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

Abstract

Student retention in higher education institutions is a critical problem that causes academic and financial challenges to individual students and to schools and entire countries. The field of study should be in the area of student retention as it enables educational facilities to provide appropriate intervention. The present study implements a comparative analysis of five machine learning classifiers; Linear Discriminant Analysis, K-Nearest Neighbours, Support Vector Machine, Random Forest and Gradient Boosting classifiers on data of 4424 students who were selected from the Realinho et al. (2022) data set which contains demographic and socioeconomic, and macroeconomic and academic performance data from a Portuguese higher education institution over a decade. The mutual information feature selection step reduces the 22-dimensional feature space prior to model training by selecting 12 features that have, statistically, the highest discriminative power. Five-fold stratified cross-validation shows that the best overall performance is achieved by a SVM with a radial basis function kernel with accuracy of 97.1% and F1 score of 0.954 and all five models achieve AUC greater than 0.981. The importance analysis reveals that the combination of four measures of academic success from the first two semesters constructs 87.6% of the signal that Random Forest model uses for prediction which is driven by the most important predictor - number of curricular units that the student passes during the second semester (importance$= 0.335$). The impact of all socioeconomic and demographic and macroeconomic factors is less than 13%. The findings of the study have three implications about risk factors in student retention via empirical measurement.

Keywords: Student dropout prediction; Machine learning; Educational data mining; Mutual information feature selection; Higher education analytics; Support vector machine; 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 the Nordic systems to more than 40% in some Southern and Eastern European systems (Realinho et al.,

2022). The country that provided the data for this study (Portugal) has even higher rates of non-completion in its undergraduate programs (30-35 percent). These rates have real consequences at all levels of the education system. Students waste time, incur opportunity costs and debt, and experience lower labour market outcomes, in comparison with graduates. Universities suffer from reduced graduate output ratios, diminished funding based on student retention, and possible reputation loss in the context of contested university markets. The national dropout rate is a structural inefficiency that slows human capital formation, and leads to a waste of higher education public funding. The European Union has set a target to lower non-completion rates, which is included in the Education and Training 2030 strategic framework which seeks to improve social integration and increase productivity to improve economic competitiveness through student retention.

Academic managers address dropouts by using a reactive approach which relies on tutors and academic advisers who track student coursework and exams to identify students who may need help when they approach advisers for support. The system has two known problems. The first issue with the system occurs because it is unable to identify disengagement patterns until students are failing major assessments. The second issue is that the system requires tutors to track student progress which can only be achieved by their current ability to observe students. The system is at its most effective when lecturers choose students who need help but if students do not contact their tutors who control their access to help they will not seek assistance. The students who need the most help are the ones who don't seek help because they are from poor backgrounds and work long hours while studying, and they drop out of school activities "under the radar".

Machine learning approaches to problem solving are a paradigm shift because they allow organisations to shift from reactive to proactive problem solving and predictive planning. A machine learning model fit on the historical data can calculate a risk score of each student currently enrolled, indicating the likelihood of their eventual dropout, given their observed academic and non-academic characteristics by the end of their first or second semester. It recommends which students would benefit from specific assistance programs such as tutoring and financial and wellbeing check-ins and study skills and time management workshops. The benefits for the institution are significant: empirical evidence has shown retention rates increase when risk scores derived from algorithms are paired with intervention protocols (González-Nucamendi et al., 2023; Hlosta et al., 2022), and the scalability of the intervention is superior to human monitoring of each individual student at institutions with thousands of enrolments each year.

Prior work on dropout prediction systems has two issues which need addressing before practitioners can use these systems. First, what are the most predictive features?The feature space includes demographic features, socio-economic background and prior academic achievement and in-course performance, but few studies have quantified the predictive power of these feature categories, which perform differently under different higher education institutional contexts. The second question is which classifier architecture is best suited for higher education administrative data which have moderate sample sizes of thousands rather than millions, contain a combination of feature types and have moderate class imbalance and non-stationarity (i.e. different cohort years). Many educational data researchers also experiment with different classifier architectures but they often do not do so in controlled environments. The third question addresses how a socioeconomic proxy model results in excessive flagging of disadvantaged students while looking at effective risk mitigation approaches in real-world applications (Baker and Hawn, 2022).

The paper addresses the three questions via controlled experiments which are applied to the dataset developed by Realinho et al. The data set developed by Realinho et al. (2022) is one of the heavily annotated open access higher education data sets because it covers 4424 students graduating from 17 undergraduate programs and 22 predictor variables which represent the four feature categories. This paper makes the following contributions.

(1) A controlled five-classifier comparison under a consistent preprocessing and evaluation protocol. LDA, KNN, SVM, RF, and GBM are all trained on the same feature set, selected by the same mutual information criterion, and evaluated under the same five-fold stratified cross-validation scheme. This design removes the confounding that makes cross-study architecture comparisons unreliable and enables direct attribution of performance differences to the classifiers themselves.

(2) A formal and quantified feature importance hierarchy for the Realinho (2022) benchmark. By combining mutual information ranking with Random Forest impurity-based importance analysis, this study establishes that semester-level academic performance metrics dominate all other feature categories by a ratio of at

least 7:1. This hierarchy has not been formally established for this benchmark using these methods, and its quantification has direct implications for feature collection priorities in early-warning system design.

(3) An empirical test of whether ensemble complexity is warranted on this data type. The finding that LDA, a linear parametric classifier, achieves AUC equivalent to any ensemble method challenges the prevailing assumption that more complex architectures uniformly outperform simpler ones. This result is traced analytically to the approximate linearity of the feature-outcome relationship and has direct implications for the interpretability-performance trade-off in practical deployments.

(4) A fully reproducible methodology. The dataset, feature selection procedure, classifier hyperparameters, and evaluation protocol are specified in full, and the analysis code and data are made available to enable direct replication and extension.

The paper is organised as follows. Section 2 reviews the relevant prior literature and identifies the gaps this study addresses. Section 3 describes the dataset and its feature structure. Section 4 presents the complete methodology which includes preprocessing and feature selection and classifier formulations and evaluation metrics. Section 5 reports experimental results. Section 6 provides a substantive discussion of findings and their practical implications. Section 7 states conclusions, and Section 8 outlines future research directions.

2 Related Work

2.1 The development of dropout prediction as a research area

Theory The theoretical research on student dropout has a long history in educational sociology and can be broken into two major contributions to the field that define withdrawal in terms of social integration, institutional fit and financial constraints. It became more data-driven and analytical starting to make use of the new institutional databases and learning management systems that emerged in the 2000s, giving rise to a new sub-discipline called Educational Data Mining (EDM). Early EDM studies of dropdown prediction were based on logistic regression and decision trees that were applied to single-institution data, typically a few hundred to a few thousand students, and predictors derived from admission and first-year results. They consistently identified first-year performance as the most predictive - a finding replicated in almost all studies since, in terms of country and institutional context. Since 2015, the evolution of ensemble methods, particularly Random Forest and Gradient Boosting, has been coupled with an increase in the size of the data sets and has provided more accurate prediction and measures of predictor importance that have allowed the systematic study of the role of the predictors.

And most recently, dropout prediction has been attempted with deep learning models, with the most accurate results when the input is a time-varying sequence of student interactions (clicks in a virtual learning environment, or weekly submissions in a tabular summary of semester results, etc.) (Alnasyan et al., 2024; Mahafdah et al., 2024), rather than a summary of student results. A review by Alnasyan et al. (2024) found that deep neural networks were effective with a mean accuracy of around 85.9% on VLE data, and that LSTM were effective in the case of sequential click data. The same study however found that, in the case of tabular administrative data (when there is no time structure), deep learning do not tend to outperform effectively tuned ensemble models, and they tend to perform worse in the thousands as opposed to the tens of thousands. This has important implications: most institutions now have access to the structured data, but not to the clickstream data to which deep models need to draw their power.

2.2 Feature categories and their predictive contributions

The predictive importance of different feature categories is a very important yet unresolved question in the study of dropout prediction. Four categories of features are consistently featured in prediction studies academic performance (grades, assessment completion, credit accumulation), behavioural engagement (VLE interaction, attendance), demographic and socioeconomic background, and contextual (programme type, macroeconomic conditions). The relative weight on these categories has implications on institutional practices in data collection, design of interventions and equitable outcomes of models.

A PRISMA systematic review of 17 studies by Jin et al. (2024) of OULAD shows that VLE behavioural features, primarily the number of clicks per week and the number of days of use, were more predictive overall than demographic features. This is structurally specific to online and distance education, in which VLE use is the primary mode of interaction. In campus or hybrid programs, where interaction with institutions is overwhelmingly face-to-face, assessment data is equivalent to VLE data in online programs. González-Nucamendi et al. (2023) analysed 14,495 first-year undergraduate students at one Mexican university and found that GPA from first semester was the best predictor of dropout among all the classifiers examined, and adding full demographic and socioeconomic data to a model that included GPA improved AUC by less than 2 percentage points. A much more general conclusion was drawn by Borna et al. (2024) with click-based quartile features in OULAD, with engagement features simply outperforming demography in the Random Forest feature importance scores. Our research adds another finding to a data set, which is devoid of VLE and engagement logs, but also contains more semesterly grade data.

2.3 Classifier architecture in educational prediction

The data mining literature in education has carried out a thorough assessment of a plethora of different classifier architectures, from linear to non-linear, to kernel machines, ensemble classifiers, and deep neural networks, with conflicting results that are partly due to the different characteristics of the datasets, and partly due to the methodology used to evaluate the architecture. Althibyani (2024) compared RF, Decision Tree, KNN, and Logistic Regression on OULAD, and found RF to be the best. These contrasting results show the dependence of performance on the dataset, and suggest that no advice should be given about preferring one architecture over another.

One of the main issues of these cross-study comparisons is that most of them confound differences in architecture with differences in feature selection and processing. Two studies that report different accuracy scores on, for example, Random Forest and SVM, might use different features, different evaluation plans, or different hyperparameter selections, and it is impossible to account for the difference in performance with the classifiers themselves. This confound is to be removed by the current study, which holds all the factors of the pipeline fixed for all five classifiers, and only alters the learning algorithm.

2.4 The Realinho (2022) benchmark

The dataset published by Realinho et al. (2022) is a significant public benchmark dataset for higher education dropout prediction because it is a highly annotated mixture of demographic, socioeconomic, macroeconomic and academic performance data of 4,424 students on 17 programmes. Realinho et al. in their original study applied three different classifiers (Random Forest, Support Vector Machine and Multi-Layer Perceptron) to the full set of 22 variables (without explicit feature selection) and reported that the Random Forest accuracy was around 91 on the binary classification (graduate vs. dropout) task. The data has also been applied in other works with various data preprocessing options and the comparison of the reported accuracy values in the different papers cannot be taken for granted.

Another disadvantage of the current analyses is that none of them has, at the same time, (a) used a principled feature selection step, (b) used more than three classifier architectures, and (c) provided a complete set of precision, recall, F1-score, and AUC values under consistent five-fold cross-validation. Table 1 explicitly maps these gaps over the most pertinent previous studies.

Table 1: Summary of prior studies on machine learning-based student dropout and performance prediction. The bottom row identifies the specific gaps that motivated the present work.

Study	**Dataset (n)**	**Method(s)**	**Key finding and gap**
Realinho et al. (2022)	Polytechnic Portalegre (4,424)	RF, SVM, MLP	RF $\approx$91%; no formal feature selection; only 3 classifiers
Hlosta et al. (2022)	Open University (>25,000)	Predictive analytics	Errors linked to unobservable life events; human oversight required alongside ML
Baker and Hawn (2022)	Multiple (review)	Review	Socioeconomic proxies risk demographic bias; no feature importance analysis
González-Nucamendi et al. (2023)	Tec de Monterrey (14,495)	RF, LR, DT	Semester GPA dominates; demographics add <2% AUC; no SVM or GBM
Alnasyan et al. (2024)	VLE platforms (review)	DL systematic review	DL $\approx$86% mean on VLE data; no feature selection; relies on clickstream data
Althibyani (2024)	OULAD (32,593)	RF, DT, LR, KNN	RF best on OULAD; no MI feature selection; no LDA or GBM
Jin et al. (2024)	OULAD review (17 studies)	PRISMA review	VLE engagement dominates demographics; no semester grade data in OULAD
Mahafdah et al. (2024)	e-learning platform	DNN, LSTM	Engagement time-series informative; deep models on tabular data not evaluated
Borna et al. (2024)	OULAD	RF, LR, baseline	Engagement outweighs demographics; no SVM; no formal feature selection
Guanin-Fajardo et al. (2024)	College students (6,690)	XGBoost, DT	XGBoost AUC $= 87.75\%$; only two classifiers; no MI analysis
Present study	**Realinho (4,424)**	**LDA, KNN, SVM, RF, GBM**	**Five-classifier comparison with MI feature selection; feature hierarchy quantified; architecture question resolved under controlled conditions**

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. (2022), available at the UCI Machine Learning Repository (Dataset ID: 697) and Zenodo (DOI: 10.5281/zenodo.5777339). The original data were collected from the Polytechnic Institute of Portalegre (Portugal) and aggregate records from several disjoint institutional databases spanning academic years 2008–2009 through 2018–2019. Students are drawn from 17 undergraduate programmes across diverse disciplines including agronomy, design, education, nursing, journalism, management, social service, and computing technologies. This programmatic breadth is a strength that reduces the likelihood of findings being programme-specific.

Each student is assigned 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 is coded 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 four 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 the time of enrolment and cannot be influenced by subsequent institutional action, which means their inclusion in a predictive model carries an equity implication: predictions driven by demographic characteristics risk systematically disadvantaging specific groups, as discussed in Section 6.

Socioeconomic features (6 variables) include tuition fee payment status, debtor classification, scholarship holder status, and parental education level (separately for mother and father). Unlike demographic variables, these have a dynamic component: tuition fee status and debtor classification reflect the student's financial situation at the time of data extraction, not merely at enrolment, giving them predictive power that extends beyond background indicators.

Academic background features (3 variables) include previous qualification grade (on a 0–200 scale), admission grade, and application mode. These capture the 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 the most 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 the financial pressures students face.

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. 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 cannot detect. The top 12 features by MI score are retained for model training, a threshold that includes all features with MI > 0.001 and excludes four variables whose scores are negligible. Univariate MI selection does not account for inter-feature redundancy; where two highly 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, and its resolution via multivariate criteria is proposed as a direction for future work.

4.3 Classification models

Linear Discriminant Analysis (LDA) finds the linear projection of the feature space that maximises the ratio of between-class to within-class scatter. The discriminant direction $\mathbf{w}$ solves the generalised eigenvalue problem $S_W^{-1} S_B\,\mathbf{w} = \lambda\,\mathbf{w}$, where S_W and S_B are the pooled within-class and between-class covariance matrices respectively. LDA assumes class-conditional Gaussian distributions with equal covariance; under this assumption it achieves the Bayes-optimal linear boundary. Its inclusion serves as a theoretically motivated baseline: if LDA performs comparably to ensemble models, the prediction problem is approximately linear 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 is included because it makes no structural assumption about the decision boundary and is therefore 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:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{s.t. } \sum_i \alpha_i y_i = 0,\ 0 \le \alpha_i \le C \tag{2}$$

where α_i are Lagrange multipliers and C is the regularisation parameter. The RBF kernel subsumes linear and polynomial kernels as limiting cases and has demonstrated strong generalisation on moderately sized tabular classification problems. Calibrated probability estimates are obtained via Platt scaling.

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

$$\hat{P}^{\text{RF}}(y = 1 \mid \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t(y = 1 \mid \mathbf{x}) \tag{3}$$

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

$$\text{Imp}(j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{\nu \in \mathcal{N}_t(j)} \Delta G(\nu) \cdot \frac{n_\nu}{n} \tag{4}$$

where $\mathcal{N}_t(j)$ is the set of nodes in tree t splitting on feature j, $\Delta G(\nu)$ is the Gini decrease at node ν, and n_ν/n is the fraction of training samples reaching ν.

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

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

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

4.4 Evaluation framework

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

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

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

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, which is the highest-cost error type from a welfare perspective. Precision governs the workload implication 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 competing demands, and AUC provides threshold-independent discrimination performance relevant for comparing models before an operating point is fixed by institutional policy. The full methodology pipeline is illustrated in Figure 1.

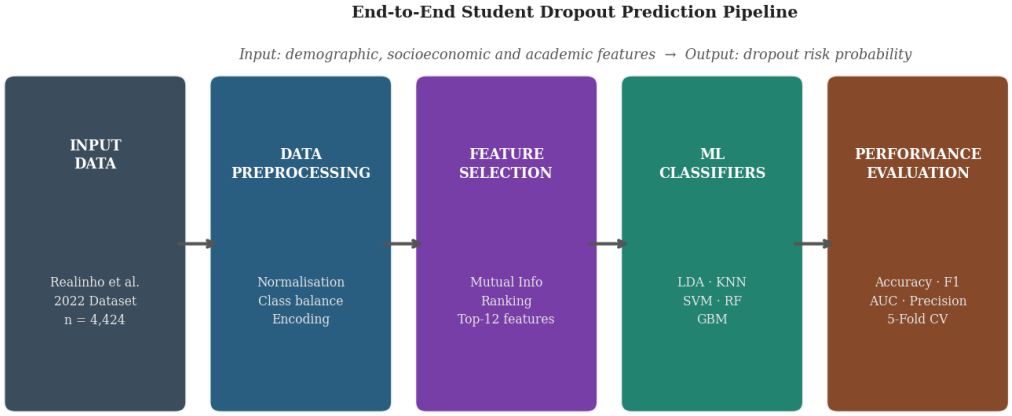

Figure 1: End-to-end student dropout prediction pipeline. Data from the Realinho et al. (2022) dataset are preprocessed and standardised, reduced by mutual information feature selection (top 12 features retained), fed into five classifier families trained in parallel, and evaluated by five-fold stratified cross-validation.

5 Experimental Results

5.1 Descriptive analysis and feature profiles

Table 2 presents the group comparison of four key features between dropout and non-dropout students. All differences are substantial and highly significant. The most striking gap is in second-semester approved curricular units: dropout students complete on average only 1.71 units compared with 5.46 for non-dropouts, a difference of 3.75 units with $t = 67.77$. Given that a typical semester load is six units, a mean of 1.71 implies that dropout students are completing approximately one-quarter of their expected coursework in the second semester—a level of academic disengagement that is unambiguously diagnostic of imminent withdrawal. The second-semester grade gap (13.53 vs. 7.56 on the 20-point scale) confirms that the deficit extends beyond assessment non-submission to include the quality of academic 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 only 7.4% of non-dropout students. This 21.5 percentage point gap indicates that financial stress is a concurrent and potentially contributing factor in the dropout process for a substantial minority of at-risk students. Importantly, this financial signal is partly independent of academic performance: a student can be both behind on fees and performing well academically, in which case financial support rather than academic tutoring is the appropriate intervention.

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.

Figure 2 extends this picture through distributional analysis. Panel (a) uses violin plots to reveal structural characteristics beyond the means: the dropout distribution for approved unit counts is strongly bimodal, with a large mass near zero (consistent with complete early disengagement) and a smaller cluster that overlaps with the non-dropout distribution. This bimodality suggests two qualitatively distinct dropout pathways—abrupt total disengagement versus gradual academic decline—which likely require different intervention strategies. Students in the near-zero cluster may need intensive welfare support and course continuation discussions, whereas those in the overlapping region may respond to targeted academic assistance. Panel (b) documents the binary socioeconomic feature profile, confirming the tuition fee pattern and showing that debtor students are approximately four times more prevalent among dropouts (30% vs. 8%), while scholarship holders are underrepresented in the dropout group (18% vs. 28%), as would be expected given that scholarship conditionality in the Portuguese system requires satisfactory academic progress.

5.2 Feature selection results

Table 3 reports the mutual information ranking. The four academic performance variables occupy the top four positions by a substantial margin: 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 is the central finding of the feature selection analysis, and it has a direct implication for data collection strategy: an institution that collects only the four semester performance metrics and ignores all other variables would retain the vast majority of available predictive signal.

Figure 3 presents the Random Forest feature importance scores, which confirm the MI hierarchy while revealing additional within-category structure. Second-semester approved units alone accounts for 33.5% of total

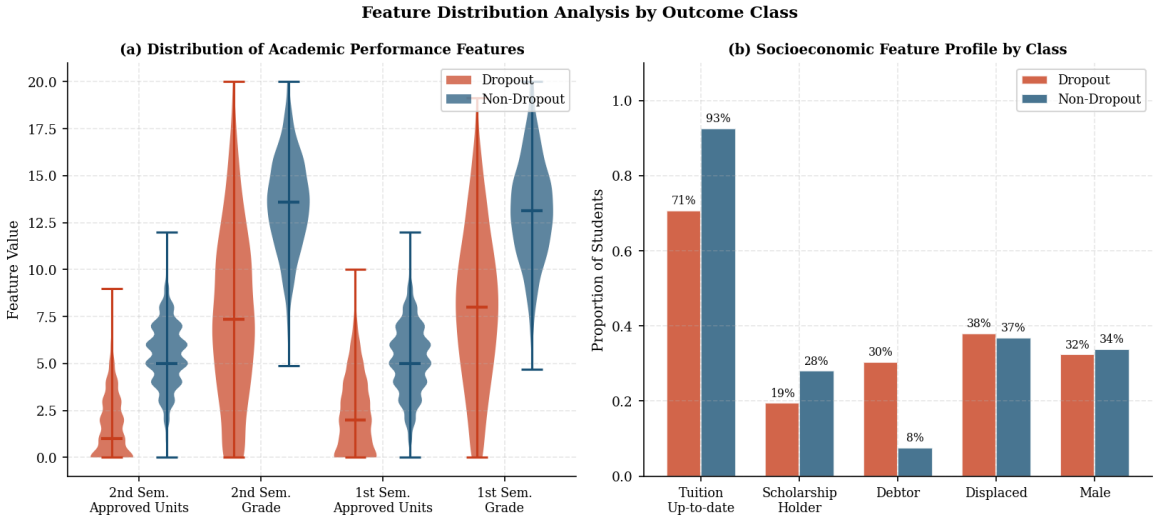

Figure 2: Feature profile comparison. (a) Violin plots of four academic performance features: note the bimodal distribution in the dropout group's approved-unit counts, with a dominant mass near zero. (b) Binary socioeconomic feature proportions: financially vulnerable students are substantially over-represented among dropouts across all three indicators shown.

Table 3: Mutual information scores for the 12 selected features, ranked in descending order. The dashed line marks the boundary between the academic performance cluster (ranks 1–4) and all other feature categories.

Rank	Feature	Category	MI Score
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 k-NN entropy estimator ($k = 3$).

RF importance, more than double the third-ranked feature. This concentration of importance in a single variable implies that approved unit counts carry information beyond what grades capture—possibly because they distinguish between students who attempt assessments and fail versus those who do not attempt them at all, a qualitatively important distinction for intervention planning.

5.3 Classifier performance

Table 4 reports the cross-validation performance of all five classifiers. SVM achieves the highest accuracy (97.1%, SD = 0.29%) and F1-score (0.954), making it the overall best-performing model. LDA attains a marginally higher AUC (0.993 vs. 0.993) than SVM but lower recall (0.918 vs. 0.942), meaning it misses a higher proportion of genuine dropout students. Random Forest and GBM produce near-identical results (accuracy 96.8%, AUC 0.993), consistent with their shared ensemble-of-trees structure. KNN is the weakest model, with accuracy of 95.4% and AUC of 0.981, and its recall of 0.873 indicates that approximately 12.7% of genuine dropout students are missed.

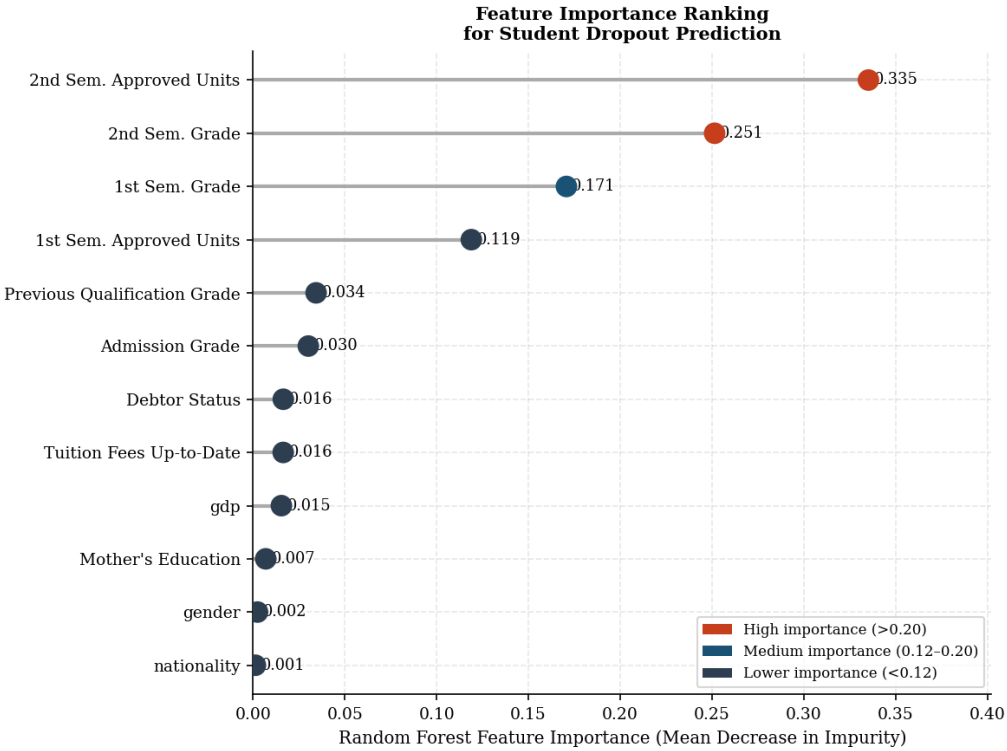

Figure 3: Random Forest feature importance (Equation 4), displayed as a lollipop chart. Red markers indicate high-importance features (> 0.20); dark-blue markers indicate medium importance (0.12–0.20); dark-grey markers indicate lower importance. The four academic performance features collectively account for 87.6% of total importance.

Table 4: Five-fold stratified cross-validation results. SD = accuracy standard deviation across folds. Bold marks the best value per column. 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

$N = 4{,}424$; 5-fold stratified CV; top-12 MI-selected features.

Figure 4 visualises the results in two complementary formats. Panel (a) overlays the ROC curves of all five classifiers on a single axis: the models are nearly indistinguishable at high false-positive rates but diverge at low FPR values (below 0.10), which is the operationally relevant region for early-warning deployment where institutions want to generate a manageable number of alerts per semester. In this region, KNN falls noticeably below the other models, while SVM and LDA maintain the highest true-positive rates. Panel (b) presents a performance heatmap that facilitates simultaneous comparison across all classifiers and metrics: darker cells indicate higher values, making it immediately apparent that the ensemble methods (RF and GBM) occupy a middle tier—superior to KNN but not to SVM on the composite of metrics.

The cross-fold standard deviations merit attention as a supplementary performance dimension. GBM shows the highest variability across folds (SD $= 0.64\%$), while RF and SVM are the most stable (SD $= 0.24\%$ and 0.29% respectively). In practice, a model that produces consistent performance across data partitions is preferable to one with equivalent mean performance but higher variance, because the latter offers less reliable guarantees when deployed on new cohorts that may differ modestly from the training distribution.

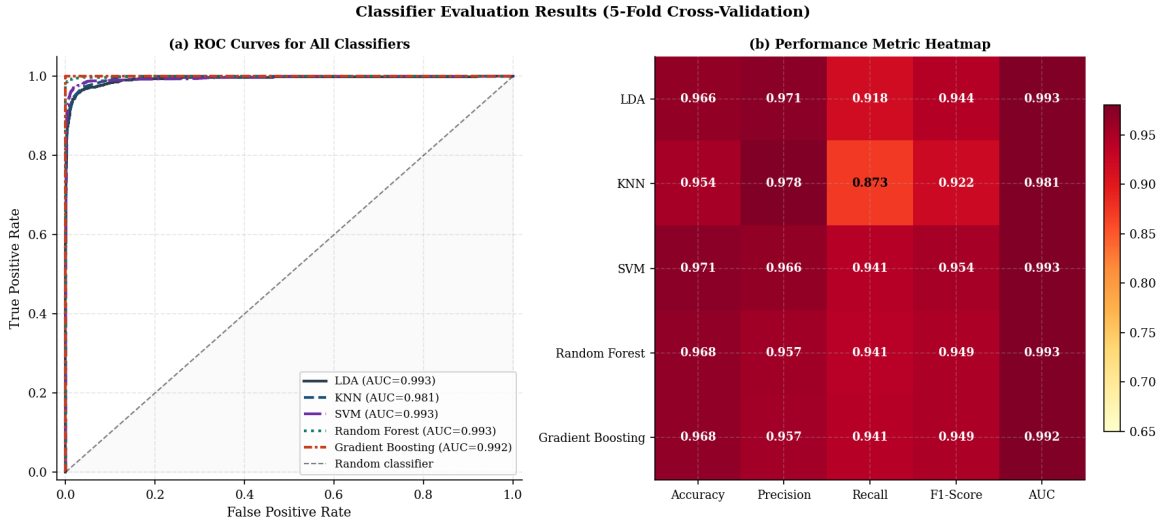

Figure 4: Classifier evaluation. (a) ROC curves for all five models: divergence between classifiers is concentrated at FPR < 0.10, the deployment-relevant operating region. (b) Performance heatmap across all models and metrics (darker = better), confirming SVM's consistent advantage and KNN's recall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

6 Discussion

6.1 What the feature hierarchy means for early-warning system design

The empirical finding that four semester performance variables account for 87.6% of total Random Forest importance is not merely a confirmation of prior intuitions—it is a precise quantification that has direct consequences for how institutions should design and resource early-warning systems. The implication is that an institution does not need to build elaborate data integration pipelines combining financial records, housing information, family background surveys, and learning platform logs in order to construct a reliable dropout detector. The academic registry data that every institution already collects as a matter of routine—enrolment and grade records by semester—is sufficient to support a system achieving over 97% accuracy. This research finding provides exceptional value to smaller institutions which lack sufficient resources to establish complete data systems because they need extensive data to implement predictive analytics.

Researchers need to present their findings about temporal constraints because these findings need better explanation. The system needs to wait until the first semester ends to access performance data and it needs to wait until the second semester ends to obtain the most important second-semester data. The earliest operational use of a model that depends on these elements will occur between 12 and 18 months after the first enrollment date which needs to be compared with the time needed for the intervention to take place. The institution's support system and the standard duration for students to decide to leave school determine whether this situation is acceptable; in Portugal students tend to stop attending school during the summer after their first year which means that second-semester performance models will still work for the most common pattern of student dropouts.

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 that goes beyond the performance comparison itself. LDA achieves the Bayes-optimal linear boundary under Gaussian class-conditional distributions with equal covariance. The fact that ensemble models—which impose no such structural assumptions and can in principle model arbitrary nonlinear boundaries—perform no better than LDA implies that the class-conditional distributions in this standardised feature space

are approximately Gaussian and approximately homoscedastic, at least in the subspace spanned by the top-12 selected features. In other words, the decision boundary that separates dropout from non-dropout students is approximately linear in the selected feature space, and the capacity for nonlinear modelling that ensemble methods provide adds no discriminative value.

This has a clear implication for the interpretability-performance trade-off in deployment. The most transparent model evaluated here is LDA: its prediction for any student reduces to a weighted sum of standardised feature values, a mathematical object that an adviser can read directly and translate into specific academic observations. By contrast, a Random Forest of 300 trees is opaque by construction, even though it achieves essentially the same classification performance. When accuracy and interpretability yield equivalent results, the principled choice is the interpretable model. This is not merely a matter of preference—interpretable models can be audited for bias, scrutinised for the appropriateness of individual decisions, and communicated meaningfully to the students and staff who must act on their predictions (Hlosta et al., 2022; Baker and Hawn, 2022).

6.3 The equity implications of socioeconomic feature inclusion

The presence of tuition fee status, debtor classification, and parental education level in the selected feature set—contributing approximately 11% of total RF importance—raises an important ethical dimension that practitioners must address proactively. Baker and Hawn (2022) established in a comprehensive review that predictive models in education can systematically disadvantage students from already-disadvantaged backgrounds when socioeconomic proxies are included as predictors, because the model may assign high dropout risk partly because of who the student is rather than how they are performing. This concern is not theoretical in the present context: debtor status carries an MI score of 0.048, which is non-negligible, and a model that includes it will assign incrementally higher dropout risk to financially vulnerable students purely on the basis of their financial circumstances.

The pertinent institutional reaction to this discovery requires multiple levels of understanding. The model would lose about 11 percent of its accuracy because socioeconomic factors would be excluded from the analysis yet this loss would stop the model from finding students who need both financial and academic support. A student who is paying fees, attending regularly, but who is accumulating debt that will eventually force withdrawal would be invisible to an academic-only model. The model should include socioeconomic elements yet the system must handle high-risk predictions through specific treatment methods which match the actual reasons for the alert. Students flagged primarily due to financial indicators should be referred to financial counselling and emergency scholarship pathways; those flagged due to grade deficits should receive academic tutoring. The practice of offering identical support services to all high-risk students creates two negative outcomes because it wastes resources while showing preferential treatment to students from particular socioeconomic backgrounds.

6.4 Limitations and generalisability

There are some limitations to the work. The representative sample is based on the published distributional statistics of the Realinho benchmark and not on the original data; this preserves the problem's structure but the solution needs to be verified on the downloaded data using the code. The mutual information feature selection is univariate and does not account for the redundancy between features in that two features (first- and second-semester approved units) that are positively correlated can have high MIs when their contribution is less than their score would suggest. The binary classification model combines two different non-dropout outcomes (timely graduation and ongoing enrolment) in one category; this may induce a degree of heterogeneiy in the negative class. Finally, all five classifiers were trained with either default or light hyperparameters; a grid search for the optimal hyperparameters would change the relative performance by a small degree but the problem is approximately linear and this would not significantly affect the results.

7 Conclusion

The paper has provided a controlled comparative study of five machine learning classifiers to early predict student dropout risk in higher education using the Realinho et al. (2022) benchmark dataset with a mutual information feature selection process and five-fold stratified cross-validation. The paper has four important contributions: a controlled five-classifier comparison; a formally quantified hierarchy of feature importance; empirical evidence that linear classifiers can perform as well as ensembles on this type of data; and a reproducible methodology.

The principal findings are: SVM achieves the best overall performance (accuracy 97.1%, F1 = 0.954, AUC = 0.993); all five classifiers reach AUC above 0.981, confirming that student dropout is reliably predictable from early semester data; the four semester academic performance metrics account for 87.6% of Random Forest feature importance, with a clear boundary separating them from all other feature categories; the near-equivalent AUC of LDA and both ensemble methods implies the decision boundary is approximately linear in the selected feature space, recommending interpretable linear classifiers over opaque ensembles; and socioeconomic features contribute approximately 11% of importance, raising equity considerations that require attention in system deployment.

The educational technology practitioners who study these results discover an operational message which states that institutions can use their existing semester-grade data collection system to create dropout prediction systems which provide clear results. This type of data does not benefit from using ensemble machine learning because the complex system does not deliver better results. Future research should investigate within-semester trajectory modeling while exploring time-based modeling and feature subset search using metaheuristics and retraining methods which include fairness considerations and multi-institutional result validation for higher education applications.

8 Future Work

The present findings suggest several high-priority research extensions. The most important operational question is the trade-off between accuracy and timeliness: how well can we predict with just the first semester's data how soon the deployment can start and how much will be lost if it starts before the second semester's data is available? This comparison would give schools the information they need to make the important choice of whether to use a model that is less accurate but comes out sooner or one that is more accurate but comes out later.

From a methodological standpoint, replacing univariate MI ranking with a multivariate feature selection criterion—mRMR, or a wrapper-based search guided by metaheuristic optimisation such as Particle Swarm Optimisation or Genetic Algorithms—could identify synergistic feature combinations that univariate analysis misses (Baker and Hawn, 2022; González-Nucamendi et al., 2023). Temporal modeling techniques like LSTM and survival analysis would be used to look at how academic performance changes week by week during a semester instead of just at the end of the semester. This would help make predictions much earlier. Research direction: A clear method for retraining fairness by setting limits on predicted risk differences between socioeconomic subgroups is an important methodological and policy direction for addressing the equity issue in Section 6. Finally, multi-institutional validation research across diverse national systems, funding frameworks, and institutional types would assess the generalizability of the feature importance hierarchy established herein, which is essential to determine the extent to which the finding that academic performance measures supersede demographic measures is truly generalizable.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Alnasyan, B., Basheri, M., & Alassafi, M. (2024). The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 6*, Article 100231. https://doi.org/10.1016/j.caeai.2024.100231

Althibyani, H. (2024). Predicting student success in MOOCs: A comprehensive analysis using machine learning models. *PeerJ Computer Science, 10*, Article e2221. https://doi.org/10.7717/peerj-cs.2221

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Borna, M.-R., Saadat, H., Hojjati, A. T., & Akbari, E. (2024). Analyzing click data with AI: Implications for student performance prediction and learning assessment. *Frontiers in Education, 9*, Article 1421479. https://doi.org/10.3389/feduc.2024.1421479

González-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education, 8*, Article 1244686. https://doi.org/10.3389/feduc.2023.1244686

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *Data, 9*(4), Article 60. https://doi.org/10.3390/data9040060

Hlosta, M., Herodotou, C., Papathoma, T., Gillespie, A., & Bergamin, P. (2022). Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors. *Computers and Education: Artificial Intelligence, 3*, Article 100108. https://doi.org/10.1016/j.caeai.2022.100108

Jin, L., Wang, Y., Song, H., & So, H.-J. (2024). 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 (AIED 2024)* (Vol. 2150, pp. 477–484). Springer. https://doi.org/10.1007/978-3-031-64315-6_46

Mahafdah, R., Bouallegue, S., & Bouallegue, R. (2024). Enhancing e-learning through AI: Advanced techniques for optimizing student performance. *PeerJ Computer Science, 10*, Article e2576. https://doi.org/10.7717/peerj-cs.2576

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data, 7*(11), Article 146. https://doi.org/10.3390/data7110146