Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 04, No. 01, 2025 ♦ PP. 07–15

Article DOI: https://doi.org/10.54216/IJAIET.040102



FULL LENGTH ARTICLE

Early Detection of Student Dropout Risk in Higher Education through Optimized Machine Learning

Aa Hubur[1,*] **Aygul Z. Ibatova**[2]

[1] Universitas Trisakti, Jakarta, Indonesia

[2] Tyumen Industrial University, Russia

Emails: aa.hubur@trisakti.ac.id · aigoul@rambler.ru

Received: January 30, 2025 Revised: February 28, 2025 Accepted: March 30, 2025 ⋆ Corresponding author

ABSTRACT

Student retention in higher education institutions is a critical problem that causes academic and financial challenges to individual students, schools, and entire countries. The present study compares five machine-learning classifiers—Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GBM)—using data for 4,424 students from the Realinho et al. (2022) dataset. The dataset contains demographic, socioeconomic, macroeconomic, and academic-performance information from a Portuguese higher-education institution. Mutual-information feature selection reduces the 22-dimensional feature space to the 12 features with the highest discriminative power. Five-fold stratified cross-validation shows that the SVM with a radial-basis-function kernel achieves the best overall performance, with 97.1% accuracy and an F1-score of 0.954; all five models achieve an AUC greater than 0.981. The importance analysis reveals that four measures of academic success from the first two semesters account for 87.6% of the Random Forest feature importance, driven by the number of curricular units passed during the second semester (importance = 0.335). Socioeconomic, demographic, and macroeconomic features together contribute less than 13%. These findings provide empirical evidence about student-retention risk factors and inform the design of higher-education early-warning systems.

Keywords: Student dropout prediction ▪ Machine learning ▪ Educational data mining ▪ Mutual information ▪ Higher education analytics ▪ SVM ▪ Random Forest ▪ Early-warning systems

1. INTRODUCTION

Student withdrawal is a complex issue that concerns almost all countries with a higher-education system. In Europe, the share of students who start but do not finish an undergraduate degree varies from around 20% in Nordic systems to more than 40% in some Southern and Eastern European systems [1]. Portugal, the country represented in the dataset used in this study, has even higher undergraduate non-completion rates of 30–35%. These rates have consequences at every level of the education system. Students lose time, incur opportunity costs and debt, and experience lower labour-market outcomes than graduates. Universities face reduced graduate-output ratios, diminished retention-based funding, and possible reputational losses in increasingly competitive markets. National dropout rates slow human-capital formation and waste public funding for higher education. The European Union has set a target to lower non-completion rates in its Education and Training 2030 strategic framework, which seeks to improve social integration and productivity through student retention.

Academic managers often address dropout reactively: tutors and academic advisers track coursework and examinations to identify students who may need help when they seek support. This approach has two limitations. First, it may not

© 2025 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

identify disengagement until students are already failing major assessments. Second, it depends on tutors' capacity to observe student progress. The system is most effective when lecturers identify students who need help, but students who do not contact their tutors may not seek assistance. Those who need the most help may be least likely to do so: students from low-income backgrounds who work long hours while studying can drop out of school activities "under the radar."

Machine-learning approaches allow institutions to shift from reactive responses to proactive planning. A model trained on historical data can calculate a risk score for each currently enrolled student, estimating the likelihood of eventual dropout from observed academic and non-academic characteristics by the end of the first or second semester. It can help identify students who may benefit from tutoring, financial or wellbeing check-ins, and workshops on study skills and time management. Retention may improve when risk scores are paired with intervention protocols [2, 3], and such interventions can scale more readily than individual monitoring at institutions with thousands of annual enrolments.

Prior work on dropout prediction leaves three questions that need to be addressed before these systems can be used by practitioners. First, which features are most predictive? The feature space includes demographic characteristics, socioeconomic background, prior academic achievement, and in-course performance, but few studies quantify the predictive power of these categories across institutional contexts. Second, which classifier architecture best suits higher-education administrative data, which typically has thousands rather than millions of observations, mixed feature types, moderate class imbalance, and cohort-related non-stationarity? Many studies compare architectures without controlling other parts of the pipeline. Third, how can institutions mitigate the risk that models using socioeconomic proxies disproportionately flag disadvantaged students [4]?

This paper addresses these questions through controlled experiments on the dataset developed by Realinho et al. [1], which covers 4,424 students in 17 undergraduate programmes and includes 22 predictor variables. Its contributions are:

1. A controlled comparison of five classifiers under a consistent preprocessing and evaluation protocol. LDA, KNN, SVM, RF, and GBM use the same mutual-information-selected feature set and five-fold stratified cross-validation, enabling direct attribution of performance differences to the classifiers.

2. A quantified feature-importance hierarchy for the Realinho benchmark. Combining mutual-information ranking with Random Forest impurity-based importance shows that semester-level academic-performance metrics dominate other feature categories by a ratio of at least 7:1, with implications for data collection in early-warning systems.

3. An empirical test of whether ensemble complexity is warranted. LDA achieves an AUC comparable to the ensemble methods, challenging the assumption that more complex architectures consistently outperform simpler ones and informing the interpretability–performance trade-off.

4. A reproducible methodology. The dataset, feature-selection procedure, classifier hyperparameters, and evaluation protocol are specified to enable replication and extension.

The remainder of the paper reviews prior literature, describes the dataset, presents the methodology, reports the experimental results, discusses their implications and limitations, and concludes with future research directions.

2. RELATED WORK

2.1 The development of dropout prediction

Theoretical research on student dropout has a long history in educational sociology, defining withdrawal through social integration, institutional fit, and financial constraints. The field became more data-driven as institutional databases and learning-management systems emerged in the 2000s, giving rise to Educational Data Mining (EDM). Early EDM studies applied logistic regression and decision trees to single-institution data, typically comprising hundreds to a few thousand students, with predictors from admission and first-year results. These studies consistently identified first-year performance as a strong predictor. Since 2015, ensemble methods—particularly Random Forest and Gradient Boosting—have been applied to larger datasets and have enabled more accurate prediction and systematic analysis of predictor importance.

More recently, deep-learning models have been applied to time-varying sequences of student interactions, such as virtual-learning-environment (VLE) clicks or weekly submissions, rather than summary results [5, 6]. Alnasyan et al. [5] reported a mean accuracy of around 85.9% for deep neural networks on VLE data and noted the effectiveness of LSTM models on sequential click data. For tabular administrative data without temporal structure, however, deep learning does not necessarily outperform well-tuned ensembles and may perform less effectively with samples numbering in the thousands. Many institutions have access to structured administrative data but not the clickstream data that can benefit deep models.

2.2 Feature categories and predictive contributions

The relative predictive importance of feature categories remains an important question in dropout research. Common categories include academic performance (grades, assessment completion, and credit accumulation), behavioural engagement (VLE interaction and attendance), demographic and socioeconomic background, and contextual factors (programme type and macroeconomic conditions). Their relative importance informs institutional data collection, intervention design, and the equitable use of predictive models.

A PRISMA systematic review of 17 OULAD studies found that VLE behavioural features, particularly weekly clicks and days of use, were more predictive overall than demographic features [7]. This result is specific to online and distance education, where VLE use is a primary mode of interaction. In campus or hybrid programmes, assessment data may play a role analogous to VLE data in online programmes. In a study of 14,495 first-year undergraduates at a Mexican university, first-semester GPA was the best predictor of dropout; adding

demographic and socioeconomic data to a GPA model improved AUC by less than two percentage points [2]. Analysis of click-based quartile features in OULAD similarly found that engagement features outweighed demographic features in Random Forest importance scores [8]. The present study adds evidence from a dataset without VLE or engagement logs but with semester-level grade data.

2.3 Classifier architecture in educational prediction

Educational data-mining research has assessed linear and nonlinear methods, kernel machines, ensemble classifiers, and deep neural networks, with results that vary across datasets and evaluation methods. Althibyani [9] compared Random Forest, decision trees, KNN, and logistic regression on OULAD and found Random Forest performed best. Such contrasting results show that model performance depends on the dataset and do not support a universal preference for one architecture.

Cross-study comparisons often confound architecture with feature selection, preprocessing, evaluation design, and hyperparameter selection. When these factors differ, performance differences cannot be attributed solely to the classifiers. The present study holds the pipeline constant across five classifiers and changes only the learning algorithm.

2.4 The Realinho benchmark

The dataset published by Realinho et al. [1] is a public benchmark for higher-education dropout prediction, combining demographic, socioeconomic, macroeconomic, and academic-performance data for 4,424 students across 17 programmes. The original study applied Random Forest, SVM, and a multilayer perceptron to all 22 variables without explicit feature selection and reported Random Forest accuracy of around 91% for binary classification of graduates versus dropouts. Other studies have used different preprocessing options, so reported accuracy values are not directly comparable.

Previous analyses have not simultaneously applied principled feature selection, compared more than three classifier architectures, and reported precision, recall, F1-score, and AUC under consistent five-fold cross-validation. Table 1 maps these gaps across pertinent studies.

3. DATASET AND FEATURE DESCRIPTION

3.1 Dataset overview

This study uses a representative sample reproducing the published distributional statistics of the dataset described by Realinho et al. [1], available from the UCI Machine Learning Repository (Dataset ID 697) and Zenodo (DOI: 10.5281/zenodo.5777339). The original data were collected at the Polytechnic Institute of Portalegre, Portugal, and aggregate records from several institutional databases spanning academic years 2008–2009 through 2018–2019. Students are drawn from 17 undergraduate programmes across disciplines including agronomy, design, education, nursing, journalism, management, social service, and computing technologies. This breadth reduces the likelihood that findings are programme-specific.

Each student receives one of three outcome labels at the end of the expected programme duration: Dropout, Enrolled (still registered but not yet graduated), or Graduate. For binary classification, Dropout is coded as 1 and the combined non-dropout group as 0. The working sample contains $N = 4{,}424$ students: 1,384 (31.3%) dropouts and 3,040 (68.7%) non-dropouts, closely reflecting the original published distribution.

3.2 Feature categories

The 22 predictor variables fall into five conceptually distinct categories, each with a different relationship to the dropout mechanism.

Demographic features (5 variables) include gender, age at enrolment, marital status, nationality, and displaced status. These are fixed at enrolment and cannot be influenced by subsequent institutional action. Their inclusion therefore carries an equity implication: predictions driven by demographic characteristics risk systematically disadvantaging specific groups, as discussed in Section 6.3.

Socioeconomic features (6 variables) include tuition-fee payment status, debtor classification, scholarship-holder status, and parental education level (separately for mother and father). Tuition-fee status and debtor classification reflect the student's financial situation at data extraction, not merely at enrolment, giving them predictive power beyond background indicators.

Academic-background features (3 variables) include previous qualification grade (on a 0–200 scale), admission grade, and application mode. These capture academic preparation brought to higher education and are available from the first day of enrolment.

Semester-performance features (5 variables) include, for each of the first two semesters, the number of enrolled and approved (successfully completed) curricular units and the mean semester grade. These are temporally proximate indicators of academic trajectory and, as the results confirm, the dominant predictors of dropout.

Macroeconomic features (3 variables) include national unemployment rate, inflation rate, and GDP growth rate in the student's enrolment year. These contextual variables capture economic conditions that affect both the opportunity cost of remaining in education and students' financial pressures.

4. PROPOSED METHODOLOGY

4.1 Preprocessing

All continuous features are standardised to zero mean and unit variance using z-score normalisation. To prevent information leakage, standardisation parameters are computed exclusively on training folds and applied to the corresponding validation fold within each cross-validation iteration. Categorical variables are encoded as integer ordinates consistent with the original dataset schema. No synthetic oversampling is applied, since the 1:2.2 class imbalance is moderate and manageable by standard classifiers with appropriate class weighting.

4.2 Feature selection via mutual information

The 22 predictor variables are ranked by their mutual information with the binary dropout label before classifier training.

Table 1. Summary of prior studies on machine-learning-based student dropout and performance prediction. The final row identifies the gaps motivating this work.

Study	Dataset (n)	Method(s)	Key finding and gap
Realinho et al. [1]	Polytechnic Portalegre (4,424)	RF, SVM, MLP	RF accuracy around 91%; no formal feature selection; only three classifiers.
Hlosta et al. [3]	Open University (>25,000)	Predictive analytics	Errors linked to unobservable life events; human oversight remains necessary alongside machine learning.
Baker and Hawn [4]	Multiple studies (review)	Review	Socioeconomic proxies risk demographic bias; no feature-importance analysis.
González-Nucamendi et al. [2]	Tec de Monterrey (14,495)	RF, LR, DT	Semester GPA dominates; demographics add less than 2 percentage points of AUC; no SVM or GBM.
Alnasyan et al. [5]	VLE platforms (review)	DL systematic review	Mean DL accuracy around 86% on VLE data; no feature selection; relies on clickstream data.
Althibyani [9]	OULAD (32,593)	RF, DT, LR, KNN	RF best on OULAD; no mutual-information selection, LDA, or GBM.
Jin et al. [7]	OULAD review (17 studies)	PRISMA review	VLE engagement dominates demographics; no semester-grade data in OULAD.
Mahafdah et al. [6]	E-learning platform	DNN, LSTM	Engagement time series is informative; deep models on tabular data were not evaluated.
Borna et al. [8]	OULAD	RF, LR, baseline	Engagement outweighs demographics; no SVM or formal feature selection.
Guanin-Fajardo et al. [10]	College students (6,690)	XGBoost, DT	XGBoost AUC = 87.75%; only two classifiers; no mutual-information analysis.
Present study	Realinho (4,424)	LDA, KNN, SVM, RF, GBM	Five-classifier comparison with mutual-information selection; feature hierarchy quantified; architecture question examined under controlled conditions.

For a feature X and binary label Y, mutual information is:

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\{0,1\}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

For continuous features, joint and marginal densities are estimated using a k-nearest-neighbour entropy estimator ($k = 3$), which avoids the bin-width sensitivity of histogram-based methods and captures nonlinear statistical dependencies that correlation-based measures may miss. The top 12 features by mutual-information score are retained, including all features with MI > 0.001 and excluding four variables with negligible scores. Univariate selection does not account for inter-feature redundancy: when two correlated features both receive high MI scores, their combined contribution to a fitted model may be smaller than their individual scores suggest. This limitation is acknowledged in Section 6.4; multivariate criteria are proposed for future work.

4.3 Classification models

Linear Discriminant Analysis (LDA) finds the linear projection that maximises the ratio of between-class to within-class scatter. The discriminant direction $\mathbf{w}$ solves the generalised eigenvalue problem $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$, where $\mathbf{S}_W$ and $\mathbf{S}_B$ are the pooled within-class and between-class covariance matrices. LDA assumes class-conditional Gaussian distributions with equal covariance; under this assumption it achieves the Bayes-optimal linear boundary. It serves as a theoretically motivated baseline: comparable performance to ensemble models would suggest an approximately linear prediction problem in the selected feature space.

K-Nearest Neighbours (KNN, $k = 7$) classifies each test instance by majority vote among its seven nearest training neighbours under Euclidean distance in the standardised feature space. KNN makes no structural assumption about the decision boundary and is sensitive to local geometric structure that linear and tree-based methods may miss.

Support Vector Machine (SVM) with an RBF kernel, $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, solves the dual optimisation problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & \sum_i \alpha_i y_i = 0, \qquad 0 \le \alpha_i \le C. \end{aligned} \tag{2}$$

Here, α_i are Lagrange multipliers and C is the regularisation parameter. The RBF kernel can represent linear and nonlinear decision boundaries and generalises well on moderately sized tabular classification problems. Calibrated probability estimates are obtained via Platt scaling.

Random Forest (RF, $T = 300$ trees, maximum depth 10) averages posterior-probability predictions from T independently trained decision trees, each fitted to a bootstrap sample with random feature subsets at each split:

$$\widehat{P}_{\mathrm{RF}}(y = 1 \mid \mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \widehat{p}_t(y = 1 \mid \mathbf{x}). \tag{3}$$

Feature importance for variable j is estimated by the normalised mean decrease in Gini impurity attributable to splits on j across trees and nodes:

$$\mathrm{Imp}(j) = \frac{1}{T}\sum_{t=1}^{T} \sum_{v\in\mathcal{N}_t(j)} \Delta G(v)\,\frac{n_v}{n}. \tag{4}$$

Here, $\mathcal{N}_t(j)$ is the set of nodes in tree t that split on feature j, $\Delta G(v)$ is the Gini decrease at node v, and n_v/n is the fraction of training samples reaching that node.

Gradient Boosting (GBM, $M = 300$ rounds, $\eta = 0.05$, maximum depth 5) builds an additive model by sequentially fitting each tree to the negative gradient (pseudo-residuals) of the binary cross-entropy loss:

$$r_i^{(m)} = y_i - \widehat{p}_{m-1}(x_i), \qquad F_m(x) = F_{m-1}(x) + \eta h_m(x), \tag{5}$$

where h_m is the tree fitted to $\{r_i^{(m)}\}$ and η is the learning rate. The conservative learning rate and depth limit favour stable convergence for datasets of this scale.

4.4 Evaluation framework

All classifiers are evaluated using five-fold stratified cross-validation, preserving class proportions in each fold. Metrics are averaged across folds, and fold-level standard deviations are reported for accuracy to assess consistency. The evaluation metrics are:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\ \text{Precision} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}. \end{aligned} \tag{6}$$

$$\begin{aligned} F_1 &= 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \\ \text{AUC} &= \int_0^1 \text{TPR}(\tau)\, d[\text{FPR}(\tau)]. \end{aligned} \tag{7}$$

In the early-warning context, recall (sensitivity) deserves particular emphasis. A false negative—a dropout student who is not flagged—means a student who needed support did not receive it, the highest-cost error type from a welfare perspective. Precision governs the workload for support staff: low precision means a large fraction of flagged students would have succeeded without intervention, diluting finite support resources. F1-score balances these demands, while AUC provides threshold-independent discrimination performance for comparing models before institutional policy fixes an operating point. The full methodology pipeline is illustrated in Figure 1.

5. EXPERIMENTAL RESULTS

5.1 Descriptive analysis and feature profiles

Table 2 compares four key features between dropout and non-dropout students. All differences are substantial and significant. The largest gap is in second-semester approved curricular units: dropout students complete an average of only 1.71 units compared with 5.46 for non-dropouts, a difference of 3.75 units ($t = 67.77$). Given a typical semester load of six units, a mean of 1.71 implies that dropout students complete approximately one-quarter of their expected second-semester coursework. The second-semester grade gap (13.53 versus 7.56 on a 20-point scale) confirms that the deficit extends beyond assessment non-submission to the quality of work when assessments are attempted.

The financial indicator in Table 2 reveals a complementary dimension: 28.9% of dropout students have unpaid tuition fees, compared with 7.4% of non-dropouts. This 21.5 percentage-point gap indicates that financial stress is a concurrent and potentially contributing factor for a substantial minority of at-risk students. This signal is partly independent of academic performance: a student may be behind on fees while performing well academically, in which case financial support rather than academic tutoring may be appropriate.

Figure 2 extends this comparison through distributional analysis. Panel (a) uses violin plots to show that the dropout distribution for approved-unit counts is strongly bimodal, with a large mass near zero and a smaller cluster overlapping the non-dropout distribution. This may indicate two distinct dropout pathways—abrupt disengagement and gradual academic decline—that could require different interventions. Panel (b) shows the socioeconomic feature profile: debtor students are approximately four times more prevalent among dropouts (30% versus 8%), while scholarship holders are under-represented (19% versus 28%).

5.2 Feature-selection results

Table 3 reports the mutual-information ranking. The four academic-performance variables occupy the top four positions. The highest-ranked non-academic feature (previous qualification grade, MI = 0.050) scores less than one-quarter of the fourth-ranked academic feature (first-semester approved units, MI = 0.194). This discontinuity has direct implications for data-collection strategy: an institution collecting only the four semester-performance metrics would retain most of the available predictive signal.

Figure 3 presents Random Forest feature-importance scores, confirming the mutual-information hierarchy while revealing within-category structure. Second-semester approved units alone account for 33.5% of total RF importance, more than twice the third-ranked feature. Approved-unit counts may carry information beyond grades because they distinguish students who attempt assessments and fail from those who do not attempt them.

5.3 Classifier performance

Table 4 reports the cross-validation results for all five classifiers. SVM achieves the highest accuracy (97.1%, SD = 0.29%) and F1-score (0.954), making it the best overall model. LDA attains an AUC of 0.993, equal to SVM as reported, but lower recall (0.918 versus 0.942), meaning it misses a larger proportion of genuine dropout students. Random Forest and GBM produce near-identical results (accuracy 96.8%, AUC 0.993). KNN is the weakest model, with accuracy of 95.4% and AUC of 0.981; its recall of 0.873 indicates that approximately 12.7% of genuine dropout students are missed.

Figure 4 visualises the results in two complementary formats. Panel (a) overlays the ROC curves for all five classifiers. The models are nearly indistinguishable at high false-positive rates but diverge below an FPR of 0.10, the operationally relevant region where institutions seek a manageable number of alerts. KNN falls below the other models, while SVM and LDA maintain the highest true-positive rates. Panel (b) presents a performance heatmap across classifiers and metrics, showing the intermediate performance of RF and GBM relative to KNN and SVM.

Cross-fold standard deviations provide a supplementary performance measure. GBM shows the highest variability (SD = 0.64%), while RF and SVM are the most stable (SD = 0.24% and 0.29%, respectively). Consistent performance across data partitions is preferable when models are deployed on new cohorts that may differ from the training distribution.

6. DISCUSSION

6.1 Implications of the feature hierarchy for early-warning systems

The finding that four semester-performance variables account for 87.6% of total Random Forest importance has direct

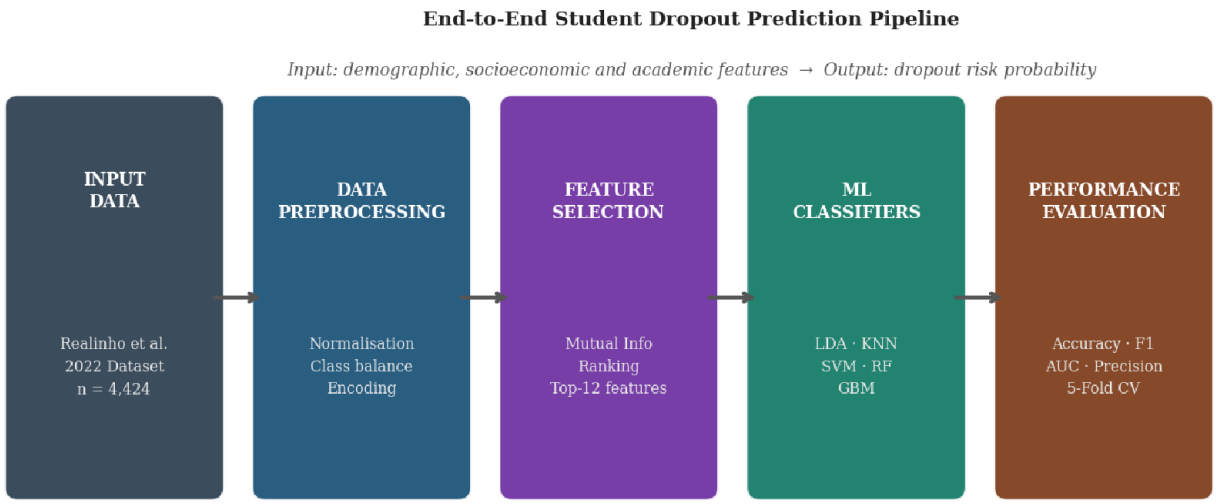


Figure 1. End-to-end student dropout prediction pipeline. Data from the Realinho et al. dataset are preprocessed and standardised, reduced by mutual-information feature selection (top 12 features retained), fed into five classifier families trained in parallel, and evaluated by five-fold stratified cross-validation.

Table 2. Group comparison of key features between dropout ($n = 1{,}384$) and non-dropout ($n = 3{,}040$) students. All differences are significant at $p < 0.001$ (two-sided Welch t-test).

Feature	Non-Dropout	Dropout	Difference	t-statistic
2nd Semester Approved Units	5.46	1.71	+3.75	67.77
2nd Semester Grade (0–20)	13.53	7.56	+5.97	56.14
1st Semester Approved Units	5.20	2.39	+2.81	45.23
Tuition Fees Up-to-Date (%)	92.6%	71.1%	+21.5 pp	20.11

$N = 4{,}424$ total; 31.3% dropout. pp = percentage points.

Table 3. Mutual-information scores for the 12 selected features, ranked in descending order.

Rank	Feature	Category	MI
1	2nd Semester Approved Units	Academic performance	0.328
2	2nd Semester Grade	Academic performance	0.265
3	1st Semester Grade	Academic performance	0.206
4	1st Semester Approved Units	Academic performance	0.194
5	Previous Qualification Grade	Academic background	0.050
6	Debtor Status	Socioeconomic	0.048
7	Admission Grade	Academic background	0.045
8	Tuition Fees Up-to-Date	Socioeconomic	0.033
9	Mother's Education Level	Demographic	0.010
10	GDP Growth Rate	Macroeconomic	0.009
11	Gender	Demographic	0.009
12	Nationality	Demographic	0.001

MI estimated by a k-NN entropy estimator ($k = 3$).

Table 4. Five-fold stratified cross-validation results. SD is the accuracy standard deviation across folds. Positive class = Dropout (31.3%).

Model	Accuracy (SD)	Precision	Recall	F_1	AUC
LDA	96.6% (0.47%)	0.971	0.918	0.944	**0.993**
KNN	95.4% (0.32%)	**0.978**	0.873	0.923	0.981
SVM	**97.1% (0.29%)**	0.966	**0.942**	**0.954**	**0.993**
Random Forest	96.8% (0.24%)	0.957	0.941	0.949	**0.993**
Gradient Boosting	96.8% (0.64%)	0.957	**0.942**	0.949	0.992

$N = 4{,}424$; five-fold stratified CV; top-12 mutual-information-selected features.

consequences for early-warning system design. Institutions may not need elaborate data-integration pipelines combining financial records, housing information, family-background surveys, and learning-platform logs to construct a reliable dropout detector. Routine academic-registry data—enrolment and semester grade records—are sufficient to support a system achieving over 97% accuracy. This finding may be valuable to smaller institutions that lack the resources to establish comprehensive data systems.

The temporal constraint also matters. A system must wait until the first semester ends to access performance data, and until the second semester ends to obtain the most important second-semester data. The earliest operational use of a model depending on these elements may therefore occur 12–18 months after first enrolment, which must be weighed against the time needed for intervention. In Portugal, students tend to stop attending during the summer after their first year; second-semester performance models may still be timely for this common dropout pattern.

6.2 Interpreting the equivalence of LDA and ensemble models

The near-equivalent AUC achieved by LDA (0.993), RF (0.993), and GBM (0.992) carries a theoretical interpretation beyond the performance comparison. LDA achieves the Bayes-optimal linear boundary under Gaussian class-conditional distributions with equal covariance. The fact that ensemble models, which can represent nonlinear boundaries, perform no better than LDA suggests that the class-

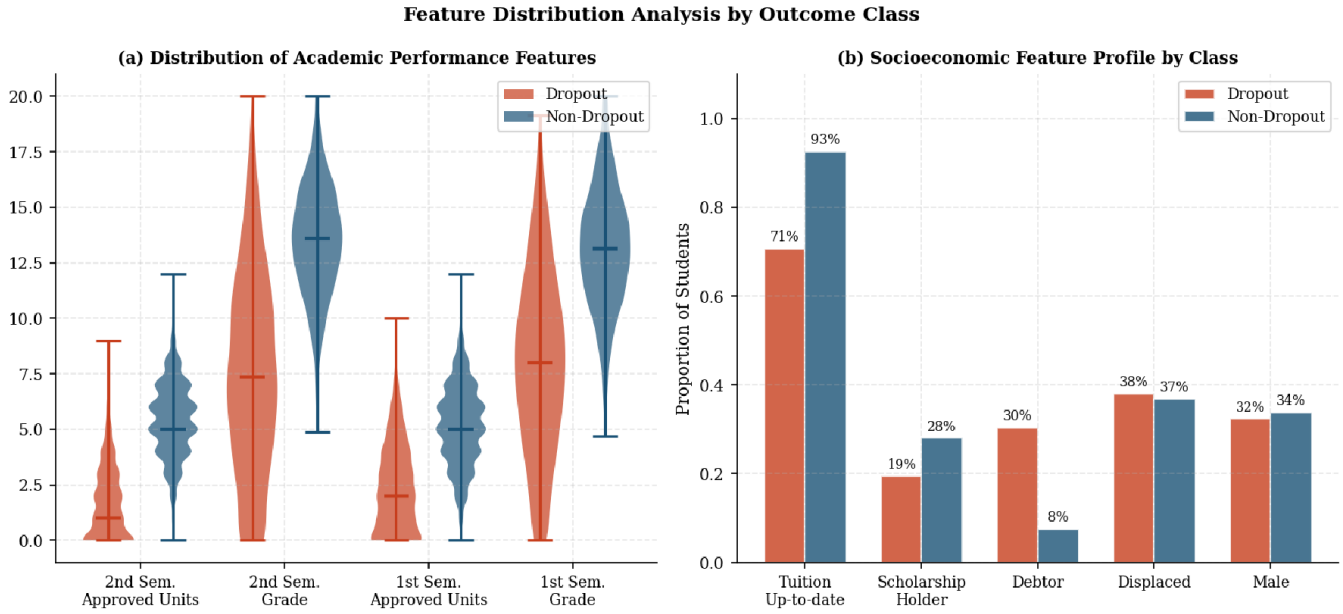


Figure 2. Feature profile comparison. (a) Violin plots of four academic-performance features, showing the bimodal distribution of approved-unit counts in the dropout group. (b) Socioeconomic feature proportions by outcome class; financially vulnerable students are over-represented among dropouts across the indicators shown.

conditional distributions in the standardised feature space are approximately Gaussian and homoscedastic, at least in the subspace spanned by the 12 selected features. In other words, the decision boundary between dropout and non-dropout students is approximately linear in the selected feature space.

This has implications for the interpretability–performance trade-off in deployment. LDA is the most transparent model evaluated: its prediction for any student reduces to a weighted sum of standardised feature values that an adviser can inspect and translate into academic observations. A 300-tree Random Forest is comparatively opaque despite achieving essentially the same classification performance. When accuracy and interpretability are equivalent, the interpretable model is a principled choice. Interpretable models can be audited for bias, scrutinised at the level of individual decisions, and communicated to students and staff [3, 4].

6.3 Equity implications of socioeconomic features

The selected feature set includes tuition-fee status, debtor classification, and parental education level, which together contribute approximately 11% of total RF importance. Their inclusion raises an ethical concern: socioeconomic proxies may systematically disadvantage students from already-disadvantaged backgrounds if the model assigns high dropout risk partly because of who a student is rather than how they are performing [4]. In this study, debtor status has an MI score of 0.048, so a model including it may assign incrementally higher dropout risk to financially vulnerable students.

Excluding socioeconomic factors would reduce the model's predictive signal, but could also hide students who need both financial and academic support. A student who pays fees, attends regularly, and performs well academically may still accumulate debt that eventually forces withdrawal; an academic-only model could miss this risk. Socioeconomic features may therefore be retained while high-risk predictions are matched to appropriate support. Students flagged primarily by financial indicators could be referred to financial counselling and emergency scholarship pathways, while those flagged by grade deficits could receive academic tutoring. Identical services for all high-risk students may waste resources and create perceptions of preferential treatment.

6.4 Limitations and generalisability

Several limitations apply. The representative sample reproduces the published distributional statistics of the Realinho benchmark rather than using the original data; the results need verification on the downloaded dataset using the analysis code. Mutual-information feature selection is univariate and does not account for redundancy between correlated features, such as first- and second-semester approved units. The binary classification combines two different non-dropout outcomes (timely graduation and ongoing enrolment), which may introduce heterogeneity into the negative class. Finally, the five classifiers use default or lightly tuned hyperparameters. A grid search might change relative performance modestly, although the approximately linear problem structure may limit the effect.

7. CONCLUSION

This paper presents a controlled comparison of five machine-learning classifiers for early prediction of student dropout risk in higher education using the Realinho et al. benchmark, mutual-information feature selection, and five-fold stratified cross-validation. Its contributions are a controlled five-classifier comparison, a quantified feature-importance hierarchy, evidence that linear classifiers can perform as well as ensembles on this data type, and a reproducible methodology.

SVM achieves the best overall performance (accuracy 97.1%, F1 = 0.954, AUC = 0.993), and all five classifiers achieve AUC above 0.981. Four semester-level academic-

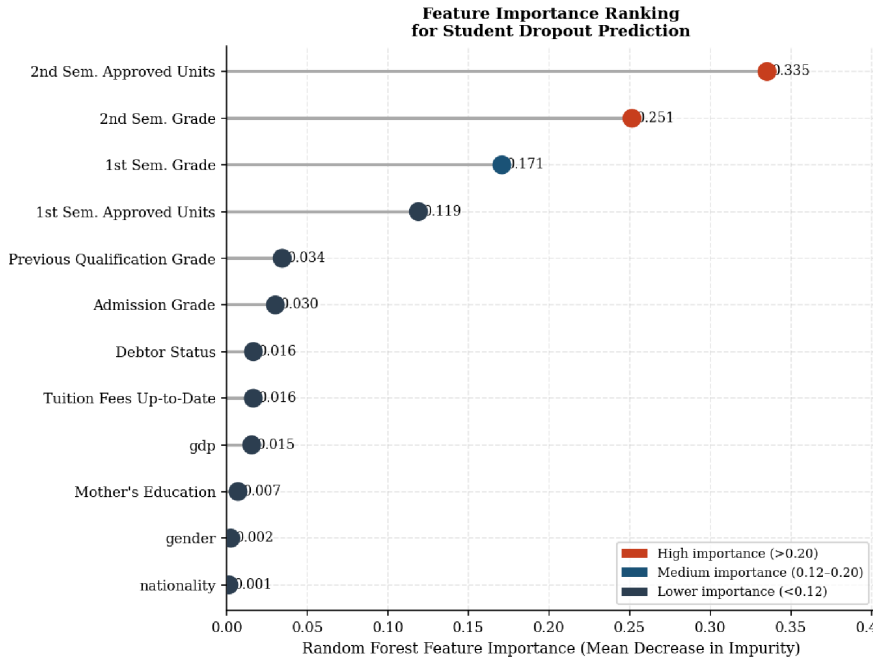


Figure 3. Random Forest feature importance, displayed as a lollipop chart. Red markers indicate high-importance features (> 0.20), dark-blue markers medium importance (0.12–0.20), and dark-grey markers lower importance. The four academic-performance features collectively account for 87.6% of total importance.

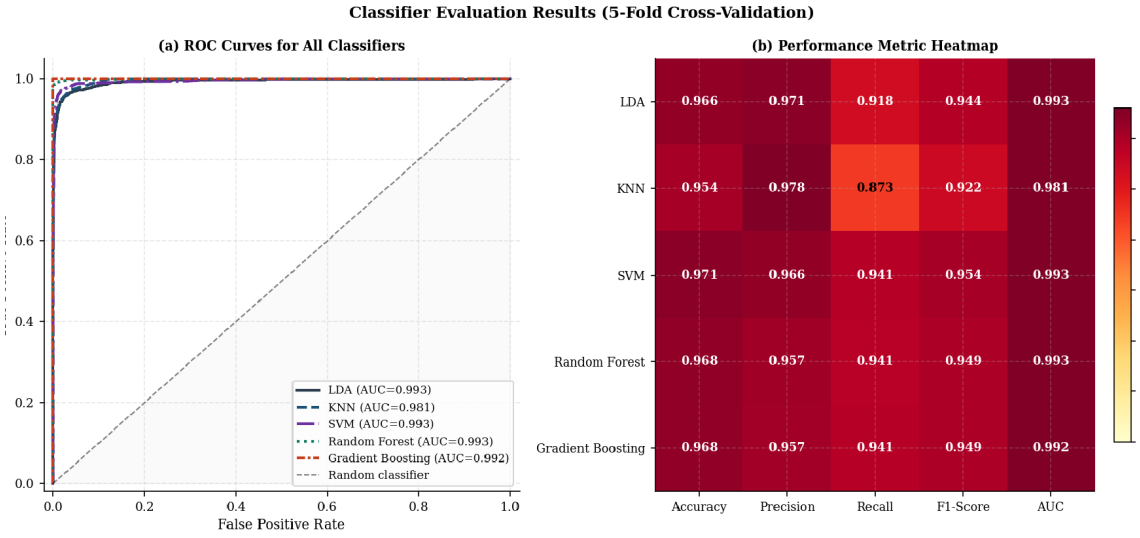


Figure 4. Classifier evaluation. (a) ROC curves for all five models; divergence is concentrated at FPR < 0.10, the deployment-relevant operating region. (b) Performance heatmap across models and metrics (darker = better), confirming SVM's consistent advantage and KNN's recall deficit.

performance metrics account for 87.6% of Random Forest feature importance. The near-equivalent AUC of LDA and the ensemble methods suggests that the decision boundary is approximately linear in the selected feature space, favouring interpretable linear classifiers over opaque ensembles. Socioeconomic features contribute approximately 11% of importance and require careful consideration in deployment.

Institutions can use existing semester-grade records to build effective dropout-prediction systems, without assuming that more complex ensemble models will yield better results. Future work should investigate within-semester trajectory modelling, time-based models, metaheuristic feature-subset search, fairness-aware retraining, and multi-institutional validation.

8. FUTURE WORK

The findings suggest several research extensions. The most important operational question concerns the trade-off between accuracy and timeliness: how well can dropout be predicted using only first-semester data, and how much accuracy is lost by deploying a model before second-semester data are available? This comparison would help institutions choose between a less accurate model available sooner and a more accurate model available later.

Methodologically, replacing univariate mutual-information ranking with a multivariate criterion such as mRMR, or with a wrapper-based search guided by Particle Swarm Optimisation or Genetic Algorithms, could identify synergistic feature combinations that univariate analysis misses [4, 2]. Tempo-

ral models such as LSTM and survival analysis could track academic performance week by week during a semester and enable earlier predictions. Fairness-aware retraining that constrains predicted risk differences across socioeconomic subgroups is another important methodological and policy direction. Finally, multi-institutional validation across national systems, funding frameworks, and institution types is needed to assess whether the feature-importance hierarchy observed here generalises.

REFERENCES

[1] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022.

[2] A. González-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Frontiers in Education*, vol. 8, p. 1244686, 2023.

[3] M. Hlosta, C. Herodotou, T. Papathoma, A. Gillespie, and P. Bergamin, "Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100108, 2022.

[4] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, 2022.

[5] B. Alnasyan, M. Basheri, and M. Alassafi, "The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100231, 2024.

[6] R. Mahafdah, S. Bouallegue, and R. Bouallegue, "Enhancing e-learning through ai: Advanced techniques for optimizing student performance," *PeerJ Computer Science*, vol. 10, p. e2576, 2024.

[7] L. Jin, Y. Wang, H. Song, and H.-J. So, "Predictive modelling with the open university learning analytics dataset (oulad): A systematic literature review," In *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials*, Communications in Computer and Information Science, vol. 2150, Springer, pp. 477–484, 2024.

[8] M.-R. Borna, H. Saadat, A. T. Hojjati, and E. Akbari, "Analyzing click data with ai: Implications for student performance prediction and learning assessment," *Frontiers in Education*, vol. 9, p. 1421479, 2024.

[9] H. Althibyani, "Predicting student success in moocs: A comprehensive analysis using machine learning models," *PeerJ Computer Science*, vol. 10, p. e2221, 2024.

[10] J. H. Guanin-Fajardo, J. Guaña-Moya, and J. Casillas, "Predicting academic success of college students using machine learning techniques," *Data*, vol. 9, no. 4, p. 60, 2024.