 Hyperparameters were selected using stratified five-fold cross-validation conducted on the training
475 data only. Balanced accuracy was used as the tuning criterion because it assigns equal importance
476 to class-specific recall and is therefore less dominated by the majority class when class frequencies
477 are unequal [106]. The calibration and test observations were excluded from all hyperparameter
478 selection to preserve the validity of subsequent performance assessment and conformal calibration.
479 The selected hyperparameter values for each model and target formulation are reported in Table 2.

480 The prespecified search grid was:

- 481
- Logistic regression: $C \in \{0.01, 0.1, 1, 10\}$.

- 482 • SVM (RBF): $C \in \{0.1, 1, 10\}$ and $\gamma \in \{\text{scale}, \text{auto}\}$.
- 483 • Neural network: L_2 penalty $\alpha \in \{0.0001, 0.001, 0.01\}$.
- 484 • Random forest: $n_{\text{estimators}} \in \{50, 100, 200\}$ and $\text{max_depth} \in \{5, 10, \text{None}\}$.
- 485 • XGBoost: learning rate $\eta \in \{0.01, 0.1, 0.2\}$, $\text{max_depth} \in \{3, 5, 7\}$, and $n_{\text{estimators}} \in$
486 $\{50, 100, 200\}$.
- 487 • LightGBM: learning rate $\eta \in \{0.01, 0.1, 0.2\}$, $\text{num_leaves} \in \{31, 50, 70\}$, and $n_{\text{estimators}} \in$
488 $\{50, 100, 200\}$.

489 Class imbalance was addressed using inverse-frequency class or observation weights, as supported
490 by the relevant estimator. Under the commonly used balanced weighting rule, the weight for class
491 k is

$$492 \quad w_k = \frac{n}{Kn_k},$$

493 where n is the number of training observations, K is the number of classes, and n_k is the number
494 of training observations in class k . Weights were determined from the training data only.

495 Predictive performance metrics

496 Performance on the held-out test set was summarized using accuracy, balanced accuracy, macro-
497 averaged F1 score, multiclass ROC-AUC, and logarithmic loss. The combination of label-based
498 and probability-based metrics is preferable to reliance on a single measure, particularly for
499 multiclass data with unequal class frequencies [107].

500 Accuracy

501 For multiclass classification, accuracy is

502

$$\text{Accuracy} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}(\hat{y}_i = y_i).$$

503

Although intuitive, overall accuracy can be dominated by the majority class when the class

504

distribution is imbalanced.

505

Balanced accuracy

506

Balanced accuracy is the unweighted mean of class-specific recalls,

507

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FN_k}.$$

508

This gives each class equal contribution regardless of prevalence [106].

509

Macro-averaged F1 score

510

For class k ,

511

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k},$$

512

and

513

$$F1_k = 2 \frac{\text{Precision}_k \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}.$$

514

The macro-averaged F1 score is

515

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K F1_k.$$

516

Macro averaging prevents large classes from mechanically dominating the summary [107].

517 Multiclass ROC-AUC

518 For binary classification, the area under the receiver operating characteristic curve can be
519 expressed as

520
$$\text{AUC} = \int_0^1 \text{TPR}(u) d\text{FPR}(u).$$

521 For the multiclass setting, class-specific one-vs-rest AUC values were computed and macro-
522 averaged across classes so that each class contributed equally [108]. The averaging convention
523 must be stated explicitly because macro and prevalence-weighted AUC answer different questions.

524 Logarithmic loss

525 Multiclass logarithmic loss was calculated as

526
$$\text{Log Loss} = -\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \sum_{k=1}^K y_{ik} \log(\hat{p}_{ik}),$$

527 where $y_{ik} = 1$ if observation i belongs to class k and 0 otherwise, and $\hat{p}_{ik}$ is the predicted
528 probability for class k . Lower log loss indicates better probabilistic prediction under the
529 logarithmic scoring rule. It should not be interpreted as a pure measure of calibration because it
530 reflects both the probability assigned to the realized class and the sharpness of the predictive
531 distribution.

532 Split conformal prediction for multiclass classification

533 Statistical guarantee

534 Conformal prediction converts the output of an arbitrary predictive model into prediction sets with
535 finite-sample marginal coverage under exchangeability [31, 33, 81]. Let (X_i, Y_i) denote
536 exchangeable observations and let $C_\alpha(X_{n+1})$ be a conformal prediction set for a future observation.
537 Split conformal prediction is constructed so that

$$538 \quad \Pr\{Y_{n+1} \in C_\alpha(X_{n+1})\} \geq 1 - \alpha,$$

539 where $\alpha \in (0,1)$ is the target miscoverage level. This is a *marginal* guarantee over the joint
540 randomness of the calibration and future observations; it does not imply exact coverage within
541 every class or covariate-defined subgroup.

542 Nonconformity score

543 For a fitted classifier that returns class probabilities $\hat{p}_y(\mathbf{x})$, the inverse-probability nonconformity
544 score for calibration observation $(\mathbf{x}_i, y_i)$ was

$$545 \quad s_i = 1 - \hat{p}_{y_i}(\mathbf{x}_i).$$

546 Smaller values indicate that the fitted classifier assigned greater probability to the observed class.
547 This score is simple and model agnostic, although more adaptive classification scores are available
548 when set efficiency or conditional behavior is a primary objective.

549 Finite-sample conformal quantile

550 Let $m = |\mathcal{D}_{\text{cal}}|$ and let

$$s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(m)}$$

denote the ordered calibration scores. The split-conformal threshold must use the finite-sample corrected order statistic

$$k = \lceil (m + 1)(1 - \alpha) \rceil,$$

$$\hat{q}_\alpha = s_{(k)},$$

provided $k \leq m$; equivalently, software may compute the empirical quantile at level $\lceil (m + 1)(1 - \alpha) \rceil / m$ using the conservative higher-order-statistic convention. This correction is important. Replacing it with an ordinary empirical $(1 - \alpha)$ quantile can lose the stated finite-sample coverage guarantee.

Prediction-set construction

For a test observation x , the prediction set was

$$C_\alpha(x) = \{y \in \mathcal{Y} : 1 - \hat{p}_y(x) \leq \hat{q}_\alpha\},$$

or, equivalently,

$$C_\alpha(x) = \{y \in \mathcal{Y} : \hat{p}_y(x) \geq 1 - \hat{q}_\alpha\}.$$

Prediction sets were evaluated at nominal coverage levels 0.90, 0.95, and 0.99, corresponding to $\alpha = 0.10, 0.05$, and 0.01 , respectively.

567 Evaluation of conformal prediction sets

568 Coverage alone is insufficient for comparing conformal predictors because a trivial set containing
569 every class has perfect coverage but no practical informativeness. We therefore jointly assessed
570 empirical validity and efficiency.

571 Empirical coverage

572 Empirical marginal coverage on the test set was

573
$$\widehat{\text{Coverage}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}\{y_i \in \mathcal{C}_\alpha(x_i)\}.$$

574 The theoretical target is $1 - \alpha$. Because empirical coverage is itself estimated from a finite test
575 sample, an observed value slightly below the nominal level in one split does not by itself imply
576 failure of the theoretical conformal guarantee.

577 Average prediction-set size

578 Efficiency was summarized by

579
$$\text{Average Set Size} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}_\alpha(x_i)|.$$

580 Among methods with comparable coverage, smaller average sets are more informative.

581

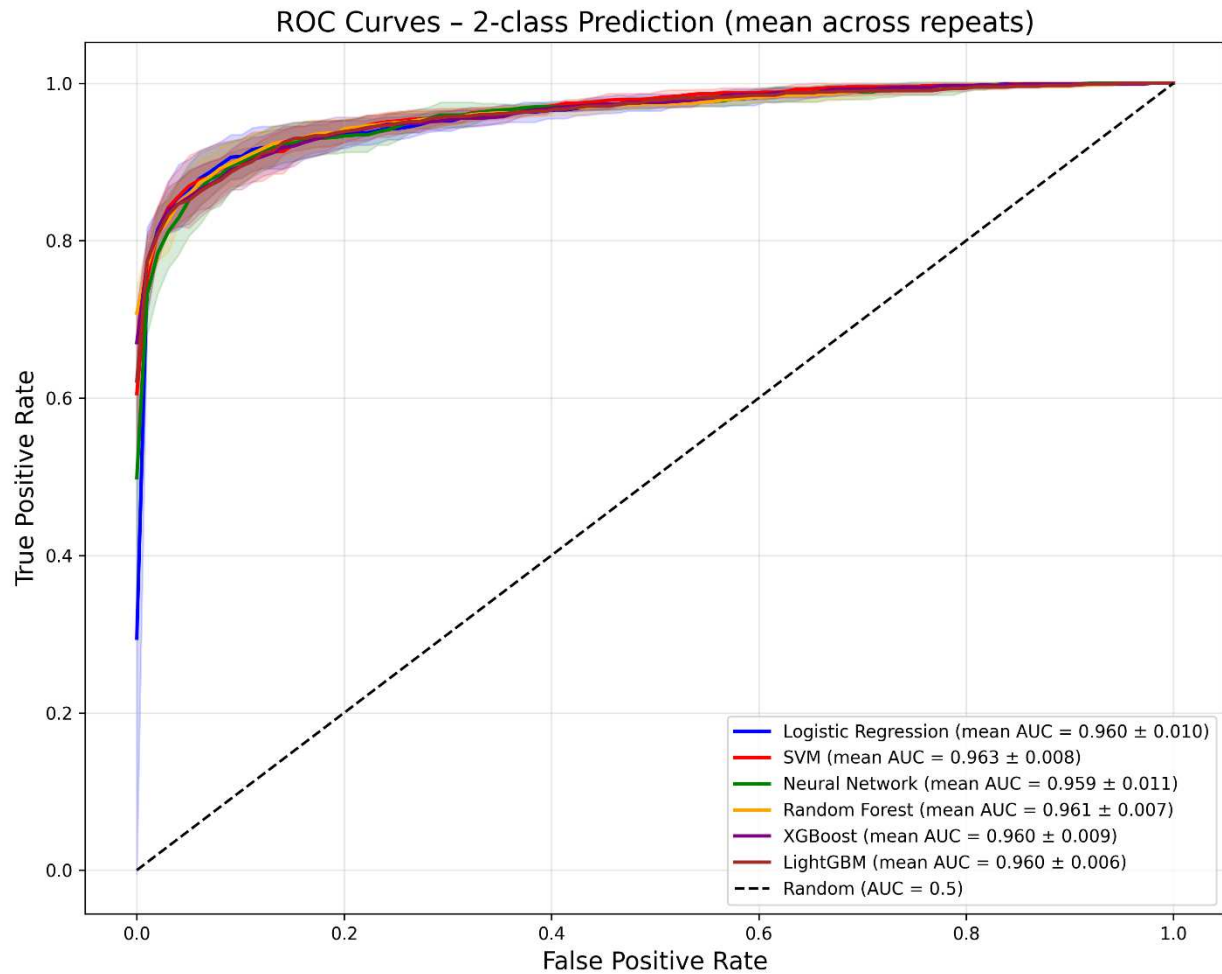


Figure 2. Receiver operating characteristic (ROC) curves for the binary classification task distinguishing students who graduated from those who dropped out.

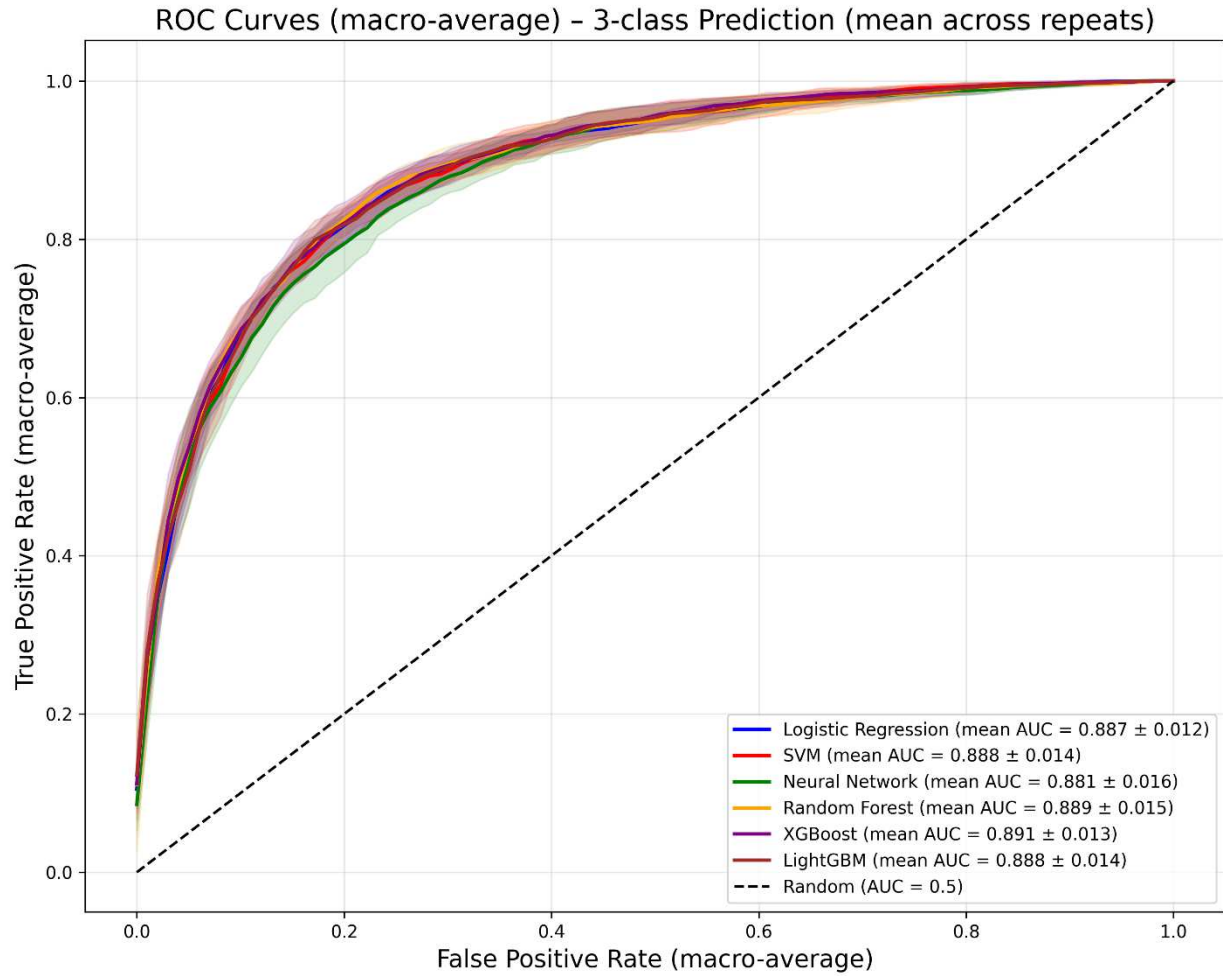


Figure 3. Receiver operating characteristic (ROC) curves for the multiclass classification task involving Graduate, Dropout, and Enrolled outcomes.

598 *Table 3. Baseline model performance for the 2-class and 3-class classification tasks. Values are means across five repeated random*
599 *splits. ROC-AUC for the 3-class task is reported using a multiclass one-vs-rest formulation.*

Target	Model	Accuracy	Balanced accuracy	F1-macro	ROC-AUC	Log loss
2-Class	Logistic Regression	0.8936	0.8832	0.8871	0.9287	0.2981
	SVM	0.8917	0.8809	0.8850	0.9410	0.2683
	Neural Network	0.8917	0.8741	0.8831	0.9278	0.3001
	Random Forest	0.8972	0.8770	0.8882	0.9412	0.2721
	XGBoost	0.9009	0.8817	0.8924	0.9416	0.2730
	LightGBM	0.9064	0.8946	0.9002	0.9353	0.2794
3-Class	Logistic Regression	0.7620	0.7343	0.7227	0.9160	0.5802
	SVM	0.7410	0.7126	0.7013	0.9209	0.5269
	Neural Network	0.7997	0.7269	0.7364	0.9165	0.5230
	Random Forest	0.7937	0.7514	0.7496	0.9254	0.5376
	XGBoost	0.7892	0.7111	0.7185	0.9198	0.5184
	LightGBM	0.7771	0.7605	0.7453	0.9243	0.5763

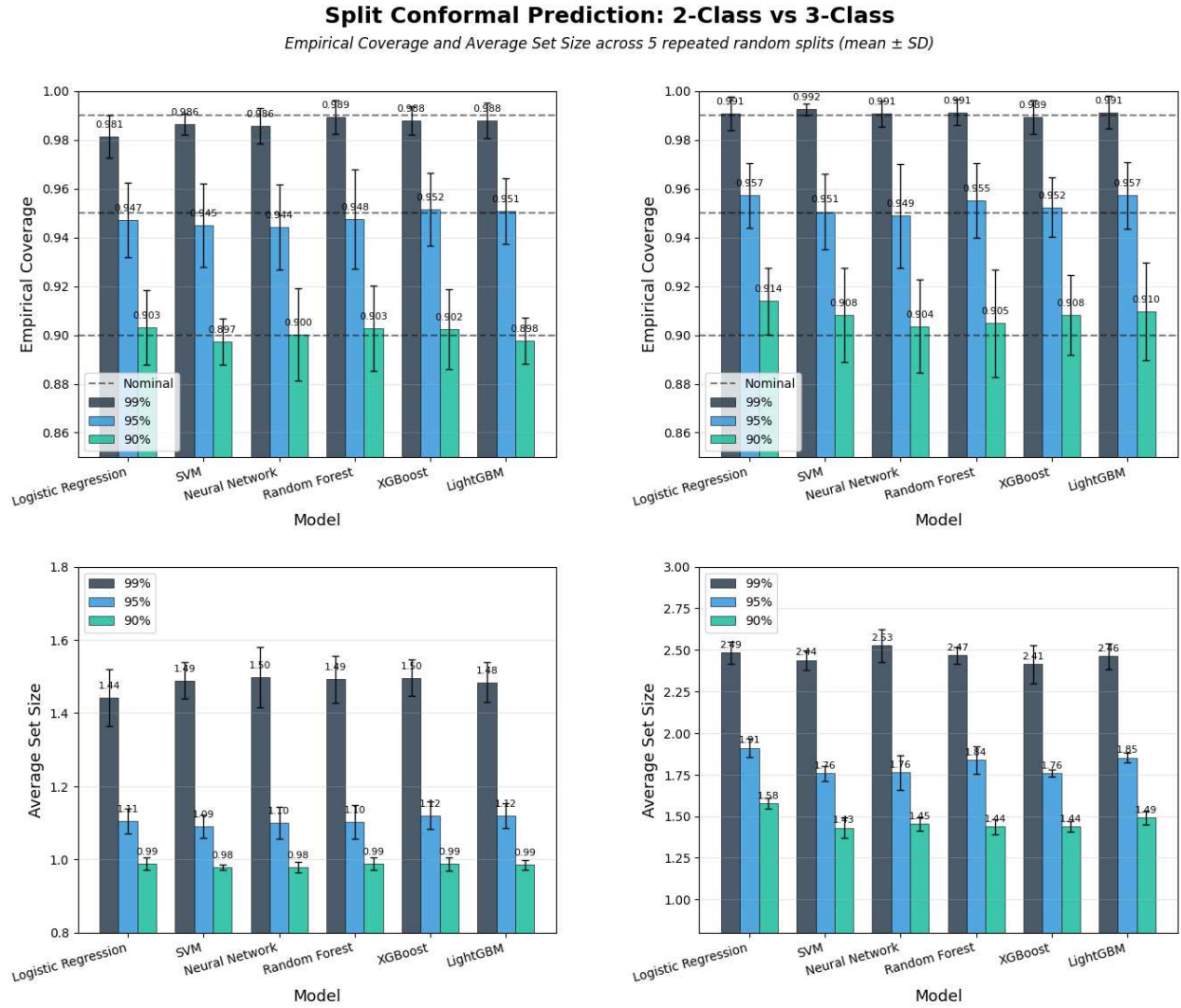


Figure 4. Conformal prediction results across nominal confidence levels for the binary and multiclass classification tasks. The figure summarizes empirical coverage and average prediction set size, illustrating the trade-off between predictive reliability and prediction-set efficiency.

Results

Baseline Model Evaluation

Table 3 summarizes baseline performance across the six models for the binary and multiclass classification tasks. In the binary task, all models performed well, with accuracy ranging from 0.8917 to 0.9064. LightGBM showed the strongest overall classification profile, achieving the highest accuracy (0.9064), balanced accuracy (0.8946), and macro-F1 (0.9002). XGBoost and Random Forest were similarly competitive in discrimination, with XGBoost yielding the highest ROC-AUC (0.9416) and Random Forest showing a nearly identical ROC-AUC (0.9412). These small differences suggest that several ensemble models captured the main signal in the data, rather than one model clearly dominating all criteria.

The multiclass task was more challenging, with accuracy ranging from 0.7410 to 0.7997 and macro-F1 ranging from 0.7013 to 0.7496. The Neural Network achieved the highest raw accuracy (0.7997), whereas LightGBM had the best balanced accuracy (0.7605) and Random Forest produced the highest macro-F1 (0.7496) and ROC-AUC (0.9254). Because balanced accuracy and macro-F1 are less influenced by class imbalance, these results indicate that Random Forest and LightGBM offered the most stable multiclass performance, with no single model uniformly outperforming the others.

Figures 2 and 3 show the ROC curves for the binary and multiclass tasks, respectively. The figures support the tabulated results by showing strong overall discrimination, particularly for the ensemble models. The binary task curves indicate clearer separation between Graduate and

Dropout outcomes, whereas the multiclass curves reflect the greater difficulty of distinguishing the Enrolled group from the two terminal outcomes.

Overall, LightGBM was retained as a key model for further interpretation because it combined strong binary classification performance, the best balanced accuracy in the multiclass task, and efficient handling of mixed feature types. Random Forest and XGBoost remained important comparators because they performed similarly well on several discrimination and calibration-related metrics.

Uncertainty Quantification via Conformal Prediction

Split conformal prediction was used to quantify prediction uncertainty for all models. Supplementary Table 2 and Figure 4 summarize empirical coverage and average prediction-set size across the 99%, 95%, and 90% confidence levels.

For the binary task, prediction sets were highly efficient, with average set sizes close to one across most models and confidence levels. Coverage was generally close to nominal levels, although several models showed slight under-coverage at the highest confidence level. XGBoost and LightGBM provided a favorable balance between valid coverage and compact prediction sets at the 95% level.

For the multiclass task, valid coverage was achieved for most model-confidence combinations, but average prediction sets were larger than in the binary task. This pattern reflects the greater ambiguity of the Enrolled category, which is less clearly separated from Graduate and Dropout outcomes. Random Forest and LightGBM showed particularly balanced performance when both classification accuracy and conformal validity were considered.

Taken together, the results suggest that LightGBM is a strong primary choice for this application because it performed consistently well across predictive accuracy, balanced class-wise performance, and interpretability. However, the differences among the leading ensemble models were modest. Therefore, Random Forest and XGBoost should be viewed as competitive alternatives, especially when the priority is multiclass discrimination or conformal-set efficiency.

Feature Importance Analysis

Model interpretation was conducted for the leading ensemble models to identify the main drivers of student outcome prediction. The SHAP summary plot in Figure 5 and the feature-importance plot in Figure 6 both showed that academic progression carried the strongest predictive signal.

The most influential features were second-semester approved curricular units, average approved units, grades, and admission performance, indicating that accumulated academic progress was central to the model decisions. Financial features also contributed meaningfully. Being up to date on tuition payments tended to move predictions toward more favorable outcomes, whereas greater financial stress shifted predictions toward higher dropout risk. Scholarship status showed additional separation, suggesting that financial context refined risk classification beyond academic performance alone. Because several dominant predictors are observed after enrollment, these findings support risk updating during the academic year rather than prediction at entry.

The direction of the SHAP effects was consistent with the study context. Higher academic achievement and stronger credit completion were associated with lower predicted dropout risk, while lower academic progression and greater financial vulnerability increased risk.

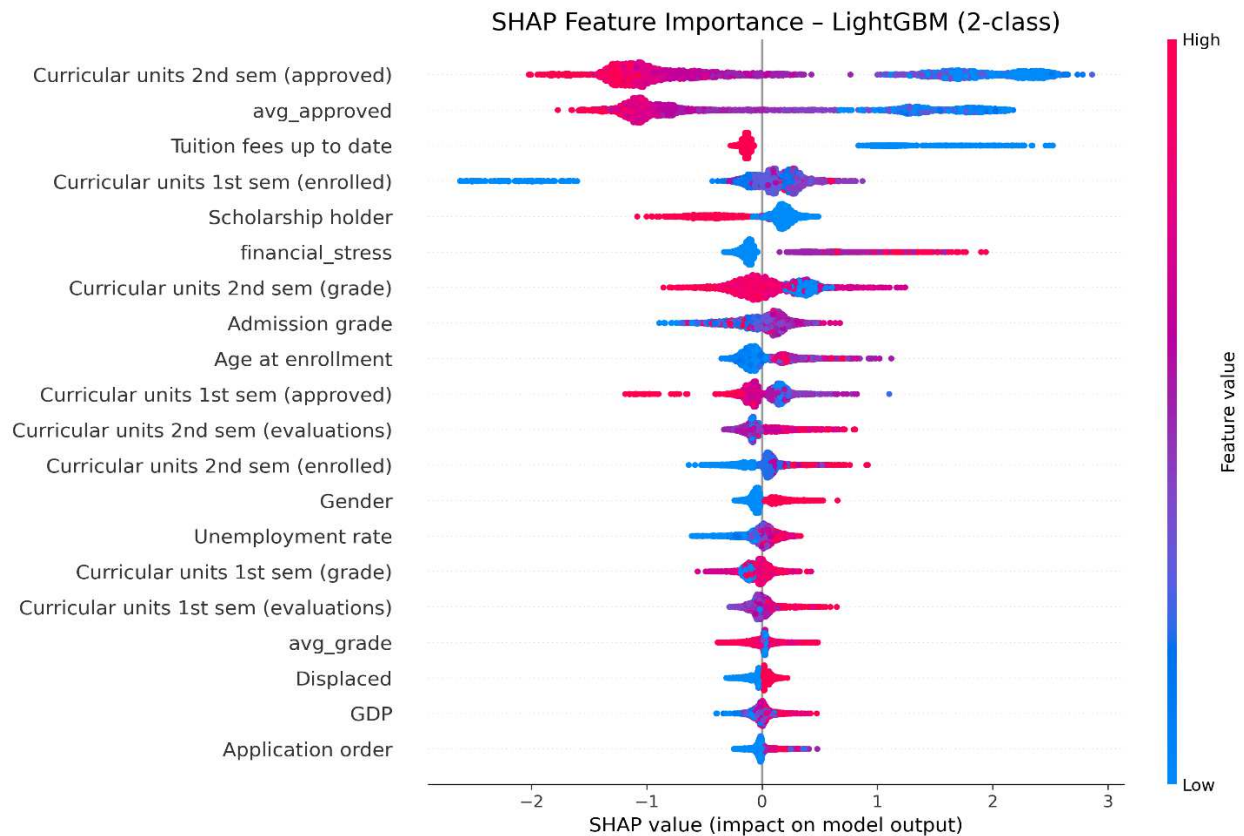


Figure 5. SHAP summary plot showing the relative contribution of predictor features to model predictions. Features are ranked by overall importance, with SHAP values indicating both the magnitude and direction of each feature's influence on predicted student outcomes.

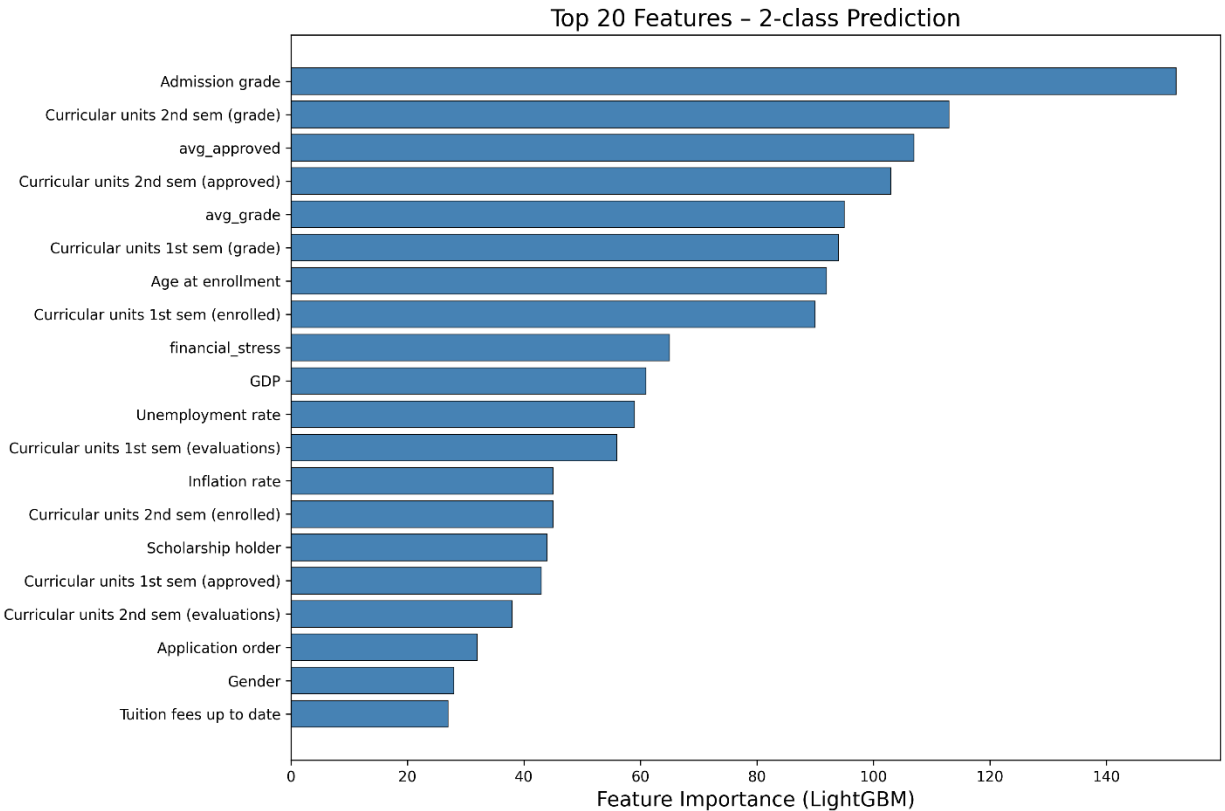


Figure 6. Feature importance plot for the LightGBM model. Features with higher importance values contributed more strongly to model decision-making, highlighting the academic, financial, demographic, and macroeconomic predictors most relevant to student outcome classification.

Summary of Key Findings

Overall, the findings show that machine learning models can distinguish student outcomes with good accuracy while conformal prediction adds an interpretable measure of uncertainty. The binary Graduate vs Dropout task was more stable and produced more compact prediction sets than the multiclass task, where the Enrolled category introduced greater ambiguity.

Among the evaluated models, LightGBM offered the most consistent overall profile, particularly for binary classification and balanced multiclass performance. Random Forest and XGBoost

remained close competitors, with Random Forest performing strongly in the multiclass setting and XGBoost showing strong discrimination and efficient conformal behavior in selected settings. The interpretability analyses further indicated that academic progression and financial status were the strongest contributors to prediction, suggesting that early academic performance and financial vulnerability are central signals for identifying students who may need support.

Discussion

Positioning the predictive results within prior research

With respect to RQ1, the findings are consistent with previous studies on student dropout and academic outcome prediction, while adding an uncertainty-aware and interpretable modeling framework. In the binary Graduate versus Dropout task, the leading ensemble models performed similarly well. LightGBM achieved the highest accuracy (0.9064), balanced accuracy (0.8946), and macro-F1 (0.9002), while XGBoost had the highest ROC-AUC (0.9416). These values fall within the range reported by earlier studies of the same dataset (Table 1). They are higher than those reported by Ridwan et al. for XGBoost (accuracy approximately 0.87; F1 approximately 0.81) [25], comparable to those of Romero and Liao, who used the same Graduate versus Dropout formulation (XGBoost F1 0.907; Random Forest ROC-AUC 0.935) [14], and slightly below the ROC-AUC of 0.957 reported by Liu et al., who used five-fold cross-validation and a dropout versus non-dropout definition that merges the Enrolled and Graduate students [27]. They are also consistent with the 86–88% accuracies that dominate the broader literature [59] and with the strong performance of boosting and ensemble methods reported elsewhere [2, 13, 15, 22–24]. Because outcome definitions, preprocessing, class balancing, tuning, and validation designs differ, these

comparisons are indicative rather than strict. Nevertheless, the overall pattern is clear: once second-semester academic information is available, several tree-based ensembles reach a similar performance ceiling, and the small differences among them should not be overinterpreted [26, 60]. By averaging results over five repeated splits, the present study provides more stable estimates than single-split evaluations, which can be sensitive to the particular partition used.

The multiclass task was more difficult, mainly because the Enrolled group represents an unresolved academic trajectory. Neural Network had the highest overall accuracy (0.7997), LightGBM had the highest balanced accuracy (0.7605), and Random Forest had the highest macro-F1 (0.7496) and multiclass ROC-AUC (0.9254). These results exceed the macro-F1 of approximately 0.70–0.72 and macro ROC-AUC of approximately 0.88–0.90 reported by Morsu using only enrollment-time information [28], which is consistent with the finding of Martins et al. that predictions improve once first-semester performance becomes available [61]. Similar difficulties with the Enrolled category have been reported by Alameri [23] and Morsu [28], and Martins et al. [15] showed that strong overall model performance can coexist with weaker classification of a smaller or less distinct class. Previous studies have generally treated this difficulty as a shortcoming of the classifier. The theoretical literature suggests an alternative interpretation: if departure is a longitudinal process [17, 43], some still-enrolled students are not yet distinguishable from future graduates or future dropouts on the basis of the available information. In such cases, the appropriate model output is not a more confident label but an honest statement that several outcomes remain plausible.

The main predictors also matched prior evidence and theory. SHAP results indicated that academic progression, especially second-semester approved curricular units, average approved units, grades, and admission performance, had the strongest influence on prediction. Financial factors, including

tuition-payment status, financial stress, and scholarship status, added meaningful information beyond academic performance. These findings align with those of Romero and Liao [14], Arora [13], Alameri [23], Villar and de Andrade [22], and Liu et al. [27], as well as with syntheses identifying credits and grades as the most consistent predictors of persistence [47]. They also correspond to the two pillars of classic attrition theory: academic integration, reflected in curricular progression, and the environmental and financial pressures emphasized by Bean and Metzner and by Cabrera et al. [45, 46]. Academic progression is therefore the clearest proximal signal of student outcome, while financial vulnerability adds a practically relevant and potentially modifiable dimension of risk.

What conformal prediction adds to dropout prediction

Regarding RQ2, the central difference between this study and the work summarized in Table 1 is that the present framework communicates predictive reliability for each student. Many student-risk models provide class labels or probabilities without clearly communicating how reliable they are, so that two students may receive the same high-risk label even when the evidence behind those predictions differs substantially. Split conformal prediction addresses this limitation by producing prediction sets with a prespecified marginal coverage target under the exchangeability assumption [34–36], allowing the model to identify cases in which one outcome is strongly supported and cases in which multiple outcomes remain plausible.

The conformal results showed compact prediction sets for the binary task. At 95% nominal coverage, XGBoost achieved empirical coverage of 0.952 with an average set size of 1.121, while LightGBM achieved coverage of 0.951 with an average set size of 1.120. Because a binary set contains at most two labels, these averages imply that at least 9–12% of students, depending on

the model, received the two-label set {Graduate, Dropout}, whereas a single outcome was supported at the 95% level for nearly all of the remaining students. This is information that accuracy alone cannot provide: an accuracy of 0.906 indicates that roughly one prediction in ten is wrong, but not which one, whereas the conformal sets identify a specific group of students for whom the evidence is insufficient. At the 90% level, average set sizes were slightly below one (0.979–0.989), implying that at least 1–2% of students received empty sets, meaning that no outcome was sufficiently consistent with the calibration data at that level; such cases can be used to flag atypical records or possible data errors. Empirical coverage did not meet the nominal target uniformly across models and levels. At the 99% level, all binary models showed slight undercoverage (0.981–0.989), while at the 95% level only XGBoost and LightGBM met the nominal target; at the 90% level, LightGBM and SVM fell slightly below it. These small deviations should be interpreted in the context of finite-sample variability in test-set coverage rather than as evidence that the conformal framework failed. Overall, the results show the trade-off between target coverage and set efficiency: moving from 95% to 99% coverage increased the average binary set size from approximately 1.1 to approximately 1.5, implying that at least 44–50% of students would receive an undetermined set.

In the multiclass task, conformal prediction made the ambiguity around the Enrolled category explicit. At 95% coverage, average set sizes were 1.758 for XGBoost, 1.839 for Random Forest, and 1.852 for LightGBM, which implies that, for these three models, at least 38–43% of students received sets containing two or three outcomes. Three-class accuracies of 0.74–0.80 in the present study would conceal this fact: roughly one label in four or five would be wrong, without any indication of which. Instead of forcing each student into one label, the prediction sets show when more than one trajectory remains plausible. This provides an operational counterpart to the process

view of dropout and is particularly useful for early-warning systems, in which ambiguous sets can signal the need for advisor review before any intervention is assigned.

Compared with other approaches to uncertainty in educational prediction, conformal prediction has several practical advantages. Unlike Bayesian deep learning [78], it does not require a specific model class or prior specification; unlike bootstrap analyses of aggregate metrics [60], it provides a statement about each individual prediction; and unlike probability calibration alone [79], it offers an explicit finite-sample coverage guarantee under exchangeability. Elmasry et al. recently showed that split conformal prediction can be added to a calibrated dropout classifier with near-nominal coverage [92]. The present results extend that evidence by showing, across six model families, two outcome formulations, and three coverage levels, that validity is largely model-agnostic whereas efficiency is not: models with similar accuracy produced different average set sizes. For institutions, this suggests that a model intended for an early-warning system should be selected not only on accuracy or ROC-AUC, but also on the size of the conformal sets it produces at the coverage level the institution chooses. Evidence from human-subject experiments, in which prediction sets improved human decisions relative to single labels [86–88], further suggests that sets may be better suited to supporting advisors than single risk scores, although their benefit depends on keeping sets small and on the difficulty of the cases [89].

Combining uncertainty with explanation

Regarding RQ3, combining conformal prediction with SHAP also changes how explanations should be used. A detailed explanation can create false precision if the prediction itself is uncertain. A more responsible workflow would first examine the prediction set and then use SHAP explanations for confident predictions, or to guide the review of ambiguous cases, for example by

showing whether ambiguity arises from mixed academic signals or from missing financial information. This approach is better suited to high-stakes educational decisions than relying only on accuracy, probabilities, or feature-importance rankings, and it complements governance-oriented explanation pipelines that surface ambiguous cases for human review [56, 63].

The SHAP analysis showed which features the fitted models used and the direction of their contributions, but it did not establish that modifying those features would change student outcomes. This distinction is important for educational decision-makers: an explanation can support case review, yet it cannot by itself justify an intervention. Liu et al. [27] similarly showed that SHAP can improve interpretability in dropout prediction. The contribution of the present study is therefore the joint presentation of model explanations and conformal uncertainty, not evidence that the identified predictors are causal intervention targets. For example, the role of tuition-payment status or financial stress should not be interpreted as proof that a specific financial intervention would reduce dropout by a corresponding amount. Romero and Liao [14] used propensity-score methods to estimate the effect of scholarship holding, underscoring that causal claims require designs beyond predictive modeling. In this study, SHAP is best viewed as a tool for understanding model predictions and identifying potentially actionable domains, not as evidence of causal impact.

Implications for institutional policy and practice

The results have several implications for institutional leaders, student-support services, and policymakers. First, conformal outputs can be translated directly into a tiered decision protocol (Table 4). Students whose set contains only Dropout can be prioritized for proactive outreach, such as academic advising, tutoring, and financial counseling. Students whose set contains only

Graduate can remain under routine monitoring without being labeled. Students whose set contains several outcomes, including Dropout, should be referred for human review, in which advisors collect additional information, such as attendance, learning management system activity, or a brief conversation, before deciding on an intervention. Empty sets should trigger a data-quality check. This protocol operationalizes the recommendation that machine learning should serve as decision support rather than as an autonomous decision-maker [59, 69], and it directs scarce advising time toward the students for whom human judgment adds the most value.

Table 4. A proposed tiered decision protocol that translates conformal prediction sets into institutional actions. The proportions of students in each tier depend on the chosen coverage level and model; the percentages shown are lower bounds derived from the average set sizes in the present study (95% coverage unless otherwise stated).

Prediction set	Interpretation	Suggested institutional action
{Dropout}	Dropout is the only outcome supported at the chosen coverage level	Prioritize proactive outreach: academic advising, tutoring, and financial counseling; review SHAP explanation to tailor support
{Graduate}	Graduation is the only supported outcome	Routine monitoring; no risk label communicated
{Enrolled} or {Enrolled, Graduate}	Continuing trajectory with no dropout signal	Monitor progression; light-touch check-ins at the next milestone
Sets including Dropout and another outcome ($\geq 9\text{--}12\%$ of binary cases; $\geq 38\text{--}43\%$ of three-class cases include two or more outcomes)	Evidence is insufficient to distinguish trajectories	Refer to advisor for human review; collect additional information before assigning an intervention
Empty set ($\approx 1\text{--}2\%$ of binary cases at 90%)	No outcome is consistent with the calibration data	Check data quality and atypical profiles; do not act on the model output

Second, the coverage level becomes an explicit and transparent policy lever. Choosing $1 - \alpha$ is equivalent to choosing how often the institution is willing to miss the true outcome, and the corresponding set sizes reveal the advising workload that this choice implies. In the binary task, moving from 90% to 99% coverage increased the average set size from approximately 0.99 to approximately 1.49, so that at least 44–50% of students would require review at the highest level. Institutions can therefore align the coverage level with their advising capacity and with the relative

costs of false reassurance and unnecessary intervention, rather than relying on an arbitrary probability threshold whose reliability is unknown.

Third, the prominence of tuition-payment status, financial stress, and scholarship status suggests that financial monitoring should be integrated into retention strategies. Overdue tuition payments are recorded administratively and could act as early triggers for contact with financial-aid offices, emergency grants, or flexible payment plans. This is consistent with attrition theory [45, 46] and with the quasi-causal evidence that scholarship holding is associated with a lower probability of dropout in these data [14], although the causal effects of specific financial interventions should be evaluated directly.

Fourth, because the dominant predictors are second-semester measures, the framework is best used as an in-year updating tool rather than as an admission-time screen. Institutions should generate predictions at several points, for example at enrollment, after the first semester, and after the second semester, and recalibrate the conformal threshold at each point, consistent with evidence that prediction quality changes over the first academic year [54, 61].

Fifth, uncertainty-aware prediction should be embedded in equity-oriented governance. Marginal coverage can conceal lower coverage or larger sets for particular groups, and the literature shows that educational prediction models can disadvantage historically underserved students [73, 75, 76]. Institutions should therefore audit coverage and set size by gender, scholarship status, displacement, international status, and other relevant characteristics; consider group-conditional or Mondrian conformal procedures where equal reliability is a substantive objective [82–84]; and monitor downstream decisions, since even well-calibrated sets can lead to disparate actions [90]. Students should be informed about how predictions are used, and communication should avoid deficit-based labels [69, 70].

Finally, at the system level, funding bodies, quality-assurance agencies, and ministries that encourage or procure predictive analytics could require vendors and institutions to report measures of predictive reliability, such as empirical coverage and set size, alongside accuracy, and to evaluate early-warning systems with experimental or quasi-experimental designs. The mixed evidence on the effectiveness of early-alert systems [67, 68] shows that accurate prediction does not automatically translate into improved retention. Uncertainty-aware, explainable, and evaluated decision support is more likely to contribute to equitable completion, in line with the aims of Sustainable Development Goal 4 on inclusive and equitable quality education.

Limitations and future directions

Several limitations should be noted. First, the data come from a single Portuguese higher education institution and cover students enrolled between 2008/2009 and 2018/2019. Institutional policies, tuition systems, academic programs, and student populations may differ across settings. As noted by Romero and Liao [14] and broader reviews [26], external validity remains a key challenge. The repeated random splits used here improve internal stability but do not establish transportability to other institutions or future cohorts.

A definitive assessment of generalizability requires external validation using data from other institutions with different programs, admission systems, financing structures, and student populations. Until such data are available, the present estimates should be treated as institution specific. A future multi-institution study should retain institution-level holdouts, recalibrate conformal thresholds locally where appropriate, and report both discrimination and coverage before any deployment claim is made.

Second, several important predictors were first- and second-semester academic-performance measures. These features improve accuracy but may limit the opportunity for earlier intervention, as Liu et al. [27] also noted. Future studies should evaluate prediction at clear time points, such as enrollment, early semester, first semester, and second semester, and incorporate longitudinal indicators from learning-management systems, assessments, attendance, and advising records [2, 28].

Third, our conformal method provides marginal coverage, not equal coverage for every class or subgroup. Marginal guarantees can coexist with undercoverage of minority classes or of disadvantaged student groups, an issue with direct equity implications [73, 84]. Future work should report class-specific and subgroup-specific coverage and set size, and consider class-conditional, Mondrian, or conditional conformal methods [82, 83] when equal reliability across Graduate, Dropout, and Enrolled students, or across demographic groups, is a substantive goal.

Fourth, split conformal prediction requires a separate calibration set, reducing the data available for model training. With 4,424 observations, this reduction may be meaningful. Future work could examine cross-conformal or cross-validation-based conformal procedures, Jackknife+ methods [109], and probability-calibration approaches such as Platt scaling, isotonic regression, or temperature scaling before conformalization.

Another limitation is the exchangeability assumption underlying the reported finite-sample coverage guarantee. The dataset spans ten academic years, during which student composition, programs, institutional policies, and economic conditions may have changed. Random train-calibration-test splits mix observations across cohorts and can therefore obscure temporal distribution shift. The marginal coverage results should be interpreted as valid only to the extent that calibration and future observations are exchangeable. Temporal holdout analyses, year-

specific coverage estimates, weighted conformal prediction under covariate shift, and online or adaptive conformal procedures are needed to evaluate and mitigate departures from this assumption.

In addition, the comparisons with previous studies presented in Table 1 and in the Discussion are indirect. The cited studies differ in outcome definitions, inclusion of the Enrolled class, preprocessing, class-imbalance handling, tuning, and validation design, and some are preprints or theses. A shared benchmarking protocol for this dataset, including standardized splits and the reporting of uncertainty metrics, would allow more rigorous comparisons.

Finally, predictive accuracy and coverage do not prove that an early-warning system improves student outcomes. Future work should evaluate these models prospectively in human-in-the-loop settings, where high-confidence predictions may trigger routine outreach, ambiguous sets may be referred to advisors, and uncertain cases may remain unclassified until more information is available. Such systems should be assessed for fairness, intervention burden, false-positive costs, resource allocation, and actual effects on retention.

Conclusion

This study shows that machine learning can effectively distinguish student academic outcomes, while also emphasizing that predictive performance alone is not sufficient for educational decision support. Tree-based ensemble models performed strongly, particularly for the binary Graduate-versus-Dropout task, whereas the multiclass task was more difficult because the Enrolled category represents a more ambiguous and transitional academic state. Consistent with prior work, academic progression was the dominant predictive signal, while financial vulnerability added

921 meaningful information. SHAP analysis further showed that curricular-unit progression, grades,
922 admission performance, tuition-payment status, and related financial factors contributed
923 substantially to model predictions.

924 The main contribution of this work is the integration of predictive modeling, conformal uncertainty
925 quantification, and SHAP-based explanation within a unified framework, positioned against an
926 extensive body of research on dropout theory, educational data mining, and early-warning systems.
927 Previous studies of the same dataset reached similar levels of accuracy but reported only aggregate
928 point metrics. Instead of providing only a predicted class or probability, conformal prediction
929 produced prediction sets that indicate when the available evidence supports a confident outcome
930 and when multiple outcomes remain plausible. This was especially useful in the multiclass setting,
931 in which larger prediction sets reflected the greater uncertainty involved in distinguishing the
932 Enrolled group. SHAP complemented this information by explaining which academic and
933 financial factors shaped the predictions.

934 Together, these findings support a shift from accuracy-centered prediction toward uncertainty-
935 aware and interpretable decision support. In practice, confident predictions could inform targeted
936 outreach, while ambiguous prediction sets could signal the need for additional information or
937 advisor review rather than forcing potentially unreliable classifications. For policymakers, the
938 coverage level offers a transparent way to balance missed at-risk students against advising
939 workload, and reliability metrics such as coverage and set size should be reported alongside
940 accuracy whenever predictive analytics are used to allocate student support. Future work should
941 validate this framework across institutions and cohorts, examine class- and subgroup-specific
942 coverage, and evaluate whether combining calibrated uncertainty with interpretable predictions
943 improves the fairness, transparency, and effectiveness of student-support interventions.

944 List of abbreviations
945
946 **AI:** Artificial Intelligence
947 **AUC:** Area Under the Curve
948 **CP:** Conformal Prediction
949 **EFB:** Exclusive Feature Bundling
950 **LASSO:** Least Absolute Shrinkage and Selection Operator
951 **LightGBM:** Light Gradient Boosting Machine
952 **ML:** Machine Learning
953 **OECD:** Organization for Economic Co-operation and Development
954 **TPR:** True Positive Rate
955 **FPR:** False Positive Rate
956 **ROC:** Receiver Operating Characteristic
957 **SHAP:** SHapley Additive exPlanations
958 **SVM:** Support Vector Machine
959 **XAI:** Explainable Artificial Intelligence
960 **XGBoost:** Extreme Gradient Boosting
961 **EDM:** Educational Data Mining
962 **EWS:** Early Warning System
963 **LA:** Learning Analytics
964 **RQ:** Research Question
965

966 Declarations

967 Acknowledgments: Not applicable.

968 Funding: This research received no specific grant from any funding agency in the public,
969 commercial, or not-for-profit sectors.

970 Author contributions: Kazi Sabbir Ahmad Nahin conceptualized the study, performed the analysis,
971 interpreted the results, and drafted and revised the manuscript.

972 Competing interests: The author declares no competing interests.

973 Ethical Approval: Not Applicable.

974 Consent to participate: Not Applicable.

975 Consent to publish: Not Applicable.

976 Data Availability: The dataset is publicly available from the UCI Machine Learning Repository:
977 <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>.

978
979 Code availability: Code used for the analyses is available at

980 https://github.com/Nahin277/CP_Project_Graduate.

981 References

- 982 1. Behr A, Giese M, Teguin Kamdjou HD, Theune K. Motives for dropping out from higher
983 education—an analysis of bachelor's degree students in Germany. *Eur J Educ*.
984 2021;56(2):325–43. <https://doi.org/10.1111/ejed.12433>.
- 985 2. Realinho V, Machado J, Baptista L, Martins MV. Predicting student dropout and academic
986 success. *Data*. 2022;7(11):146. <https://doi.org/10.3390/data7110146>.
- 987 3. Casanova JR, Castro-López A, Bernardo AB, Almeida LS. The dropout of first-year STEM
988 students: is it worth looking beyond academic achievement? *Sustainability*.
989 2023;15(2):1253. <https://doi.org/10.3390/su15021253>.
- 990 4. Quinn J. Drop-out and completion in higher education in Europe among students from
991 under-represented groups. An independent report authored for the European Commission.
992 Brussels: NESET; 2013.
- 993 5. Lorenzo-Quiles O, Galdón-López S, Lendínez-Turón A. Factors contributing to university
994 dropout: a review. *Front Educ*. 2023;8:1159864.
995 <https://doi.org/10.3389/educ.2023.1159864>.
- 996 6. Vlasova H. Eye-opening college dropout rates & statistics—2022. *Admissionsly*; 2022.
997 <https://admissionsly.com/college-dropout-rates>.
- 998 7. Hanson M. College dropout rates. *Education Data Initiative*; 2021.
999 <https://educationdata.org/college-dropout-rates>.
- 1000 8. Snyder TD, de Brey C, Dillow SA. Digest of education statistics 2017 (NCES 2018-070).
1001 Washington, DC: National Center for Education Statistics; 2019.
- 1002 9. Raisman N. The cost of college attrition at four-year colleges & universities. Policy
1003 perspectives. Virginia Beach: Educational Policy Institute; 2013.

- 1004 10. Aina C, Baici E, Casalone G, Pastore F. The determinants of university dropout: a review
1005 of the socio-economic literature. *Socioecon Plann Sci.* 2022;79:101102.
1006 <https://doi.org/10.1016/j.seps.2021.101102>.
- 1007 11. Kehm BM, Larsen MR, Sommersel HB. Student dropout from universities in Europe: a
1008 review of empirical literature. *Hung Educ Res J.* 2019;9(2):147–64.
1009 <https://doi.org/10.1556/063.9.2019.1.18>.
- 1010 12. Matz SC, Bukow CS, Peters H, Deacons C, Dinu A, Stachl C. Using machine learning to
1011 predict student retention from socio-demographic characteristics and app-based
1012 engagement metrics. *Sci Rep.* 2023;13(1):5705. [https://doi.org/10.1038/s41598-023-](https://doi.org/10.1038/s41598-023-32484-w)
1013 [32484-w](https://doi.org/10.1038/s41598-023-32484-w).
- 1014 13. Arora D. Predicting students academic success and dropout using supervised machine
1015 learning. *Int J Sci Stud.* 2023;11(6):72–8.
- 1016 14. Romero S, Liao X. Statistical and machine learning models for predicting university
1017 dropout and scholarship impact. *PLoS One.* 2025;20(6):e0325047.
1018 <https://doi.org/10.1371/journal.pone.0325047>.
- 1019 15. Martins MV, Tolledo D, Machado J, Baptista LMT, Realinho V. Early prediction of
1020 student's performance in higher education: a case study. In: Rocha Á, Adeli H, Dzemyda G,
1021 Moreira F, Ramalho Correia AM, editors. *Trends and applications in information systems*
1022 *and technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing*, vol.
1023 1365. Cham: Springer; 2021. p. 166–75. https://doi.org/10.1007/978-3-030-72657-7_16.
- 1024 16. Lee E, Song Y, Oh S. Survival analysis approach for student departure of freshmen:
1025 focusing on the case of S university. *J Learn Centered Curric Instr.* 2020;20(18):235–58.
- 1026 17. Tinto V. Through the eyes of students. *J Coll Stud Retent.* 2017;19(3):254–69.
1027 <https://doi.org/10.1177/1521025115621917>.
- 1028 18. Alalawi K, Athauda R, Chiong R. Contextualizing the current state of research on the use
1029 of machine learning for student performance prediction: a systematic literature review. *Eng*
1030 *Rep.* 2023;5(12):e12699. <https://doi.org/10.1002/eng2.12699>.
- 1031 19. Namoun A, Alshantiti A. Predicting student performance using data mining and learning
1032 analytics techniques: a systematic literature review. *Appl Sci.* 2021;11(1):237.
1033 <https://doi.org/10.3390/app11010237>.
- 1034 20. Hassan MA, Muse AH, Nadarajah S. Predicting student dropout rates using supervised
1035 machine learning: insights from the 2022 national education accessibility survey in
1036 Somaliland. *Appl Sci.* 2024;14(17):7593. <https://doi.org/10.3390/app14177593>.
- 1037 21. Kim S, Choi E, Jun YK, Lee S. Student dropout prediction for university with high
1038 precision and recall. *Appl Sci.* 2023;13(10):6275. <https://doi.org/10.3390/app13106275>.
- 1039 22. Villar A, de Andrade CRV. Supervised machine learning algorithms for predicting student
1040 dropout and academic success: a comparative study. *Discov Artif Intell.* 2024;4(1):2.
1041 <https://doi.org/10.1007/s44163-023-00079-z>.

- 1042 23. Alameri F. Predicting student dropout risk using machine learning [master's thesis].
1043 Rochester: Rochester Institute of Technology; 2025.
- 1044 24. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting
1045 student's dropout in university classes using two-layer ensemble machine learning
1046 approach: a novel stacked generalization. *Comput Educ Artif Intell*. 2022;3:100066.
1047 <https://doi.org/10.1016/j.caeai.2022.100066>.
- 1048 25. Ridwan A, Priyatno AM, Ningsih L. Predict students' dropout and academic success with
1049 XGBoost. *J Educ Comput Appl*. 2024;1(2):1–8.
- 1050 26. Agyemang EF, Mensah JA, Ampomah OA, Agyekum L, Akuoko-Frimpong J, Quansah A,
1051 et al. Predicting students' academic performance via machine learning algorithms: an
1052 empirical review and practical application. *Comput Eng Intell Syst*. 2024;15(1):86–102.
1053 <https://doi.org/10.7176/CEIS/15-1-09>.
- 1054 27. Liu H, Mao M, Li X, Gao J. Model interpretability on private-safe oriented student dropout
1055 prediction. *PLoS One*. 2025;20(3):e0317726.
1056 <https://doi.org/10.1371/journal.pone.0317726>.
- 1057 28. Morsu HR. Machine learning models for predicting student dropout, enrollment, and
1058 graduation. Preprints. 2025. <https://doi.org/10.20944/preprints202512.0166.v1>.
- 1059 29. Fonti V, Belitser E. Feature selection using LASSO. Research paper in business analytics.
1060 Amsterdam: Vrije Universiteit Amsterdam; 2017.
- 1061 30. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*.
1062 2003;3:1157–82.
- 1063 31. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. New York:
1064 Springer; 2005.
- 1065 32. Angelopoulos AN, Bates S, Malik J, Jordan MI. Uncertainty sets for image classifiers
1066 using conformal prediction. In: International Conference on Learning Representations
1067 (ICLR); 2021. <https://doi.org/10.48550/arXiv.2009.14193>.
- 1068 33. Angelopoulos AN, Bates S. Conformal prediction: a gentle introduction. *Found Trends*
1069 *Mach Learn*. 2023;16(4):494–591. <https://doi.org/10.1561/22000000101>.
- 1070 34. Papadopoulos H. Inductive conformal prediction: theory and application to neural
1071 networks. In: Fritzsche P, editor. Tools in artificial intelligence. Rijeka: InTech; 2008. p.
1072 315–30. <https://doi.org/10.5772/6078>.
- 1073 35. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. 2008;9:371–421.
- 1074 36. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-
1075 free uncertainty quantification. *arXiv:2107.07511 [Preprint]*. 2021.
1076 <https://doi.org/10.48550/arXiv.2107.07511>.
- 1077 37. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive
1078 inference for regression. *J Am Stat Assoc*. 2018;113(523):1094–111.
1079 <https://doi.org/10.1080/01621459.2017.1307116>.

38. Lei J, Wasserman L. Distribution-free prediction bands for non-parametric regression. *J R Stat Soc Series B Stat Methodol.* 2014;76(1):71–96. <https://doi.org/10.1111/rssb.12021>.
39. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. *Ann Stat.* 2023;51(2):816–45. <https://doi.org/10.1214/23-AOS2276>.
40. Tibshirani RJ, Foygel Barber R, Candès E, Ramdas A. Conformal prediction under covariate shift. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Red Hook: Curran Associates; 2019.
41. Solari A, Djordjilović V. Multi split conformal prediction. *Stat Probab Lett.* 2022;184:109395. <https://doi.org/10.1016/j.spl.2022.109395>.
42. Zhou X, Chen B, Gui Y, Cheng L. Conformal prediction: a data perspective. *ACM Comput Surv.* 2025;58(2):49. <https://doi.org/10.1145/3736575>.
43. Tinto V. Dropout from higher education: a theoretical synthesis of recent research. *Rev Educ Res.* 1975;45(1):89–125. <https://doi.org/10.3102/00346543045001089>.
44. Tinto V. Leaving college: rethinking the causes and cures of student attrition. 2nd ed. Chicago: University of Chicago Press; 1993.
45. Bean JP, Metzner BS. A conceptual model of nontraditional undergraduate student attrition. *Rev Educ Res.* 1985;55(4):485–540. <https://doi.org/10.3102/00346543055004485>.
46. Cabrera AF, Nora A, Castañeda MB. College persistence: structural equations modeling test of an integrated model of student retention. *J High Educ.* 1993;64(2):123–39. <https://doi.org/10.1080/00221546.1993.11778419>.
47. Kocsis Á, Molnár G. Factors influencing academic performance and dropout rates in higher education. *Oxf Rev Educ.* 2025;51(3):414–32. <https://doi.org/10.1080/03054985.2024.2316616>.
48. Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. *WIREs Data Min Knowl Discov.* 2020;10(3):e1355. <https://doi.org/10.1002/widm.1355>.
49. Albreiki B, Zaki N, Alashwal H. A systematic literature review of student' performance prediction using machine learning techniques. *Educ Sci.* 2021;11(9):552. <https://doi.org/10.3390/educsci11090552>.
50. Batool S, Rashid J, Nisar MW, Kim J, Kwon HY, Hussain A. Educational data mining to predict students' academic performance: a survey study. *Educ Inf Technol.* 2023;28(1):905–71. <https://doi.org/10.1007/s10639-022-11152-y>.
51. Prenkaj B, Velardi P, Stilo G, Distanto D, Faralli S. A survey of machine learning approaches for student dropout prediction in online courses. *ACM Comput Surv.* 2020;53(3):57. <https://doi.org/10.1145/3388792>.
52. Delen D. A comparative analysis of machine learning techniques for student retention management. *Decis Support Syst.* 2010;49(4):498–506. <https://doi.org/10.1016/j.dss.2010.06.003>.

- 1117 53. Kemper L, Vorhoff G, Wigger BU. Predicting student dropout: a machine learning
1118 approach. *Eur J High Educ.* 2020;10(1):28–47.
1119 <https://doi.org/10.1080/21568235.2020.1718520>.
- 1120 54. Berens J, Schneider K, Görtz S, Oster S, Burghoff J. Early detection of students at risk—
1121 predicting student dropouts using administrative student data from German universities and
1122 machine learning methods. *J Educ Data Min.* 2019;11(3):1–41.
1123 <https://doi.org/10.5281/zenodo.3594771>.
- 1124 55. Krüger JGC, Britto AS Jr, Barddal JP. An explainable machine learning approach for
1125 student dropout prediction. *Expert Syst Appl.* 2023;233:120933.
1126 <https://doi.org/10.1016/j.eswa.2023.120933>.
- 1127 56. Bettahi A, Belouadha FZ, Harroud H. A modular and explainable machine learning
1128 pipeline for student dropout prediction in higher education. *Algorithms.* 2025;18(10):662.
1129 <https://doi.org/10.3390/a18100662>.
- 1130 57. Rabelo AM, Zárate LE. A model for predicting dropout of higher education students. *Data*
1131 *Sci Manag.* 2025;8(1):72–85. <https://doi.org/10.1016/j.dsm.2024.07.001>.
- 1132 58. Bañeres D, Rodríguez ME, Guerrero-Roldán AE, Karadeniz A. An early warning system to
1133 detect at-risk students in online higher education. *Appl Sci.* 2020;10(13):4427.
1134 <https://doi.org/10.3390/app10134427>.
- 1135 59. Lopez-Muñoz G, Gonzalez Serrano C, Sanchez-Ferreira C. Student dropout prediction in
1136 higher education: a systematic review of machine learning methods and risk factors. *J Inf*
1137 *Technol Educ Res.* 2026;25:20. <https://doi.org/10.28945/5776>.
- 1138 60. Bertolini R, Finch SJ, Nehm RH. Quantifying variability in predictions of student
1139 performance: examining the impact of bootstrap resampling in data pipelines. *Comput*
1140 *Educ Artif Intell.* 2022;3:100067. <https://doi.org/10.1016/j.caeai.2022.100067>.
- 1141 61. Martins MV, Baptista L, Machado J, Realinho V. Multi-class phased prediction of
1142 academic performance and dropout in higher education. *Appl Sci.* 2023;13(8):4702.
1143 <https://doi.org/10.3390/app13084702>.
- 1144 62. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances*
1145 *in Neural Information Processing Systems 30 (NIPS 2017)*. Red Hook: Curran Associates;
1146 2017. p. 4765–74.
- 1147 63. Nagy M, Molontay R. Interpretable dropout prediction: towards XAI-based personalized
1148 intervention. *Int J Artif Intell Educ.* 2024;34(2):274–300. [https://doi.org/10.1007/s40593-](https://doi.org/10.1007/s40593-023-00331-8)
1149 [023-00331-8](https://doi.org/10.1007/s40593-023-00331-8).
- 1150 64. Rudin C. Stop explaining black box machine learning models for high stakes decisions and
1151 use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206–15.
1152 <https://doi.org/10.1038/s42256-019-0048-x>.
- 1153 65. Arnold KE, Pistilli MD. Course signals at Purdue: using learning analytics to increase
1154 student success. In: *Proceedings of the 2nd International Conference on Learning Analytics*

1155 and Knowledge. New York: ACM; 2012. p. 267–70.
 1156 <https://doi.org/10.1145/2330601.2330666>.

1157 66. Jayaprakash SM, Moody EW, Lauría EJM, Regan JR, Baron JD. Early alert of
 1158 academically at-risk students: an open source analytics initiative. *J Learn Anal*.
 1159 2014;1(1):6–47. <https://doi.org/10.18608/jla.2014.11.3>.

1160 67. Villano R, Harrison S, Lynch G, Chen G. Linking early alert systems and student retention:
 1161 a survival analysis approach. *High Educ*. 2018;76(5):903–20.
 1162 <https://doi.org/10.1007/s10734-018-0249-y>.

1163 68. Plak S, Cornelisz I, Meeter M, van Klaveren C. Early warning systems for more effective
 1164 student counselling in higher education: evidence from a Dutch field experiment. *High*
 1165 *Educ Q*. 2022;76(1):131–52. <https://doi.org/10.1111/hequ.12298>.

1166 69. Slade S, Prinsloo P. Learning analytics: ethical issues and dilemmas. *Am Behav Sci*.
 1167 2013;57(10):1510–29. <https://doi.org/10.1177/0002764213479366>.

1168 70. Lawson C, Beer C, Rossi D, Moore T, Fleming J. Identification of 'at risk' students using
 1169 learning analytics: the ethical dilemmas of intervention strategies in a higher education
 1170 institution. *Educ Technol Res Dev*. 2016;64(5):957–68. [https://doi.org/10.1007/s11423-](https://doi.org/10.1007/s11423-016-9459-0)
 1171 [016-9459-0](https://doi.org/10.1007/s11423-016-9459-0).

1172 71. Tzimas D, Demetriadis S. Ethical issues in learning analytics: a review of the field. *Educ*
 1173 *Technol Res Dev*. 2021;69(2):1101–33. <https://doi.org/10.1007/s11423-021-09977-4>.

1174 72. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of
 1175 generalizability, transparency and ethics in predictive learning analytics. *Comput Educ*
 1176 *Open*. 2021;2:100060. <https://doi.org/10.1016/j.caeo.2021.100060>.

1177 73. Baker RS, Hawn A. Algorithmic bias in education. *Int J Artif Intell Educ*.
 1178 2022;32(4):1052–92. <https://doi.org/10.1007/s40593-021-00285-9>.

1179 74. Idowu JA. Debiasing education algorithms. *Int J Artif Intell Educ*. 2024;34(4):1510–40.
 1180 <https://doi.org/10.1007/s40593-023-00389-4>.

1181 75. Bird KA, Castleman BL, Song Y. Are algorithms biased in education? Exploring racial bias
 1182 in predicting community college student success. *J Policy Anal Manage*. 2025;44(2):379–
 1183 402. <https://doi.org/10.1002/pam.22569>.

1184 76. Yu R, Lee H, Kizilcec RF. Should college dropout prediction models include protected
 1185 attributes? In: *Proceedings of the Eighth ACM Conference on Learning @ Scale*. New
 1186 York: ACM; 2021. p. 91–100. <https://doi.org/10.1145/3430895.3460139>.

1187 77. Deho OB, Joksimović S, Li J, Zhan C, Liu J, Liu L. Should learning analytics models
 1188 include sensitive attributes? Explaining the why. *IEEE Trans Learn Technol*.
 1189 2023;16(4):560–72. <https://doi.org/10.1109/TLT.2022.3226474>.

1190 78. Hu Q, Rangwala H. Reliable deep grade prediction with uncertainty estimation. In:
 1191 *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. New
 1192 York: ACM; 2019. p. 76–85. <https://doi.org/10.1145/3303772.3303802>.

- 1193 79. Kwon JB. A portable, generalizable machine learning framework for long-term student
1194 dropout prediction. *IEEE Access*. 2026;14:30830–43.
1195 <https://doi.org/10.1109/ACCESS.2026.3666870>.
- 1196 80. Sadinle M, Lei J, Wasserman L. Least ambiguous set-valued classifiers with bounded error
1197 levels. *J Am Stat Assoc*. 2019;114(525):223–34.
1198 <https://doi.org/10.1080/01621459.2017.1395341>.
- 1199 81. Romano Y, Sesia M, Candès E. Classification with valid and adaptive coverage. In:
1200 *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Red Hook:
1201 Curran Associates; 2020. p. 3581–91.
- 1202 82. Vovk V. Conditional validity of inductive conformal predictors. In: *Proceedings of the*
1203 *Asian Conference on Machine Learning*. PMLR. 2012;25:475–90.
- 1204 83. Gibbs I, Cherian JJ, Candès EJ. Conformal prediction with conditional guarantees. *J R Stat*
1205 *Soc Series B Stat Methodol*. 2025;87(4):1100–26. <https://doi.org/10.1093/jrsssb/qkaf008>.
- 1206 84. Romano Y, Barber RF, Sabatti C, Candès EJ. With malice toward none: assessing
1207 uncertainty via equalized coverage. *Harv Data Sci Rev*. 2020;2(2).
1208 <https://doi.org/10.1162/99608f92.03f00592>.
- 1209 85. Gibbs I, Candès E. Adaptive conformal inference under distribution shift. In: *Advances in*
1210 *Neural Information Processing Systems 34 (NeurIPS 2021)*. Red Hook: Curran Associates;
1211 2021. p. 1660–72.
- 1212 86. Cresswell JC, Sui Y, Kumar B, Vouitsis N. Conformal prediction sets improve human
1213 decision making. In: *Proceedings of the 41st International Conference on Machine*
1214 *Learning*. PMLR. 2024;235:9439–57.
- 1215 87. Babbar V, Bhatt U, Weller A. On the utility of prediction sets in human-AI teams. In:
1216 *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*
1217 *(IJCAI-22)*. 2022. p. 2457–63. <https://doi.org/10.24963/ijcai.2022/341>.
- 1218 88. Straitouri E, Wang L, Okati N, Gomez Rodriguez M. Improving expert predictions with
1219 conformal prediction. In: *Proceedings of the 40th International Conference on Machine*
1220 *Learning*. PMLR. 2023;202:32633–53.
- 1221 89. Zhang D, Chatzimpampas A, Kamali N, Hullman J. Evaluating the utility of conformal
1222 prediction sets for AI-advised image labeling. In: *Proceedings of the CHI Conference on*
1223 *Human Factors in Computing Systems*. New York: ACM; 2024. p. 1–19.
1224 <https://doi.org/10.1145/3613904.3642446>.
- 1225 90. Cresswell JC, Kumar B, Sui Y, Belbahri M. Conformal prediction sets can cause disparate
1226 impact. In: *International Conference on Learning Representations (ICLR)*; 2025.
1227 <https://doi.org/10.48550/arXiv.2410.01888>.
- 1228 91. Morsomme R, Smirnov E. Conformal prediction for students' grades in a course
1229 recommender system. In: *Proceedings of the Eighth Symposium on Conformal and*
1230 *Probabilistic Prediction and Applications*. PMLR. 2019;105:196–213.

- 1231 92. Elmasry MMS, Gafar MG, Elsabagh MA. Trustworthy educational risk modeling with
1232 calibrated probabilities, conformal uncertainty, explainable AI, and graph-based
1233 refinement. *Inventions*. 2026;11(3):65. <https://doi.org/10.3390/inventions11030065>.
- 1234 93. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via
1235 coordinate descent. *J Stat Softw*. 2010;33(1):1–22. <https://doi.org/10.18637/jss.v033.i01>.
- 1236 94. Agresti A. Foundations of linear and generalized linear models. Hoboken: Wiley; 2015.
- 1237 95. Fox J. Applied regression analysis and generalized linear models. 3rd ed. Thousand Oaks:
1238 SAGE Publications; 2015.
- 1239 96. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
1240 <https://doi.org/10.1023/A:1010933404324>.
- 1241 97. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd
1242 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New
1243 York: ACM; 2016. p. 785–94. <https://doi.org/10.1145/2939672.2939785>.
- 1244 98. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient
1245 gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30
1246 (NIPS 2017). Red Hook: Curran Associates; 2017. p. 3146–54.
- 1247 99. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*.
1248 2010;36(11):1–13. <https://doi.org/10.18637/jss.v036.i11>.
- 1249 100. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat*
1250 *Methodol*. 1996;58(1):267–88. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- 1251 101. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network
1252 classification models: a methodology review. *J Biomed Inform*. 2002;35(5–6):352–9.
1253 [https://doi.org/10.1016/s1532-0464\(03\)00034-0](https://doi.org/10.1016/s1532-0464(03)00034-0).
- 1254 102. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min*
1255 *Knowl Discov*. 1998;2(2):121–67. <https://doi.org/10.1023/A:1009715923555>.
- 1256 103. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey
1257 on support vector machine classification: applications, challenges and trends.
1258 *Neurocomputing*. 2020;408:189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>.
- 1259 104. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
- 1260 105. Scornet E, Hooker G. Theory of random forests. *Annu Rev Stat Appl*. 2026;13:99–121.
1261 <https://doi.org/10.1146/annurev-statistics-112723-034707>.
- 1262 106. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its
1263 posterior distribution. In: 2010 20th International Conference on Pattern Recognition.
1264 Istanbul: IEEE; 2010. p. 3121–4. <https://doi.org/10.1109/ICPR.2010.764>.
- 1265 107. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification
1266 tasks. *Inf Process Manag*. 2009;45(4):427–37. <https://doi.org/10.1016/j.ipm.2009.03.002>.

- 1267 108. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple
1268 class classification problems. *Mach Learn.* 2001;45(2):171–86.
1269 <https://doi.org/10.1023/A:1010920819831>.
- 1270 109. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Predictive inference with the jackknife+.
1271 *Ann Stat.* 2021;49(1):486–507. <https://doi.org/10.1214/20-AOS1965>.
1272

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFiles.docx](#)