Predicting University Dropout from Pre-Enrollment Data: A Comparison of Interpretable Machine Learning Models

Alberto Cecchi
PhD Candidate
Alma Mater Europaea University
Maribor, Slovenia
alberto.cecchi@almamater.si

Matjaž Gams
Institut »Jožef Stefan«
Jamova 39
Ljubljana, Slovenija
matjaž.gams@ijs.si

Abstract

University dropout remains one of the most persistent challenges in European higher education, particularly in open-access systems where socioeconomic vulnerabilities compound early attrition. Most predictive approaches rely on post-enrollment performance data, which arrives too late for low-cost intervention. This paper tests two hypotheses: (H1) that dropout risk can be predicted above chance using exclusively pre-enrollment features, each validated for temporal availability at the admission stage; and (H2) that simpler, inherently interpretable models can match the performance of more complex ensemble methods. Using the UCI dataset from a Portuguese institution, we compare Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). A one-sample Wilcoxon signed-rank test indicates that every model's per-fold AUC exceeds the 0.50 baseline ($p = 0.031$, the minimum attainable with five folds), supporting H1. A paired Wilcoxon test finds no detectable performance difference between LR and RF ($W = 7.0$, $p = 1.00$; mean F1 difference +0.0002), consistent with H2 and with the preference for simpler models under the EU AI Act. SHAP analysis identifies scholarship status, age at enrollment, and gender as the dominant predictors, with scholarship absence and older age acting as additive, independent risk factors. These findings indicate that meaningful predictive signal exists before enrollment and that interpretable models can support proactive intervention strategies.

Keywords

educational data mining, student dropout, pre-enrollment prediction, interpretable machine learning, explainable AI, SHAP, EU AI Act, temporal validity

1 Introduction

University dropout is one of the most persistent challenges facing higher education in Europe. Within the European Higher Education Area, dropout rates remain particularly high in countries with open-access admission systems such as Italy and Slovenia, where structural and socioeconomic vulnerabilities contribute to significant early attrition [11, 2]. Despite decades of research, the problem remains only partially addressed, in part because the dominant approaches in educational data mining rely on post-enrollment behavioral signals — such as first-semester grades and attendance — that become available too late for timely intervention [15, 13].

A smaller but growing body of work has begun to explore whether meaningful predictive signal can be found in data already available at the admission stage. Nielsen et al. demonstrate feasibility using Danish application-time data, and Aina shows that household cultural and financial background carries significant weight in shaping university dropout in Italy [10, 1]. However, these studies do not perform explicit temporal validation of the features they employ, nor do they frame their findings within the regulatory constraints introduced by the EU AI Act, which classifies educational decision-support systems as high-risk and requires transparency and auditability [7].

This paper addresses both gaps. Using the UCI dataset from a Portuguese higher education institution [12], we compare three interpretable machine learning models — Logistic Regression, Decision Tree, and Random Forest — trained exclusively on features whose availability at the admission stage has been individually verified. The study tests two hypotheses. **H1 (Pre-enrollment sufficiency)**: pre-enrollment data alone carries sufficient predictive signal to identify students at risk of dropout at rates significantly above chance. **H2 (Simplicity–performance equivalence)**: inherently interpretable models perform comparably to more complex ensemble methods. A one-sample Wilcoxon signed-rank test assesses H1, and a paired Wilcoxon test compares Logistic Regression and Random Forest for H2. SHAP analysis identifies the dominant predictors and their directional effects.

The paper makes three contributions. First, it introduces a feature-level temporal validation procedure tailored to admission-stage prediction, ensuring that every input variable is genuinely observable before enrollment. Second, it provides a direct comparison between simple interpretable models and a stronger ensemble baseline under the same pre-enrollment constraint. Third, it interprets the resulting predictions using SHAP and discusses their implications for transparency and compliance in educational decision support. The remainder of the paper is organized as follows. Section 2 reviews related work, Section 3 describes the data and feature selection, Section 4 outlines the methodology, Section 5 presents the results, and Sections 6 and 7 discuss the findings and conclude the paper.

2 Related Work

Dropout prediction is a well-established area within educational data mining. Comprehensive surveys document a wide range of approaches, the majority of which rely on post-enrollment behavioral data — first-semester grades, attendance, and learning management system activity [13, 3]. While these models often achieve strong performance, they depend on information that becomes available only after the critical window for early intervention has passed [15].

A smaller body of work has explored dropout prediction using data available at or before admission. Nielsen et al. show that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.29

Danish admission-time records can support meaningful classification, while Aina documents how parental background and household financial conditions shape dropout in the Italian system [10, 1]. These studies confirm the viability of an upstream approach, but neither performs explicit temporal validation of individual features nor addresses the regulatory context of the EU AI Act [7].

The question of interpretability is particularly relevant. Rudin argues that in high-stakes domains, inherently interpretable models should be preferred over black-box models with post-hoc explanations, as the latter introduce an additional layer of approximation that can obscure the true decision logic [14]. Barredo Arrieta et al. provide a taxonomy of explainability approaches, while Lundberg and Lee introduce SHAP as a theoretically grounded framework for feature attribution [4, 8]. In the educational domain, the tension between model complexity and interpretability remains underexplored, especially in connection with emerging legal requirements [6]. The UCI Predict Students' Dropout and Academic Success dataset provides a shared benchmark, but prior studies have not combined feature-level temporal validation with an EU AI Act compliance framing.

3 Data

3.1 Dataset and Preparation

The analysis uses the UCI Predict Students' Dropout and Academic Success dataset, accessed via the UCI Machine Learning Repository (dataset ID 697) [12]. It contains 4,424 records from a Portuguese higher education institution, described by 36 features and a three-class target variable (Dropout, Enrolled, Graduate). Students classified as still enrolled (n = 794) were excluded, as their outcome remains unresolved. This yields a binary classification task — Dropout (1) vs. Graduate (0) — with 3,630 observations and a class distribution of approximately 61% graduates and 39% dropouts. The dataset contains no missing values. Nominal categorical variables were explicitly cast to a categorical dtype before one-hot encoding, ensuring that integer-coded categories are treated as nominal rather than ordinal. Class imbalance was addressed through the `class_weight=balanced` parameter during model training.

3.2 Temporal Validation and Feature Selection

Since this study focuses exclusively on pre-enrollment prediction, each of the 36 original variables was assessed for genuine observability at or before admission. The guiding criterion is straightforward: if a feature is not available at the point when an intervention decision would be made, its inclusion compromises the practical validity of the model.

Validation proceeded in two stages. First, all curricular unit variables from the first and second semesters — units enrolled, approved, evaluated, and completed, as well as corresponding grade averages — were excluded as unambiguous post-enrollment signals (12 features). Second, two temporally ambiguous variables were removed: *Tuition fees up to date* and *Debtor*. Although both carry financial information, each presupposes an active enrollment relationship — a student cannot be classified as a debtor before being formally enrolled.

The resulting feature set comprises 22 variables spanning five categories (Table 1): demographic attributes, admission attributes, academic background, socioeconomic indicators, and macroeconomic context variables. After one-hot encoding of the nominal categories, the 22 original variables expanded to 215 feature columns. This design keeps the modeling setup close to the information realistically available to admissions offices and makes the prediction setting explicit before model development. In operational terms, the prediction is computed once, at the admission decision point, from data an admissions office already holds; its output is a risk flag routed to a human-mediated offer of support — academic advising, need-based aid, or mentoring — not an input to the admission decision itself, keeping the system within the assistive, human-oversight role the EU AI Act expects of high-risk educational tools.

Table 1: Pre-enrollment feature set: the 22 retained variables, all verified as observable at or before admission.

Category	Variables
Demographic (7)	marital status, nationality, age at enrollment, gender, displaced, educational special needs, international
Admission (2)	application mode, application order
Academic background (5)	course, daytime/evening attendance, previous qualification, previous qualification grade, admission grade
Socioeconomic (5)	mother's / father's qualification, mother's / father's occupation, scholarship holder
Macroeconomic (3)	GDP, unemployment rate, inflation rate

4 Methods

4.1 Model Selection

Three models were selected to represent increasing complexity within the family of interpretable classifiers. Logistic Regression serves as the simplest baseline, with directly readable coefficients. Decision Tree introduces non-linear decision boundaries while remaining visually interpretable. Random Forest aggregates multiple trees for improved generalization, at the cost of reduced direct interpretability. As a strong and widely used ensemble benchmark, RF provides a practically meaningful comparison point for simpler interpretable models. This progression allows a structured comparison: if a more complex model does not significantly outperform a simpler one, the transparency advantage becomes decisive — particularly under the EU AI Act, where inherently interpretable models reduce the need for post-hoc approximation [14].

4.2 Training and Evaluation

All models were trained on a stratified 80/20 split (training n = 2,904; held-out test n = 726; `random_state = 42`). LR used standard scaling and `max_iter = 1000`; DT used `max_depth = 5` and `min_samples_leaf = 20`; RF used 100 estimators and `max_depth = 10`. All models used `class_weight=balanced`. Performance was assessed using 5-fold stratified cross-validation on the training set, followed by evaluation on the held-out test set. Primary metrics are F1-score and AUC-ROC for the Dropout class. F1 captures the trade-off between identifying at-risk students (recall) and avoiding unnecessary intervention (precision); AUC-ROC complements F1 by measuring discriminative ability across all classification thresholds.

4.3 Statistical Comparison

Two non-parametric tests were applied. For H1, a one-sample Wilcoxon signed-rank test compared each model's per-fold AUC-ROC scores against the 0.50 random baseline (one-sided alternative). For H2, a paired Wilcoxon signed-rank test compared the per-fold F1 scores of LR and RF. The Wilcoxon test was chosen

Table 2: Model performance for the dropout class.

Model	F1 CV	AUC CV	F1 Test	AUC Test	Recall
LR	0.668 ± 0.022	0.783 ± 0.022	0.643	0.762	0.722
DT	0.645 ± 0.009	0.759 ± 0.021	0.619	0.725	0.799
RF	0.668 ± 0.019	0.800 ± 0.016	0.667	0.795	0.694

in both cases because the small number of folds ($k = 5$) does not permit a reliable assumption of normality. This also limits the resolution of the inferential claims. With $k = 5$, the minimum achievable p-value for the one-sided test is $1/32 \approx 0.031$ — a limit relevant to the interpretation of H1 below.

4.4 Feature Importance Analysis

SHAP (SHapley Additive exPlanations) TreeExplainer was applied to the Random Forest on the held-out test set [8]. SHAP was preferred over built-in feature importance because it provides both global importance rankings and directional effects. The analysis was applied to RF rather than LR because SHAP beeswarm plots offer a richer visualization of non-linear and interaction effects across the feature space.

5 Results

5.1 Model Performance

Table 2 reports performance across the three models under both cross-validation and held-out test set evaluation.

The three models achieve closely comparable F1 scores in cross-validation (0.645–0.668), with overlapping standard deviations; LR and RF are indistinguishable at the third decimal (both 0.668). RF shows a moderate advantage in AUC-ROC (0.800 CV, 0.795 test), reflecting stronger discriminative ability across thresholds. Test set results are consistent with cross-validation estimates, confirming stable generalization.

For H1, all three models achieve AUC values well above 0.50, ranging from 0.759 to 0.800 in cross-validation and from 0.725 to 0.795 on the test set. The one-sample Wilcoxon signed-rank test yields $p = 0.031$ for all three models, which is the minimum achievable under the test with $k = 5$ folds — indicating that every per-fold AUC exceeds the 0.50 baseline for every model. The consistency between cross-validation and test estimates confirms that the signal generalizes.

For H2, the paired Wilcoxon test on per-fold F1 scores of LR and RF finds no detectable difference ($W = 7.0$, $p = 1.00$; mean difference +0.0002). While a non-significant result does not by itself constitute positive evidence of equivalence, a value of $p = 1.00$ should not be over-interpreted as proof of identical performance; still, the near-zero effect size and the identical cross-validation F1 means are empirically consistent with H2. Under these conditions, model selection can be guided by transparency and auditability rather than marginal performance gains.

From a policy perspective, DT achieves the highest recall on the test set (0.80), followed by LR (0.72) and RF (0.69). If missed cases (false negatives) carry a disproportionately high cost in an intervention context, this further supports the case for simpler models in operational deployment, even when their F1 is slightly lower due to reduced precision.

5.2 SHAP Feature Importance

Figure 1 presents the SHAP analysis of the Random Forest on the held-out test set. The importance ranking reveals a clear

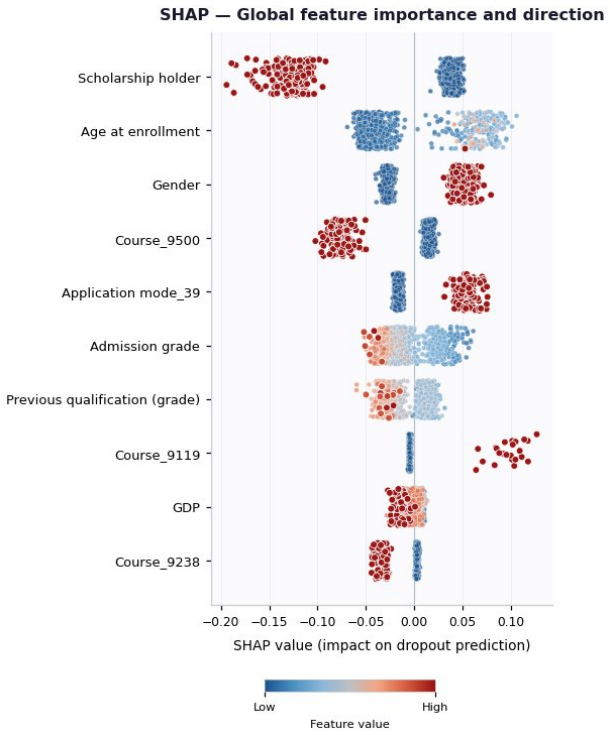

Figure 1: SHAP analysis of pre-enrollment predictors for the Random Forest on the held-out test set. Each dot represents one observation; color indicates feature value (blue = low, red = high); horizontal position shows the SHAP contribution to dropout prediction. Features are ordered by mean |SHAP|.

hierarchy. Scholarship holder is the dominant predictor (mean |SHAP| = 0.062): holding a scholarship consistently pushes predictions away from dropout (median SHAP = −0.126), while its absence shifts them toward it (median SHAP = +0.041) — a separation of 0.167 SHAP units. This is consistent with the interpretation that financial support acts as a stabilizing factor in retention [5]. Age at enrollment ranks second, with higher age associated with increased risk, aligning with the literature on non-traditional students who face competing work and family demands [15]. Gender ranks third, with male students showing increased predicted risk. This signal should be interpreted cautiously as a predictive association rather than a causal mechanism, and any operational use would require explicit fairness review.

Within the no-scholarship subgroup, the median SHAP contribution is +0.043 for older students (> 25 years, the 75th-percentile cutoff) and +0.040 for younger students — a near-null difference, indicating that scholarship absence and older age at enrollment act as additive, independent risk factors rather than interacting multiplicatively. Both risks are real and quantitatively substantial, but their combination adds rather than compounds.

Below the top three, the ranking highlights that specific degree programs and admission pathways carry identifiable signal. Three course dummies (9500, 9119, 9238) and one admission-pathway dummy (Application mode 39) enter the top 10, indicating that dropout probability varies systematically across specific programs and admission routes. Admission grade (rank 6) and previous qualification grade (rank 7) appear in the lower half

with moderate protective effects, confirming that prior academic preparation contributes to prediction but with less weight than financial support, age, and program-specific characteristics. GDP appears at rank 9, with higher values associated with slightly reduced risk.

These associations should be interpreted as statistical patterns rather than causal effects; observed relationships may reflect underlying socioeconomic structures rather than direct causal mechanisms.

6 Discussion

The results support both hypotheses. The Wilcoxon tests confirm that pre-enrollment data carries meaningful predictive signal (H1, $p = 0.031$ at the test's minimum resolution) and reveal no detectable performance difference between LR and RF (H2, $p = 1.00$; mean F1 gap +0.0002). Together, these findings have a direct practical consequence: dropout risk can be estimated at admission, and among the models compared the simplest performs as well as the most complex.

This matters because under the EU AI Act, educational decision-support systems are classified as high-risk and subject to transparency requirements [7, 6]. While LR's post-encoding feature space (215 columns) is larger than the raw 22-variable input, its coefficients remain individually inspectable and can be aggregated back to original variables for policy-facing summaries — a form of hierarchical interpretability not available for tree ensembles without post-hoc approximation. Under these conditions, choosing a simpler model is not a compromise on performance but an alignment with legal and institutional requirements.

The SHAP analysis offers specific guidance for intervention design. The dominance of scholarship status suggests that financial support mechanisms play a central role in retention [5]; indeed, causal evidence from Italian administrative data indicates that need-based grants substantially reduce first-year dropout, with roughly one third of recipients retained who would otherwise have withdrawn [9]. The additive rather than interactive relationship between scholarship absence and older age indicates that need-based aid expansion should target both groups independently, rather than only at their intersection. The relatively weak contribution of prior academic grades is itself informative: interventions focused exclusively on academic readiness may miss the socioeconomic factors that drive early attrition. Given the EU AI Act framing, operational deployment would additionally require fairness auditing — the presence of gender and age among top predictors must be accompanied by disparate impact analysis and bias mitigation safeguards, so that risk scores lead to supportive outreach rather than exclusionary decisions.

6.1 Limitations

The dataset originates from a single Portuguese polytechnic institution, constraining generalizability. The sample size ($n = 3{,}630$) does not support fine-grained subgroup analysis. Models incorporating first-semester data would likely achieve higher accuracy; the AUC-ROC of 0.795 (RF, test set) reflects the trade-off inherent in the pre-enrollment framing. The question is not whether post-enrollment data improves prediction — it does — but whether the pre-enrollment signal is sufficient to justify early intervention, which the results suggest it is.

7 Conclusion

This paper demonstrates that university dropout can be predicted with meaningful accuracy using exclusively pre-enrollment data, where every feature has been individually validated for temporal availability at admission. Both hypotheses are supported: all three models predict dropout above chance on every fold (H1, $p = 0.031$, the test's minimum resolution), and LR and RF show no detectable performance difference (H2, $p = 1.00$, mean F1 gap +0.0002) — shifting model selection from accuracy optimization to transparency and regulatory alignment.

SHAP analysis identifies scholarship status, age at enrollment, and gender as the dominant predictors, with scholarship absence and older age acting as additive, independent risk factors. Future work will address generalizability by replicating this framework on institutional data from Italian and Slovenian universities, and by extending the analysis to larger multi-program datasets within the European Higher Education Area, with formal fairness auditing of the demographic predictors before any deployment.

More broadly, the results suggest that the conventional reliance on post-enrollment performance data may not be necessary for initiating early support. If pre-enrollment signals are sufficient to identify at-risk students at admission, institutions can intervene earlier, at lower cost, and with greater alignment to the transparency standards that European regulation increasingly demands.

Code availability: https://colab.research.google.com/drive/1pXGd3dH6sSLJwxBUYj2zo7l-OdAK5S8H?usp=sharing.

References

[1] Carmen Aina. 2013. Parental background and university dropout in Italy. *Higher Education*, 65, 4, 437–456. DOI: 10.1007/s10734-012-9554-z.

[2] Carmen Aina, Eliana Baici, Giorgia Casalone, and Francesco Pastore. 2022. The determinants of university dropout: a review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79, 101102. DOI: 10.1016/j.seps.2021.101102.

[3] Mauricio Alban and David Mauricio. 2019. Predicting university dropout through data mining: a systematic literature. *Indian Journal of Science and Technology*, 12, 4, 1–12. DOI: 10.17485/ijst/2019/v12i4/139729.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[5] Alberto F. Cabrera, Amaury Nora, and Maria B. Castañeda. 1992. The role of finances in the persistence process: a structural model. *Research in Higher Education*, 33, 5, 571–593. DOI: 10.1007/BF00973759.

[6] Martin Ebers. 2025. Truly risk-based regulation of artificial intelligence: how to implement the EU's AI Act. *European Journal of Risk Regulation*, 16, 2, 684–703. DOI: 10.1017/err.2024.78.

[7] European Commission. 2024. Regulation (eu) 2024/1689 on artificial intelligence (ai act). Official Journal of the European Union, L 2024/1689. (2024).

[8] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. Vol. 30.

[9] Francesca Modena, Enrico Rettore, and Giulia Martina Tanzi. 2020. The effect of grants on university dropout rates: evidence from the Italian case. *Journal of Human Capital*, 14, 3, 343–370. DOI: 10.1086/710220.

[10] M. L. Nielsen et al. 2025. Trading off performance and human oversight in algorithmic policy: evidence from Danish college admissions. arXiv preprint arXiv:2411.15348. (2025). DOI: 10.48550/arXiv.2411.15348.

[11] Jocey Quinn. 2013. Drop-out and Completion in Higher Education in Europe. Tech. rep. European Commission.

[12] Valentim Realinho, Mónica Vieira Martins, Jorge Machado, and Luís Baptista. 2022. Predicting student dropout and academic success. *Data*, 7, 11, 146. DOI: 10.3390/data7110146.

[13] Cristóbal Romero and Sebastián Ventura. 2020. Educational data mining and learning analytics: an updated survey. *WIREs Data Mining and Knowledge Discovery*, 10, e1355. DOI: 10.1002/widm.1355.

[14] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 5, 206–215. DOI: 10.1038/s42256-019-0048-x.

[15] Vincent Tinto. 1993. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. (2nd ed.). University of Chicago Press.