

Target Leakage in Machine Learning Models that Predict Higher Education Graduation from Administrative Records

Samer Yaghi

syaghi@ucas.edu.ps

University College of Applied Sciences

Rebhi S. Baraka

Islamic University of Gaza

Research Article

Keywords: educational data mining, learning analytics, graduation prediction, target leakage, algorithmic fairness, Palestine

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11412985/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Higher education institutions in developing countries hold large volumes of student records that are rarely used to support decision-making or to identify at-risk students early. This study examines what such administrative records can and cannot reveal about graduation among IT students at the University College of Applied Sciences, Gaza, Palestine, using 4,704 records spanning 2000 to 2019. We first show that two attributes routinely present in student information systems, graduation year and graduate grade average, are deterministic functions of the outcome: retaining them drives Decision Tree, Random Forest and Gradient Boosting to exactly 100% accuracy with zero cross-validation variance, which is the mechanism behind previously reported near-perfect results on these data. Under a corrected, leakage-free protocol in which the train/test split precedes scaling and augmentation, Gradient Boosting reaches 89.16% accuracy (AUC = 0.955; 95% CI [87.4, 90.7]), Random Forest 88.81% and Decision Tree 87.18%; Random Forest and Gradient Boosting are statistically indistinguishable (McNemar, $p = 0.620$). The dominant predictor is the number of regular semesters completed ($r = +0.546$) rather than cumulative GPA ($r = +0.450$). Restricted to admission-time attributes, graduation remains predictable at 70.47% accuracy (AUC = 0.769). Recorded sex is excluded from the attribute set and used only to audit the fitted models, yielding a demographic parity difference of -0.11 and equal-opportunity differences between -0.041 and -0.068 that track a higher observed graduation base rate among female students. The study contributes a reproducible account of how outcome-determined attributes inflate reported accuracy in educational prediction.

1. Introduction

The rapid digitalization of higher education has produced unprecedented volumes of institutional data. This includes enrolment records, academic transcripts, and semester-level grade histories spanning decades. Mining this data to generate actionable intelligence has become a central priority for academic administrators and policy-makers worldwide [2, 3]. Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as established research subfields that bridge computer science and education, offering evidence-based tools to improve graduation rates and reduce dropout rates [4, 5].

Student performance and graduation prediction are among the most studied problems in EDM. The ability to identify students at risk of dropping out early enables timely intervention. In contrast, accurately predicting who will graduate provides insight into what needs to be redesigned in the curriculum and resource allocation [6, 7]. Best-practice guidance for this task is well established [8]. Yet reviews of the field repeatedly observe that methodological rigour has not kept pace with the volume of published prediction studies, and that reported accuracies are frequently difficult to interpret or reproduce [9, 10].

This article makes that concern concrete. Working with a single institutional dataset, we show that two attributes present in almost every student information system — the recorded graduation year and the graduate grade average — are populated only for students who graduated, and therefore encode the outcome itself. Retaining them produces exactly 100% accuracy for every classifier tested and a cross-

validation standard deviation of zero. Removing them yields accuracies in the high-eighties, consistent with published figures for comparable tasks [11, 12]. The contribution is therefore not a higher accuracy but a defensible one, together with two analyses that a corrected model makes possible: how early graduation can be predicted, and whether the resulting predictions behave equitably across student groups.

Two gaps frame the study. First, contributions from the Middle East remain under-represented in the higher-education technology literature [13], and evidence from conflict-affected institutions with constrained data infrastructure is scarcer still [14]. Second, prediction studies rarely report what happens when a modelling assumption fails, even though such reports are precisely what allows other institutions to avoid the same error.

The University College of Applied Sciences (UCAS) in Gaza, Palestine, addresses both gaps [1]. With twenty years of student records (2000 to 2019) across four IT specializations, UCAS provides a rare longitudinal dataset in the Arab-world higher education context. Of the 4,704 students analysed, 38.82% graduated, 21.83% remained enrolled at the time of extraction, 21.73% discontinued their studies, 15.69% were academically dismissed and 1.89% were suspended. That non-completion divides almost evenly between academic and non-academic pathways is itself informative: no purely academic model can address the whole of it.

This paper makes six contributions. First, it provides a reproducible demonstration that outcome-determined attributes in institutional records produce perfect, and entirely spurious, predictive accuracy, together with the diagnostic signature — 100% accuracy across all classifiers and zero cross-validation variance — that allows other institutions to detect the same condition. Second, it establishes corrected performance baselines for graduation prediction on a twenty-year administrative dataset: 87 to 89% with full academic history, and 71% from pre-enrolment attributes alone. Third, it reports a full inferential apparatus honestly, comprising five-fold stratified cross-validation, Wilson score intervals and McNemar's test, including the finding that two of the three pairwise model differences are not statistically significant. Fourth, it specifies a transparent and leakage-free augmentation protocol in which the train/test split precedes augmentation, and reports that Gaussian-noise augmentation degrades rather than improves performance on this dataset. Fifth, it adds a fairness audit by recorded sex, which is held out of the attribute set, reporting demographic parity, equal opportunity and subgroup error rates, which replaces the demographically framed decision rule of the earlier analysis. Sixth, it draws out policy implications and deployment guidelines for UCAS academic advisors, and releases an open repository so that the work can be reproduced [15].

This paper is organized as follows: Section 2 reviews previous research. Section 3 explains the methods and data. Section 4 covers the design of the machine learning model. Section 5 presents the experimental results. Section 6 discusses the findings, policy implications, and limitations. Section 7 concludes with future directions.

2. Related Work

2.1 Machine Learning for Student Performance Prediction

A substantial body of literature applies supervised classification to academic outcome prediction. Asif et al. [6] applied Decision Trees and Neural Networks to undergraduate records and identified first-year GPA and pre-university scores as the dominant predictors. Waheed et al. [7] applied deep learning to Virtual Learning Environment (VLE) clickstream data from UK and Saudi Arabian universities. Huang et al. [16] evaluated classification methods and learning logs for predicting students' academic performance. Chango et al. [17] review approaches to fusing multimodal data sources, among them learning-management logs, social-media interaction and demographic records, in learning analytics and educational data mining. Lopez-Bernal et al. [18] set out k-nearest neighbours, linear discriminant analysis and the perceptron as accessible entry points to binary classification in an Education 4.0 setting.

Three studies provide the most directly comparable reference points. Alyahyan and Düşteğör [8] synthesise the field into a step-by-step best-practice guide covering outcome definition, attribute selection and algorithm choice; their guidance frames the present study, and our leakage finding is best read as an empirical account of what follows when one of their conditions is violated. Akçapınar et al. [11] compared seven classifiers on the interaction data of 76 students and reached 86% accuracy, with k-nearest neighbours identifying 89% of unsuccessful students. Abu Zohair [12] modelled a deliberately small institutional dataset and reported 76.3% and 69.7% prediction rates for dissertation and course outcomes respectively. Rahmani et al. [19] reviewed 110 dropout studies published between 2013 and 2023 and grouped risk factors into demographic, course-related, technological, motivational and support-related categories. These figures place the corrected results reported in Section 5 within an established range rather than outside it.

2.2 Explainable AI and Interpretability in EDM

The shift toward model explainability has become a prominent trend in educational and AI research fields. Tiukhova et al. [20] demonstrated that explainability can also serve as a tool to evaluate the stability of prediction models for student success, while Ujkani et al. [21] showed that explainable AI techniques support the initial identification of at-risk students. Mai et al. [22] proposed a belief-rule-based explainable framework.

Explanation alone, however, does not make a prediction useful. Susnjak et al. [23] found that most learning-analytics dashboards stop at descriptive reporting and argued for integrating predictive and prescriptive layers so that a flagged student maps onto a defined action. Kleimola and Leppisaari [24], interviewing nineteen teaching professionals across four Finnish institutions, found strong recognition of analytics potential alongside very limited pedagogical realisation. Guzmán-Valenzuela et al. [9], reviewing 385 learning-analytics studies, characterise the field as producing a preponderance of analytics but very

little learning. This critique directly shapes Section 6.4, where we treat advising actionability rather than headline accuracy as the operative criterion.

2.3 Big Data Frameworks and Distributed Computing in EDM

As educational datasets grow in scale, distributed computing has become essential [3]. Agnihotri et al. [25] demonstrated Python-Spark integration for EDM tasks. Mai et al. [26] deployed a Spark-based undergraduate prediction platform in Vietnam. Fischer et al. [2] reviewed the affordances of big data in EDM, noting the complementarity between micro-level behavioral data and macro-level institutional records. We note at the outset that a dataset of 4,704 records does not require distributed execution; Section 6.5 addresses the question of institutional scalability on its own terms rather than as a performance claim.

2.4 Methodological Rigour, Leakage and Reproducibility

Target leakage arises when an attribute available to the model encodes information that would not be available at the moment of prediction — in the limiting case, when it is a deterministic function of the outcome itself. Kapoor and Narayanan [27] document leakage across seventeen scientific fields and identify it as a principal driver of irreproducible machine-learning findings, noting that near-perfect reported accuracy is one of its most reliable symptoms. In higher education specifically, Bond et al. [10] synthesised 66 secondary studies of artificial intelligence in higher education and called explicitly for increased methodological rigour in the profiling and prediction literature that dominates the field. Reported accuracies above 99% for graduation prediction should therefore be treated as a diagnostic signal rather than a result, and Section 5.2 provides a worked example of why.

2.5 Recent Advances in AI-Driven Student Success Prediction

Recent studies have increasingly combined machine learning, deep learning, and learning analytics to improve the prediction of academic outcomes. Alalawi et al. [28] proposed an extended learning analytics framework that unites predictive techniques with pedagogical interventions, highlighting the importance of linking predictions to actionable educational support. Kalita et al. [29] demonstrated that Bi-LSTM models enhanced with SHAP-based explanations can achieve both predictive strength and understandability. Farhood et al. [30] survey AI-based personalised learning and report the same pattern across the field: predictive components are plentiful, while evidence that they change instructional practice is sparse. Flanagan et al. [31] modelled early-warning prediction from e-book reading behaviour and reported an AUC of 0.81 for engagement approximately one week into a course, establishing that useful early signals precede formal assessment data — a finding our pre-enrolment analysis in Section 5.5 extends backwards to the point of admission.

2.6 Fairness and Responsible Use of Educational Prediction

Where predictive models inform advising, their error distribution across student groups becomes an operational concern rather than an abstract one. Baker and Hawn [32] review the mechanisms by which algorithmic bias enters educational systems and distinguish measurable disparity from its underlying causes. Raffaghelli et al. [33] argue that datafication in higher education requires institutions to develop critical data literacies capable of interrogating, rather than merely consuming, algorithmic output. Jones [34] proposes consent mechanisms that preserve student privacy and autonomy when institutional records are repurposed for analytics. A further concern is that fairness problems may enter before any model is fitted. Working in the same institutional setting as the present study, Yaghi and AbuSamra [35] show that interruptions to internet connectivity during online examinations depress recorded scores in a way that is detectable from standard examination logs and separable from academic performance, with disruption far more frequent in practical than in theory assessments. Because the cumulative averages our models consume are produced by such assessment processes, measurement error of this kind propagates into any model trained on them. Section 5.9 responds to this literature with a quantitative subgroup audit, and Section 6.4 places the resulting figures under institutional oversight rather than automated action.

2.7 The MENA Context and Research Gap

Studies from the Middle East and North Africa (MENA) are limited. Waheed et al. [7] included a cohort of students from a Saudi university, and Aldowah et al. [36] reviewed the use of EDM in UAE colleges. Al-Sheeb et al. [37] modelled engineering students' academic achievement at a Qatari university from cognitive and non-cognitive factors. Abu Zohair [12] provides the closest regional precedent, modelling a small dataset at an institution in Dubai. A content and authorship analysis of the educational-technology literature confirms that contributions from the Middle East, South America and Africa remain under-represented [13], and Stewart et al. [14] document the structural barriers that constrain educators' data capacity across institutional contexts. We are not aware of a comparable longitudinal analysis of institutional records from a Palestinian higher-education institution.

Despite the significant progress documented in the literature on educational data mining, learning analytics, and explainable AI, few studies have combined long-term institutional datasets with explainable predictive analytics within resource-constrained higher education environments (Saqr & López-Pernas, 2024; Alalawi et al., 2025; Tiukhova et al., 2024). Furthermore, most existing studies focus on predicting student success without addressing the deployment challenges specific to developing-country contexts. Nor do they report the methodological failure modes encountered along the way. The gap this article addresses is accordingly threefold: methodological, contextual, and concerned with the timing and equity of prediction rather than its peak accuracy.

3. Dataset and Methodology

3.1 Data Collection

The data were obtained from UCAS' student information system, with permission from the institution. The full extract comprises 65,532 enrolment records across all college programmes, and represent all students who majored in IT from 2000 to 2019 across four specializations: programming and databases, multimedia, geographic information systems (GIS), and web design and development. Data were obtained at three levels: (i) pre-enrolment data (high school records), (ii) in-program data (semester-level performance, up to 12 semesters), and (iii) outcome data (graduation status). Cohort construction proceeded in three documented steps. Restricting the extract to the four IT specializations left 5,744 records; removing students with zero regular and zero summer semesters, that is, applicants who never commenced study, left 4,823; and removing records with a missing cumulative average left the analytical cohort of 4,704. A stratified 70/30 split then produced 3,292 training and 1,412 test records. Student identifiers were replaced with anonymous IDs to comply with privacy regulations.

Eight attributes were retained for modelling: high-school branch (literary, scientific or other), high-school final percentage, IT major, university cumulative GPA (STD_AVG), regular semesters completed, summer semesters completed, age at enrolment, and region of residence. The binary target STATUS records whether the student graduated.

Recorded sex is deliberately not among them. The student information system stores a binary administrative sex category, and we hold it out of the attribute set so that no prediction is conditioned on it, retaining it only to audit how the fitted models behave across the two groups (Section 5.9). Excluding it costs nothing: Random Forest accuracy rises from 88.31% to 88.81% and Gradient Boosting falls from 89.38% to 89.16%, differences well inside the confidence intervals reported in Section 5.3.

Two attributes present in the extract were excluded before any model was fitted, and the reason is central to what follows. GRADUATE_YEAR records the year of graduation and is null for every student who did not graduate; STD_GRADEAVG records the graduate grade average and is equal to STD_AVG for every non-graduate. Neither is available at the moment a prediction would be useful, and each is a deterministic function of the outcome. Section 5.2 reports what happens when they are retained.

3.2 Exploratory Analysis

A preliminary examination of the dataset indicates that 1,826 students (38.82%) graduated, 1,027 (21.83%) remained enrolled at the time of extraction, 1,022 (21.73%) discontinued their studies, 738 (15.69%) were academically dismissed, 89 (1.89%) were suspended and 2 (0.04%) fall into other categories. By recorded sex, 73.53% of students in these programmes are male and 26.47% female. Graduates from the scientific high school stream averaged a GPA of 77.08%, compared to 74.01% for graduates from the literary stream. Graduation rates differ markedly by high-school stream: 53.88% for the scientific stream against 32.77% for the literary stream and 25.86% for other streams. Differences across the four IT majors are comparatively modest, ranging from 34.08% to 40.93%.

3.3 Data Preprocessing and Augmentation Protocol

The preprocessing for this analysis comprised five steps: (1) removal of students with zero records across all semesters; (2) numeric encoding of categorical variables (high school branch, major, recorded sex, and status); (3) stratified random sampling of the dataset into training (70%) and testing (30%) sets, performed before any transformation; (4) min-max normalization fitted on the training partition only and applied unchanged to the test partition; and (5) augmentation of the training set with Gaussian noise from a normal distribution (mean = 0, standard deviation = 0.01), resulting in a cumulative fourfold increase and a total of 13,168 training records. No augmentation of, or information from, the original testing set (n = 1,412) was used to influence the training data, thereby avoiding the possibility of data leakage.

The ordering above is material. In the implementation released with the earlier version of this work, normalization and augmentation preceded the split, so that near-duplicate copies of training records appeared in the test partition. The protocol reported here reverses that ordering, and the released repository has been corrected accordingly. Section 5.4 additionally reports results with and without augmentation, since the procedure turns out not to help.

3.4 Machine Learning Algorithms and Hyperparameters

Three classifiers were selected from a preliminary screening of nine algorithms (KNN, SVM, Neural Network, Naïve Bayes, Logistic Regression, AdaBoost, DT, RF, GBT) based on hold-out performance. Table 1 documents all hyperparameters used with the selected classifiers. Hyperparameters were held fixed at the values reported in the earlier analysis so that the change in performance reported in Section 5.3 is attributable to the attribute set rather than to tuning.

Table 1
Algorithm Hyperparameters

Hyperparameter	Decision Tree	Random Forest	GBT Classifier
n_estimators / n_trees	1	100	100
max_depth	6	None (unlimited)	3
max_features	All features	sqrt(n_features)	All features
min_samples_split	4	4	4
min_samples_leaf	2	2	2
learning_rate	—	—	0.1
subsample	—	1.0 (bootstrap)	0.8
criterion/impurity	Gini	Gini	Friedman MSE
random_state	42	42	42
Implementation	scikit-learn 1.8	scikit-learn 1.8	scikit-learn 1.8

3.5 Evaluation Framework

The model's performance was measured on the test set ($n = 1,412$) using Correct Classifications, AUC, F1 Score, Precision, and Recall, all derived from the confusion matrix. Ninety-five percent confidence intervals for CA were computed using the Wilson score interval. Stability was confirmed via 5-fold stratified cross-validation on the original dataset (4,704 records), without augmentation, with the scaler refitted inside each fold. Statistical significance of pairwise performance differences was assessed using McNemar's test ($\alpha = 0.05$), which is appropriate for comparing classifiers on the same test set [38].

Three attribute configurations were evaluated on identical splits. Configuration A retains STD_GRADEAVG and GRADUATE_YEAR and serves as a leakage diagnostic. Configuration B is the corrected specification and provides the study's main results. Configuration C restricts the model to attributes available at admission and addresses the question of how early graduation can be predicted. Fairness metrics – demographic parity difference, equal-opportunity difference, and subgroup true-positive, false-negative and false-positive rates – were computed on the Configuration B test partition, grouped by recorded sex, which is not itself an input to any model.

4. Machine Learning Model Architecture

Figure 1 shows the UCAS machine learning architecture, outlining the steps from data collection to result evaluation. Two control points are load-bearing. The leakage screen removes attributes that are populated as a consequence of the outcome, and the split precedes both scaling and augmentation so that no information from the test partition can reach the training procedure. The architecture keeps training and test data separate by performing data augmentation only after splitting the datasets. This ensures the test set contains new, unseen records, similar to real-world situations.

Table 2 provides Pearson correlations between each attribute and the graduation outcome, including the two excluded on leakage grounds and recorded sex, which is reported for completeness but not used as an input. The strongest legitimate correlate is the number of regular semesters completed ($r = + 0.546$), followed by summer semesters ($r = + 0.472$) and cumulative GPA ($r = + 0.450$). Persistence, in other words, carries more signal than attainment in this cohort. The two excluded attributes are shown for comparison: GRADUATE_YEAR correlates with the outcome at $r = + 1.000$, which is the arithmetic expression of its being the outcome. High-school final percentage ($r = + 0.023$, $p = 0.115$) and major ($r = + 0.008$, $p = 0.580$) are not significantly associated with graduation, contrary to what the earlier feature ranking implied.

Table 2
Pearson Correlation with Graduation — All Attributes (n = 4,704)

#	Pearson r	Attribute	p	Description
1	+ 0.546	NORMAL_SMTR_COUNT	< 0.001	Regular semesters completed
2	+ 0.472	SIFI_SMTR_COUNT	< 0.001	Summer semesters completed
3	+ 0.450	STD_AVG	< 0.001	University cumulative GPA
4	+ 0.170	HS_BRANCH	< 0.001	High school branch
5	−0.081	REGION	< 0.001	Region of residence
6	+ 0.069	Recorded sex (not an input)	< 0.001	Administrative sex category
7	+ 0.032	AGE	0.026	Age at enrolment
8	+ 0.023	HS_SCORE	0.115 (n.s.)	High school percentage
9	+ 0.008	MAJOR	0.580 (n.s.)	IT specialization
—	+ 1.000	GRADUATE_YEAR (excluded)	< 0.001	Outcome-determined
—	+ 0.484	STD_GRADEAVG (excluded)	< 0.001	Outcome-determined

5. Experimental Results

5.1 Experimental Setup

All experiments were executed on a single commodity workstation using Python 3 and scikit-learn 1.8. The analytical cohort of 4,704 records does not require distributed execution, and no distributed-execution timings are claimed in this article; Section 6.5 discusses institutional scalability separately. The hyperparameters for each model were defined as in Table 1, and all experiments were executed with `random_state = 42` to ensure reproducibility.

5.2 Target Leakage Diagnostic

Configuration A retains the two outcome-determined attributes identified in Section 3.1. Every classifier then classifies the entire test partition without error: 100.00% accuracy, AUC = 1.000, and a five-fold cross-validation standard deviation of exactly 0.00 for all three models. Three structurally different learners — a depth-limited single tree, a bagged ensemble and a boosted ensemble — do not converge on identical perfect performance because the problem is easy; they do so because one attribute resolves it deterministically. GRADUATE_YEAR is non-null for 1,826 of 1,826 graduates and null for 2,878 of 2,878 non-graduates, and STD_GRADEAVG equals STD_AVG for 100% of non-graduates. A decision rule of the form 'graduation year is recorded' reproduces the label exactly.

This is the mechanism behind the 99.8% accuracy reported in the earlier version of this work, and it is a general hazard rather than a local one: both attributes are standard fields in student information systems, and neither is available at the time a prediction would be useful. Figure 2 contrasts the three configurations. The practical diagnostic we would offer other institutions is the conjunction of near-perfect accuracy across heterogeneous model families with negligible cross-validation variance; where both appear, the attribute set should be audited before the result is reported [27].

5.3 Classification Performance with 95% Confidence Intervals

Figure 3 and Table 3 report the corrected results. Gradient Boosting attains the highest accuracy at 89.16% (95% CI [87.4, 90.7]; AUC = 0.955), followed by Random Forest at 88.81% (AUC = 0.954) and the Decision Tree at 87.18% (AUC = 0.933). The three confidence intervals overlap substantially, and the ordering should not be over-interpreted. These values sit within the range published for comparable tasks – 86% for early-warning classification [11] and 69.7–76.3% on a smaller institutional dataset [12] – which is the appropriate benchmark for a corrected result.

The Decision Tree attains the highest recall (0.898) at the cost of the lowest precision (0.797), producing 125 false positives. For advising deployment, where the cost of a missed at-risk student differs from the cost of an unnecessary consultation, this trade-off matters more than the accuracy ranking.

The final two columns of Table 3 report the effect of the augmentation procedure. Gaussian-noise augmentation reduces accuracy for all three classifiers, by 0.71 to 1.91 percentage points. The un-augmented results are therefore treated as primary throughout, and the augmentation protocol is retained in this article as a reported negative control rather than as a methodological contribution. Synthetic replication of a 3,292-record training partition adds no information that the original records do not already contain.

Table 3
ML Model Performance – Hold-out Test Set (n = 1,412) with 95% Confidence Intervals and the effect of augmentation

Algorithm	AUC	CA (%)	95% CI	F1	Precision	Recall	CA with x4 aug.	Δ
GBT Classifier	0.955	89.16	[87.4–90.7]	0.862	0.851	0.874	87.75	–1.41
Random Forest (RF)	0.954	88.81	[87.1–90.4]	0.859	0.840	0.880	86.90	–1.91
Decision Tree (DT)	0.933	87.18	[85.3–88.8]	0.845	0.797	0.898	86.47	–0.71

5.4 5-Fold Cross-Validation with Confidence Intervals

Table 4 shows the results of the 5-fold stratified cross-validation on the original dataset without augmentation. Fold-level accuracies cluster tightly for all three models: Random Forest $88.67 \pm 0.58\%$, Gradient Boosting $88.50 \pm 0.62\%$ and Decision Tree $87.29 \pm 0.54\%$. The hold-out estimates therefore reflect stable model behaviour rather than a favourable split. The standard deviations are an order of magnitude larger than those obtained under Configuration A, which is what a genuinely difficult classification problem looks like.

Table 4
5-Fold Stratified CV Accuracy (%) with 95% CI – Original Dataset (n = 4,704)

Algorithm	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD (95% CI)
Random Forest	89.16	89.05	87.78	88.95	88.40	88.67 ± 0.58 [88.16–89.18]
GBT Classifier	89.16	87.89	87.99	88.31	89.15	88.50 ± 0.62 [87.96–89.04]
Decision Tree	87.78	87.78	86.50	87.35	87.02	87.29 ± 0.54 [86.82–87.76]

McNemar's test on the shared test partition (Table 5) gives a different picture from the one previously reported. The Decision Tree differs significantly from both ensembles ($\chi^2 = 7.290$, $p = 0.007$ against Gradient Boosting; $\chi^2 = 4.889$, $p = 0.027$ against Random Forest), but the two ensembles are statistically indistinguishable from each other ($\chi^2 = 0.246$, $p = 0.620$). The claim that all pairwise differences are significant at $p < 0.001$ does not hold once leakage is removed, and no single model can be declared best.

Table 5
McNemar's test – corrected models on the shared test set (n = 1,412)

Comparison	χ^2	p	Interpretation
Decision Tree vs GBT Classifier	7.290	0.007	Significant
Random Forest vs Decision Tree	4.889	0.027	Significant
Random Forest vs GBT Classifier	0.246	0.620	Not significant

5.5 How Early Can Graduation Be Predicted?

Configuration C restricts the model to attributes available before a student begins study: age at enrolment, high-school stream, high-school percentage, intended major and region of residence. Gradient Boosting reaches 70.47% accuracy (95% CI [68.0, 72.8]; AUC = 0.769), Random Forest 67.92% and the Decision Tree 67.28%; five-fold cross-validation gives $70.39 \pm 0.66\%$, $68.30 \pm 1.30\%$ and $68.13 \pm 0.77\%$ respectively (Table 6). Recall falls to 0.568 for the best model, meaning that roughly two in five eventual graduates are not identified at admission.

The gap between 70% at admission and 89% once semester records accumulate quantifies the informational value of enrolment history, and it does so in a form an institution can act on. Admission-time screening is accurate enough to prioritise advising capacity but not accurate enough to justify

consequential decisions about individual students. This is consistent with early-warning work reporting an AUC of 0.81 approximately one week into a course [31] and 74% accuracy at week three [11]: useful early signals exist, but they are probabilistic.

Table 6
Pre-enrolment prediction – Configuration C (admission-time attributes only)

Algorithm	CA (%)	95% CI	AUC	F1	Recall	5-fold CV
GBT Classifier	70.47	[68.0–72.8]	0.769	0.599	0.568	70.39 ± 0.66
Random Forest	67.92	[65.4–70.3]	0.747	0.559	0.524	68.30 ± 1.30
Decision Tree	67.28	[64.8–69.7]	0.738	0.534	0.484	68.13 ± 0.77

5.6 Confusion Matrix Analysis

Figure 4 presents the confusion matrices for the three corrected models. Of the 548 students in the test partition who graduated, the Random Forest misclassifies 66 as non-graduating – a false-negative rate of 12.0%, not the 0.1% previously reported – and misclassifies 92 of 864 non-graduates as graduating. For advising, the false-negative rate is the operative figure, and 12.0% is high enough that model output must be treated as a prompt for human review rather than as a determination.

5.7 Decision Rules from the Decision Tree

The DT model (depth = 6, 79 nodes, 40 leaves) yields directly interpretable graduation prediction rules. Figure 5 illustrates the key decision paths, with thresholds shown in raw units; tree-based models are scale-invariant, so the displayed tree is structurally identical to the fitted one. The much smaller tree obtained here reflects the removal of the outcome-determined attributes, which previously supported a large number of near-pure splits. Selected rules:

The root split is on regular semesters completed (≤ 3.5), which alone accounts for 62.8% of the tree's total impurity reduction. Of the 1,955 students at or below this threshold, 23 graduated (1.18%); above it, the graduation rate is 65.59%.

Within the higher-persistence branch, the next split is age at enrolment (≤ 23.5). Students above that age graduate at 75.01% ($n = 2,217$) against 26.32% below it ($n = 532$). This split most plausibly separates completed trajectories from students still enrolled at extraction, and should be read as a censoring artefact rather than as evidence that older entrants perform better (see Section 7).

Cumulative GPA enters at the third level, splitting at 62.2%: within the older, higher-persistence branch, students above that average graduate at 82.29% ($n = 1,880$) against 34.42% below it ($n = 337$).

High-school branch and major receive zero importance in the fitted Decision Tree and region receives 0.002. Recorded sex is not among the attributes at all, so no sex-conditioned rule can arise.

The sex-differentiated rule reported in the earlier analysis is not reproduced. It was an artefact of a tree grown on outcome-determined attributes, and it is contradicted by the subgroup analysis in Section 5.9, in which female students have the higher observed graduation rate. Under the corrected specification no such rule is even expressible, because recorded sex is held out of the attribute set.

5.8 Feature Importance Analysis

Figure 6 reports feature importance for the corrected models. The number of regular semesters completed is dominant in all three (Random Forest 0.339, Decision Tree 0.628, Gradient Boosting 0.570), ahead of cumulative GPA (0.246, 0.152, 0.180). This reverses the ranking reported previously, in which GPA appeared dominant — an appearance produced by the leaked graduate-GPA attribute rather than by attainment itself. Age at enrolment ranks third (0.183 in the Random Forest), for the censoring reason set out in Section 5.7. Region, major and high-school branch together account for 5.8% of Random Forest importance and under 0.6% in each of the two tree models. This suggests that a student's demographic background adds little beyond what GPA and semester records already show. As a result, academic advising for at-risk students should focus more on academic achievement and persistence.

5.9 Fairness Audit by Recorded Sex

Because the earlier analysis surfaced a demographically conditioned rule, we report a subgroup audit of the corrected models on the test partition (Table 7). That audit uses recorded sex only to partition the test set; it is not an input to any of these models. The observed graduation base rate is 45.0% among female students and 36.6% among male students. All three models track this difference: the demographic parity difference (male minus female selection rate) is -0.110 for Random Forest, -0.106 for the Decision Tree and -0.113 for Gradient Boosting, closely matching the -0.084 base-rate gap. That a disparity of this size persists even though the attribute is withheld is itself informative: it is carried by the enrolment-history attributes rather than introduced by the model.

Equal-opportunity differences are smaller in magnitude and also favour female students: -0.041 (RF), -0.053 (DT) and -0.068 (GBT). False-negative rates are correspondingly lower for female students (0.096 versus 0.136 for Random Forest), while false-positive rates are somewhat higher (0.176 versus 0.123). The direction is consistent across all three models and is the opposite of what the earlier decision rule implied. Accuracy parity holds closely (85.98% versus 87.22% for Random Forest).

Two qualifications are necessary. First, parity in error rates does not establish that deployment would be equitable; a model that faithfully reproduces an existing differential can entrench it [32]. Second, the female subgroup comprises 371 test observations, so subgroup estimates carry wider uncertainty than the aggregate figures. Third, the category is a binary administrative field recorded by the institution; it is not self-reported, does not represent gender identity, and supports no claim about groups that a binary field cannot represent. We report these metrics as a deployment precondition and a monitoring baseline, not as evidence that the model is fair in any general sense.

Table 7

Subgroup performance by recorded sex — corrected models (test set, n = 1,412); recorded sex is not an input to any model

Model	Group	n	Accuracy (%)	TPR	FNR	FPR	Selection rate	Base rate
Random Forest	Male	1,041	87.22	0.864	0.136	0.123	0.394	0.366
Random Forest	Female	371	85.98	0.904	0.096	0.176	0.504	0.450
Decision Tree	Male	1,041	86.36	0.869	0.131	0.139	0.406	0.366
Decision Tree	Female	371	86.79	0.922	0.078	0.176	0.512	0.450
GBT Classifier	Male	1,041	87.42	0.885	0.115	0.132	0.407	0.366
GBT Classifier	Female	371	88.68	0.952	0.048	0.167	0.520	0.450

5.10 Comparison with Prior Literature

Table 8 positions the corrected results against representative prior work. At 89.16%, the best corrected model sits within the 82–91% band reported across the comparison studies rather than above it. This is the expected outcome: there is no methodological reason why administrative records from one institution should support dramatically better prediction than comparable datasets elsewhere, and the earlier apparent margin was a leakage artefact. What distinguishes this study is not accuracy but the combination of a twenty-year single-institution record, an explicit leakage diagnostic, pre-enrolment prediction, and a subgroup fairness audit in a context under-represented in the literature.

Table 8
Comparison with Representative Prior Studies in EDM and Learning Analytics

Study	Dataset	Algorithm(s)	Best CA	XAI	Fairness	Country/Context
Asif et al. [6]	Undergraduate records	DT, NN	~ 82%	No	No	Pakistan
Huang et al. [16]	Three universities	RF, XGBoost	~ 91%	No	No	Taiwan/Japan
Waheed et al. [7]	VLE logs	Deep Learning	~ 88%	No	No	Saudi Arabia/UK
Akçapınar et al. [11]	LMS interaction data	kNN, CN2, others	86%	Partial	No	Türkiye
Abu Zohair [12]	Small institutional dataset	SVM, LDA	76.3%	No	No	UAE
Tiukhova et al. [20]	Student success datasets	Explainable ML	N/A	Yes	No	Netherlands
Ujkani et al. [21]	Course success records	XAI-based models	~ 90%	Yes	No	Europe
Kalita et al. [29]	Academic performance data	Bi-LSTM + SHAP	> 90%	Yes	No	Multi-institution
Present Study	20-year UCAS dataset	RF/DT/GBT	89.2%	Partial	Yes	Palestine (Gaza)

As Table 8 shows, subgroup fairness reporting is absent from every comparison study, and none reports a negative methodological finding of the kind documented in Section 5.2. The contribution of this article is therefore located in the reliability and responsibility of the analysis rather than in its headline performance.

6. Discussion

6.1 Statistical Validity and Model Performance

The corrected models achieve 87–89% accuracy with five-fold standard deviations below one percentage point, and their confidence intervals overlap. Two of the three pairwise comparisons are not statistically significant, so no model can be declared superior on this dataset. The practical implication is that model selection should be governed by the error profile appropriate to the intended use rather than by aggregate accuracy: the Decision Tree offers the highest recall and the most legible rules, Gradient Boosting the best overall balance, and Random Forest the most stable behaviour across folds.

The false-negative rate deserves emphasis. At 12.0% for the Random Forest, roughly one in seven students who will graduate is flagged as at risk. In an advising workflow this is tolerable, since the consequence is an additional consultation. It would not be tolerable if model output were used to restrict enrolment or allocate financial support without review, and Section 6.4 is framed accordingly.

6.2 Persistence Rather Than Attainment

Once outcome-determined attributes are removed, the ordering of the classifiers reverses: Gradient Boosting (89.16%) now outperforms the Decision Tree (87.18%), and the difference is statistically significant ($p = 0.007$). The earlier explanation — that a single dominant feature rendered boosting unnecessary — described an artefact of the leaked attribute rather than a property of educational data.

The substantive finding to replace it concerns which attribute now dominates. The number of regular semesters completed carries the highest correlation with graduation ($r = +0.546$) and the highest importance in all three models, ahead of cumulative GPA ($r = +0.450$). Persistence, not attainment, is the primary signal. This aligns with qualitative accounts in which sustained engagement and time management, rather than momentary performance, determine completion [39, 40], and with dropout taxonomies in which support and motivational factors weigh alongside academic ones [19]. For advising, it shifts the operative question from whether a student is performing well to whether the student is still progressing.

6.3 Explainability and the XAI Connection

Feature relevance analysis (Fig. 6) and Decision Tree decision rules (Fig. 5) provide two ways to explain the model's predictions. This matches the trend in explainable AI research in education (Tiukhova et al., 2024; Saqr & López-Pernas, 2024), which shows that models used in education should be understandable to advisors and students, not just accurate. Future work should use SHAP value analysis on the Random Forest model to explain individual student predictions.

Recent studies show that explainability alone is not enough. Saqr and López-Pernas [41] found that even with explainable AI, educational decisions can still be misunderstood. The concern is broader still: reviews of the field find that analytics capability routinely outpaces its pedagogical realisation [9, 24], and that dashboards commonly stop at description without specifying an action [23]. An explanation that a student is flagged because of a low semester count is only useful if the institution has a defined response to a stalled enrolment trajectory. Section 6.4 therefore specifies responses rather than thresholds alone. This means that future versions of our framework should combine prediction explanations with human oversight and clear institutional procedures.

6.4 Policy Implications and Deployment Recommendations

The study's findings translate directly into three actionable policy recommendations for UCAS and comparable higher education institutions:

Early warning system: Use the corrected models to flag students for early intervention as soon as each semester ends. Because the false-negative rate is 12.0%, a flag should open a consultation rather than trigger an automated action, and the flagging threshold should be calibrated to available advising capacity rather than fixed a priori. Given that persistence is the dominant predictor, an interruption in semester progression is the single most informative trigger. Academic advisors should provide these students with priority intervention once flagged.

Admission-time triage rather than major direction: at 70.47% accuracy and 0.568 recall, pre-enrolment models are suitable for prioritising which entrants receive proactive advising, and are not suitable for steering individual students toward or away from specific majors. Graduation rates across the four IT specializations differ by less than seven percentage points, and major carries no significant bivariate association with the outcome ($r = +0.008$, $p = 0.580$), so the evidence does not support major-level guidance at all.

Financial intervention targeting: The 21.73% discontinuation rate and the 15.69% dismissal rate together indicate that non-completion is not a single phenomenon, and the discontinuation component cannot be addressed by academic prediction alone. UCAS should compare the model's at-risk flags with financial aid records to proactively offer scholarships or payment plans to students showing both academic risk and financial stress indicators.

No sex-conditioned rule appears in the corrected models, and it cannot appear as a split in any fitted model. Nonetheless, deployment should be governed by the ongoing monitoring described in Section 5.9: subgroup error rates should be recomputed at each retraining cycle, and any widening of the equal-opportunity difference should trigger review before the model continues in service. Institutions adopting such a system should also establish the consent and transparency mechanisms appropriate to repurposing administrative records for analytics [34, 33], and may find it useful to benchmark their readiness against established capability-maturity dimensions for learning-analytics adoption [42].

6.5 Socioeconomic Context

Academic dismissal (15.69%) and discontinuation (21.73%) together account for more non-completion than either does alone, and the discontinuation pathway is the larger of the two. This indicates a systemic problem extending beyond academic preparedness. The incomplete group represents an intervention target that purely academic models cannot fully address [43]. Enrolment in these programmes fluctuates across the observation window in a pattern consistent with the broader economic conditions in Gaza, and any deployment of predictive models in this setting must treat that volatility as a contextual factor rather than as noise. This is a contextual factor that must inform any deployment of predictive models in this setting. Structural constraints on institutional data capacity documented across comparable contexts [14] reinforce the case for models that operate on records institutions already hold.

6.6 Institutional Scalability

A cohort of 4,704 records trains in seconds on commodity hardware, and we make no distributed-computing claim on this basis. The relevant scalability question for a resource-constrained institution is organisational rather than computational: whether the extraction, screening, retraining and monitoring cycle described in Sections 3 and 5.9 can be sustained with existing staff and on-premise infrastructure. On that criterion the pipeline is well within reach, since it requires no cloud resources, no specialised hardware, and no data beyond what the student information system already holds [3, 42].

7. Limitations

The following limitations set the boundaries for what this study's findings can show:

Single-institution dataset: Internal 5-fold CV confirms stability, but external validity across other Palestinian or Arab institutions has not been established empirically.

Concurrent rather than prospective prediction: cumulative GPA and semester counts are recorded as of extraction, so Configuration B describes a model informed by a student's record to date rather than a genuinely prospective forecast at a fixed point in time. Configuration C addresses the extreme case of admission-time prediction, but the intermediate case — semester-by-semester forecasting — requires longitudinal restructuring of the data that the present extract does not support.

Absence of socioeconomic and engagement variables: Scholarship status, commute distance, LMS engagement, and employment status are not included in the feature set. These likely explain a significant portion of the 21.73% discontinuation rate, and their inclusion would improve model coverage.

Static model: The model does not update dynamically as new cohorts enroll. Concept drift may degrade accuracy for post-2019 cohorts without periodic retraining.

Fairness assessment scope: the audit in Section 5.9 covers recorded sex only, on 371 female test observations, and that field is a binary administrative category rather than a self-reported one. Region of residence and high-school stream are also candidate protected or proxy attributes, and disparity metrics computed on a single test partition carry sampling uncertainty that the reported point estimates do not convey.

Temporal features: Our study defined semester counts as scalar values rather than modeling the full semester-by-semester GPA trajectory as a time series. LSTM- or Transformer-based approaches would likely extract more temporal information from this longitudinal dataset than we could with scalar features alone.

Measurement error in the input attainment measure: cumulative GPA is treated here as a fixed attribute, but it is itself the product of assessment processes that can be unequal in infrastructure-constrained settings. Yaghi and AbuSamra [35] demonstrate, at the same institution, that connectivity failures during online examinations suppress recorded scores independently of preparation. Any attainment signal

affected in this way enters our models as if it were a clean measure, and the resulting predictions inherit that error.

Outcome definition and censoring: students still enrolled at extraction (21.83%) are coded as non-graduates, which conflates 'has not graduated yet' with 'will not graduate'. The age split identified in Section 5.7 is the visible trace of this conflation. The coding is conservative for early-warning purposes but depresses apparent precision, and a censoring-aware survival formulation would be preferable.

Relation to prior reporting: results reported in the first author's master's thesis were obtained under the flawed protocol corrected here and should not be cited as independent corroboration of the present findings.

8. Conclusion

This study analysed twenty years of institutional records from UCAS, Gaza, and reports three findings. First, two attributes routinely held in student information systems – graduation year and graduate grade average – are deterministic functions of the graduation outcome; retaining them drives Decision Tree, Random Forest and Gradient Boosting to exactly 100% accuracy with zero cross-validation variance, and this is the mechanism behind previously reported near-perfect results on these data. Second, once the attributes are removed, graduation is predictable at 89.16% (Gradient Boosting; AUC = 0.955), 88.81% (Random Forest) and 87.18% (Decision Tree), with the two ensembles statistically indistinguishable from each other, and with the number of regular semesters completed rather than cumulative GPA as the dominant predictor. Third, using admission-time attributes alone, graduation remains predictable at 70.47% accuracy (AUC = 0.769), quantifying how much of the signal precedes enrolment.

Recorded sex is held out of the attribute set and used only to audit the fitted models; that audit found no disadvantage to female students, since the models track a higher observed female graduation base rate, with equal-opportunity differences between -0.041 and -0.068 . The sex-conditioned decision rule reported in the earlier analysis is not reproduced, and under the corrected specification it is not expressible.

This study proposes the following practical policy recommendations for UCAS: a semester-by-semester early-warning dashboard triggered by interruptions in enrolment progression rather than by attainment alone; admission-time triage of advising capacity rather than major-level guidance, which the evidence does not support; targeted financial support for the discontinuation group that academic models cannot reach; and routine subgroup monitoring at each retraining cycle.

Future research should test the model at other universities in Palestine and the Arab world, incorporate socioeconomic and LMS engagement data, develop an online retraining system to handle changes over time, use SHAP-based explanations for individual predictions, and explore LSTM or Transformer models to leverage the full sequence of semester data. A censoring-aware survival formulation and semester-by-semester prospective evaluation are the most immediate methodological priorities. Further research

should examine the integration of explainable learning analytics dashboards with personalized student support. Alalawi et al. [28] showed that predictive analytics, combined with clear dashboards, can build advisor trust, improve transparency, and make interventions more effective. Using responsible AI frameworks is also important to make sure automated recommendations are fair, transparent, and helpful in education [41]. High prediction accuracy is necessary, but not sufficient for responsible use in higher education. Where it is achieved too easily, it may not be accuracy at all.

Declarations

Acknowledgements: The authors thank the Academic Affairs Office and the Scientific Research Department of the University College of Applied Sciences for granting access to the institutional records analysed in this study, and the Deanship of Scientific Research and Postgraduate Studies at the Islamic University of Gaza for facilitating this research.

Funding: No funding was received for conducting this study.

Dual publication: No part of this work has been published or is under consideration elsewhere. The study builds on the first author's master's thesis, as stated below; the analyses reported here are new.

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article.

Clinical trial number: not applicable.

Consent to Publish declaration: not applicable.

Consent to Participate declaration: not applicable.

Ethics Approval: The dataset was obtained from UCAS with institutional permission. No human subjects research requiring ethics approval was conducted. Student identifiers were replaced with anonymous IDs before analysis.

Prior dissemination: This article builds on the first author's master's thesis [1]. The institutional dataset is the same; all analyses reported here were newly executed for this article under a corrected, leakage-free protocol, and the results supersede those reported in the thesis. No result is reproduced from the thesis without being re-derived.

Data Availability: The dataset analysed in this study comprises administrative records for 4,704 students held by the University College of Applied Sciences, and was analysed with institutional permission granted for scientific research purposes. It is not publicly available. A synthetic 400-record sample is distributed with the code at <https://github.com/sameredu/UCAS-EDML-Graduation-Prediction> (archived at <https://doi.org/10.5281/zenodo.23436468>). Its records correspond to no real student and reproduce the two leakage relationships of Section 5.2 exactly, so the diagnostic can be verified without access to

the records; it does not reproduce the point estimates reported here, which come from the full dataset and are committed to the repository under results/.

Code Availability: Full implementation available at <https://github.com/sameredu/UCAS-EDML-Graduation-Prediction> (release v2.0.0, archived at <https://doi.org/10.5281/zenodo.23436468>). The released code executes the protocol described in Section 3: the train/test split precedes scaling and augmentation, the scaler is fitted on the training partition only, and the cross-validation, Wilson interval, McNemar and fairness routines used in Section 5 are included. Running `run_all.py` against the full extract reproduces every figure reported here; the outputs are also committed to the repository under results/. The sample data distributed with the code is synthetic and reproduces the leakage diagnostic of Section 5.2 exactly.

Authors' Contributions: S.Y. conceived the study, obtained and prepared the institutional dataset, designed and implemented the analytical pipeline, conducted all experiments and statistical analyses, prepared all figures and tables, and wrote the manuscript. R.S.B. supervised the research, advised on the study design and methodology, and critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

References

1. Yaghi S, Baraka RS. The effect of analyzing big educational data on students' performance: the case of University College of Applied Sciences [Master's thesis]. Gaza: Islamic University of Gaza; 2022.
2. Fischer C, Pardos ZA, Baker RS, Williams JJ, Smyth P, Yu R, Slater S, Baker R, Warschauer M. Mining big data in education: affordances and challenges. *Rev Res Educ.* 2020;44(1):130–60. <https://doi.org/10.3102/0091732X20903304>.
3. Ang KLM, Ge FL, Seng KP. Big educational data & analytics: survey, architecture and challenges. *IEEE Access.* 2020;8:116392–414. <https://doi.org/10.1109/ACCESS.2020.2994561>.
4. Rienties B, K hler Simonsen H, Herodotou C. Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: a need for coherence. *Front Educ.* 2020;5:128. <https://doi.org/10.3389/feduc.2020.00128>.
5. Essa A, Ayad H. Student success system: risk analytics and data visualization using ensembles of predictive models. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*. New York: ACM; 2012. pp. 158–61. <https://doi.org/10.1145/2330601.2330641>
6. Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. *Comput Educ.* 2017;113:177–94. <https://doi.org/10.1016/j.compedu.2017.05.007>.
7. Waheed H, Hassan SU, Aljohani NR, Hardman J, Alelyani S, Nawaz R. Predicting academic performance of students from VLE big data using deep learning models. *Comput Hum Behav.* 2020;104:106189. <https://doi.org/10.1016/j.chb.2019.106189>.

8. Alyahyan E, Düştégör D. Predicting academic success in higher education: literature review and best practices. *Int J Educ Technol High Educ*. 2020;17(1):3. <https://doi.org/10.1186/s41239-020-0177-7>.
9. Guzmán-Valenzuela C, Gómez-González C, Rojas-Murphy Tagle A, Lorca-Vyhmeister A. Learning analytics in higher education: a preponderance of analytics but very little learning? *Int J Educ Technol High Educ*. 2021;18(1):23. <https://doi.org/10.1186/s41239-021-00258-x>.
10. Bond M, Khosravi H, De Laat M, Bergdahl N, Negrea V, Oxley E, Pham P, Chong SW, Siemens G. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *Int J Educ Technol High Educ*. 2024;21(1):4. <https://doi.org/10.1186/s41239-023-00436-z>.
11. Akçapınar G, Altun A, Aşkar P. Using learning analytics to develop early-warning system for at-risk students. *Int J Educ Technol High Educ*. 2019;16(1):40. <https://doi.org/10.1186/s41239-019-0172-z>.
12. Abu Zohair LM. Prediction of student's performance by modelling small dataset size. *Int J Educ Technol High Educ*. 2019;16(1):27. <https://doi.org/10.1186/s41239-019-0160-3>.
13. Bond M. The International Journal of Educational Technology in Higher Education: content and authorship analysis 2010–2024. *Int J Educ Technol High Educ*. 2024;21(1):60. <https://doi.org/10.1186/s41239-024-00492-z>.
14. Stewart B, Miklas E, Szczyrek S, Le T. Barriers and beliefs: a comparative case study of how university educators understand the datafication of higher education systems. *Int J Educ Technol High Educ*. 2023;20(1):33. <https://doi.org/10.1186/s41239-023-00402-9>.
15. Yaghi S, Baraka RS. UCAS-EDML: target leakage in graduation prediction from administrative records. Version 2.0.0 [computer software]. Zenodo; 2026. <https://doi.org/10.5281/zenodo.23436468>.
16. Huang AYQ, Lu OHT, Huang JCH, Yin CJ, Yang SJH. Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interact Learn Environ*. 2020;28(2):206–30. <https://doi.org/10.1080/10494820.2019.1636086>.
17. Chango W, Lara JA, Cerezo R, Romero C. WIREs Data Min Knowl Discov. 2022;12(4):e1458. <https://doi.org/10.1002/widm.1458>. A review on data fusion in multimodal learning analytics and educational data mining.
18. Lopez-Bernal D, Balderas D, Ponce P, Molina A. Education 4.0: teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems. *Future Internet*. 2021;13(8):193. <https://doi.org/10.3390/fi13080193>.
19. Rahmani AM, Groot W, Rahmani H. Dropout in online higher education: a systematic literature review. *Int J Educ Technol High Educ*. 2024;21(1):19. <https://doi.org/10.1186/s41239-024-00450-9>.
20. Tiukhova E, Vemuri P, López Flores N, Islind AS, Óskarsdóttir M, Poelmans S, Baesens B, Snoeck M. Explainable learning analytics: assessing the stability of student success prediction models by means of explainable AI. *Decis Support Syst*. 2024;182:114229. <https://doi.org/10.1016/j.dss.2024.114229>.

21. Ujkani B, Minkovska D, Hinov N. Course success prediction and early identification of at-risk students using explainable artificial intelligence. *Electronics*. 2024;13(21):4157. <https://doi.org/10.3390/electronics13214157>.
22. Mai J, Wei F, He W, Huang H, Zhu H. An explainable student performance prediction method based on dual-level progressive classification belief rule base. *Electronics*. 2024;13(22):4358. <https://doi.org/10.3390/electronics13224358>.
23. Susnjak T, Ramaswami GS, Mathrani A. Learning analytics dashboard: a tool for providing actionable insights to learners. *Int J Educ Technol High Educ*. 2022;19(1):12. <https://doi.org/10.1186/s41239-021-00313-7>.
24. Kleimola R, Leppisaari I. Learning analytics to develop future competences in higher education: a case study. *Int J Educ Technol High Educ*. 2022;19(1):17. <https://doi.org/10.1186/s41239-022-00318-w>.
25. Agnihotri L, Mojarad S, Lewkow N, Essa A. Educational data mining with Python and Apache Spark. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)*. New York: ACM; 2016. pp. 507–8. <https://doi.org/10.1145/2883851.2883857>
26. Mai TL, Do PT, Chung MT, Thoai N. An Apache Spark-based platform for predicting the performance of undergraduate students. In: *Systems. (HPCC/SmartCity/DSS)*; 2019. pp. 191–9. <https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00041>
27. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning- based science. *Patterns*. 2023;4(9):100804. <https://doi.org/10.1016/j.patter.2023.100804>.
28. Alalawi K, Athauda R, Chiong R. An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *Int J Artif Intell Educ*. 2025;35(3):1239–87. <https://doi.org/10.1007/s40593-024-00429-7>.
29. Kalita E, Alfarwan AM, El Aouifi H, Kukkar A, Hussain S, Ali T, Gaftandzhieva S. Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. *Front Educ*. 2025;10:1581247. <https://doi.org/10.3389/educ.2025.1581247>.
30. Farhood H, Nyden M, Beheshti A, Muller S. Artificial intelligence-based personalised learning in education: a systematic literature review. *Discov Artif Intell*. 2025;5:331. <https://doi.org/10.1007/s44163-025-00598-x>.
31. Flanagan B, Majumdar R, Ogata H. Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis. *Int J Educ Technol High Educ*. 2022;19(1):41. <https://doi.org/10.1186/s41239-022-00348-4>.
32. Baker RS, Hawn A. Algorithmic bias in education. *Int J Artif Intell Educ*. 2022;32(4):1052–92. <https://doi.org/10.1007/s40593-021-00285-9>.
33. Raffaghelli JE, Manca S, Stewart B, Prinsloo P, Sangrà A. Supporting the development of critical data literacies in higher education: building blocks for fair data cultures in society. *Int J Educ Technol High Educ*. 2020;17(1):58. <https://doi.org/10.1186/s41239-020-00235-w>.

34. Jones KML. Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *Int J Educ Technol High Educ*. 2019;16(1):24. <https://doi.org/10.1186/s41239-019-0155-0>.
35. Yaghi S, AbuSamra AA. Fairness aware learning analytics for identifying internet induced inequality in online examinations. *Discov Educ*. 2026;5(1):1009. <https://doi.org/10.1007/s44217-026-02182-6>.
36. Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: a review and synthesis. *Telemat Inf*. 2019;37:13–49. <https://doi.org/10.1016/j.tele.2019.01.007>.
37. Al-Sheeb BA, Hamouda AM, Abdella GM. Modeling of student academic achievement in engineering education using cognitive and non-cognitive factors. *J Appl Res High Educ*. 2019;11(2):178–98. <https://doi.org/10.1108/JARHE-10-2017-0120>.
38. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10(7):1895–923. <https://doi.org/10.1162/089976698300017197>.
39. Xavier M, Meneses J. Persistence and time challenges in an open online university: a case study of the experiences of first-year learners. *Int J Educ Technol High Educ*. 2022;19(1):31. <https://doi.org/10.1186/s41239-022-00338-6>.
40. Yılmaz AB, Karataş S. Why do open and distance education students drop out? Views from various stakeholders. *Int J Educ Technol High Educ*. 2022;19(1):28. <https://doi.org/10.1186/s41239-022-00333-x>.
41. Saqr M, López-Pernas S. Why explainable AI may not be enough: predictions and mispredictions in decision making in education. *Smart Learn Environ*. 2024;11(1):52. <https://doi.org/10.1186/s40561-024-00343-4>.
42. Šimić D, Begičević Ređep N, Rako S, Kadoić N, Van Petegem W, Rienties B, Cabrera Lanzo N, Eichhorn M, Guàrdia L, Kučina Softić S, Tillmann A. HELA-CMM: capability maturity model for adoption of learning analytics in higher education. *Int J Educ Technol High Educ*. 2025;22(1):25. <https://doi.org/10.1186/s41239-025-00509-1>.
43. Daniel BK. *Br J Educ Technol*. 2019;50(1):101–13. <https://doi.org/10.1111/bjet.12595>. Big data and data science: a critical review of issues for educational research.

Figures

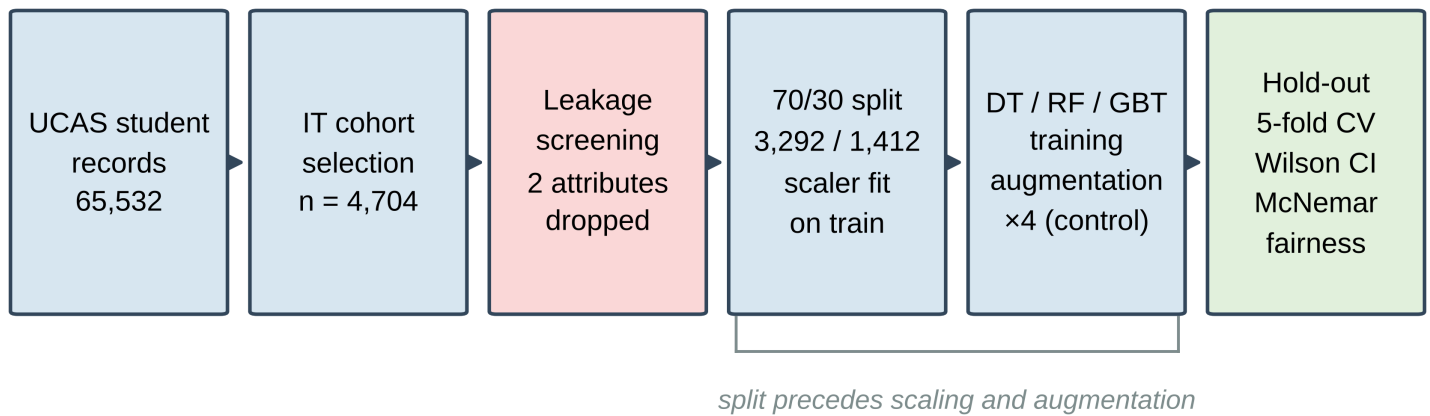


Figure 1

UCAS leakage-free analytical pipeline. The 65,532-record extract is reduced to the 4,704-record IT cohort; the leakage screen, shaded light red, removes GRADUATE_YEAR and STD_GRADEAVG; the stratified 70/30 split yields 3,292 training and 1,412 test records, with the min-max scaler fitted on the training partition only; Gaussian-noise augmentation (×4) is applied to the training set alone and reported as a negative control; evaluation combines a hold-out test, five-fold cross-validation, Wilson score intervals, McNemar's test and a fairness audit by recorded sex. The bracket marks the ordering constraint: the split precedes both scaling and augmentation

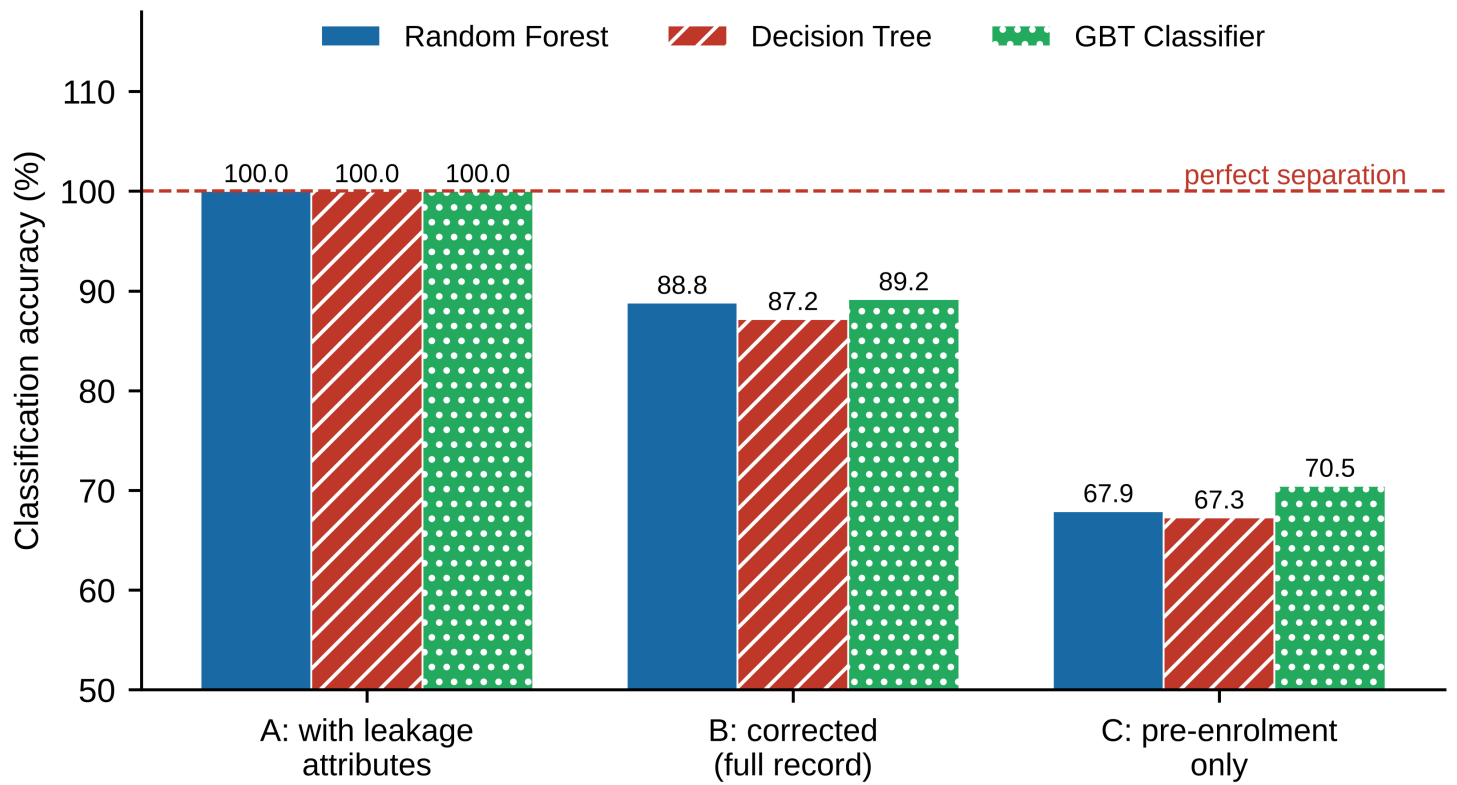


Figure 2

Effect of target leakage on classification accuracy across the three attribute configurations, hold-out test set ($n = 1,412$). The dashed line marks perfect separation. Series are distinguished by both colour and fill pattern

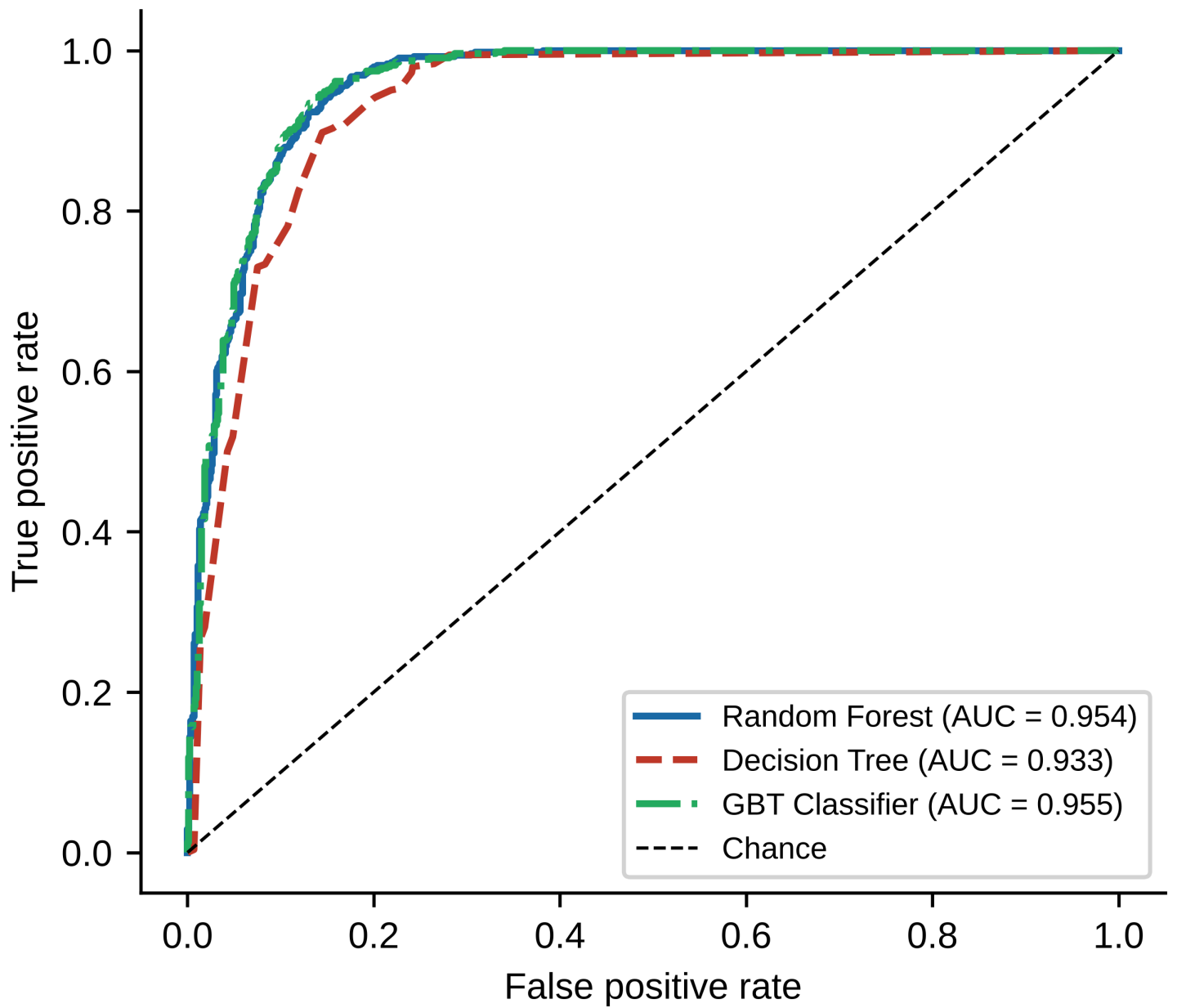


Figure 3

ROC curves — corrected models on the hold-out test set ($n = 1,412$): RF (AUC = 0.954, solid), DT (AUC = 0.933, dashed), GBT (AUC = 0.955, dash-dotted); the diagonal marks chance

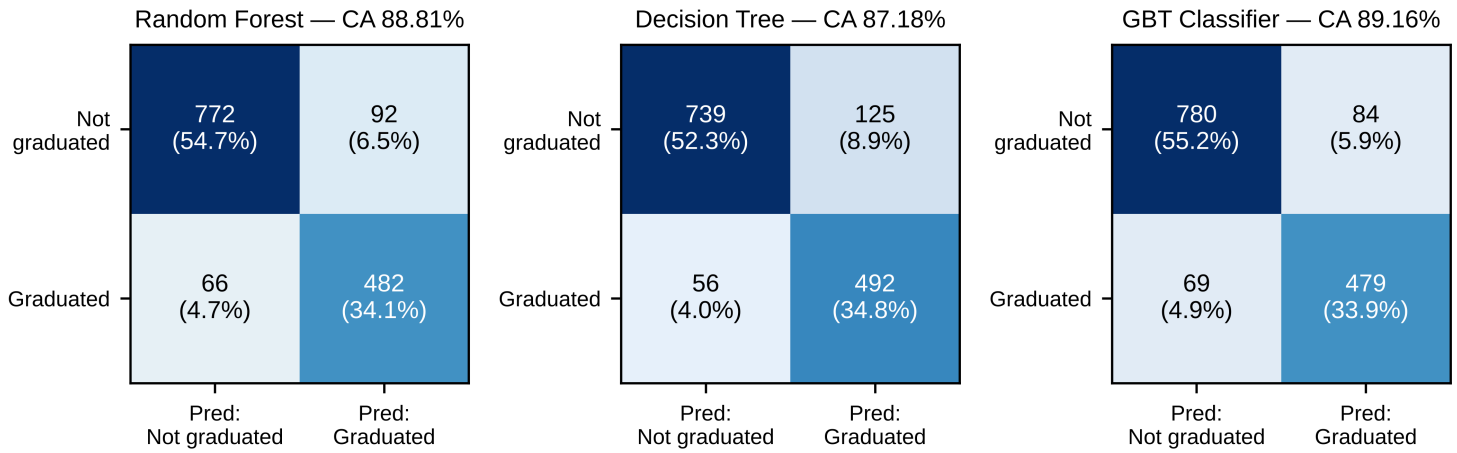


Figure 4

Confusion matrices — all three classifiers side by side for the corrected models on the hold-out test set (n = 1,412). Cells give counts with row-wise percentages of the test set in parentheses

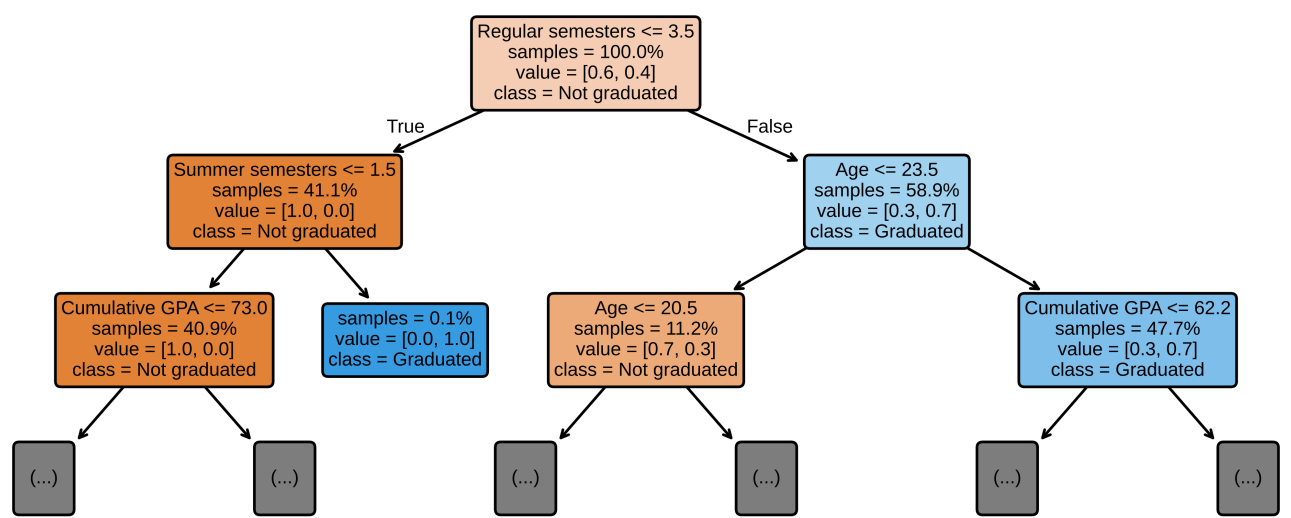


Figure 5

Partial decision tree — graduation prediction rules for the corrected model. The fitted tree has depth 6 with 79 nodes and 40 leaves; the first two levels are shown, with split thresholds in raw units. Orange nodes predict non-graduation, blue nodes graduation; grey nodes mark truncated subtrees

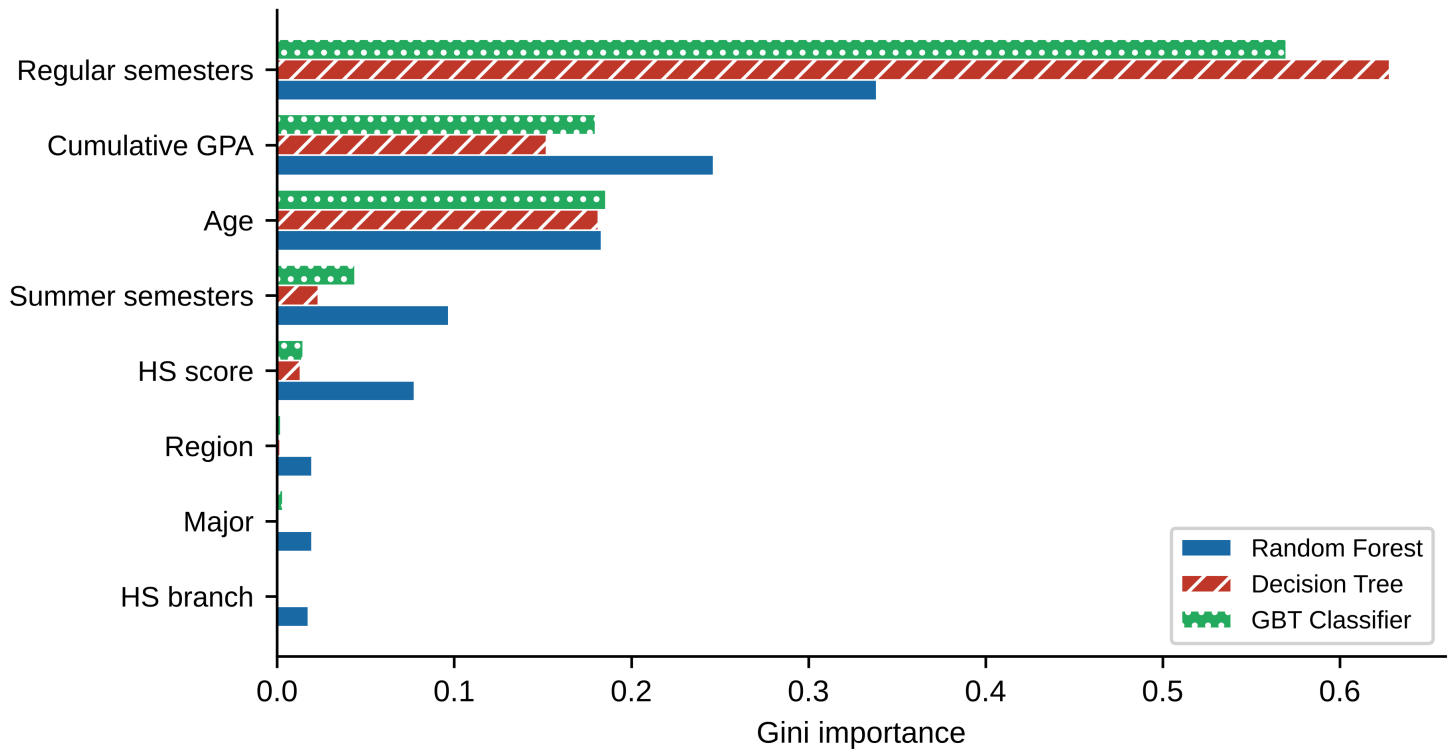


Figure 6

Feature importance comparison – RF vs GBT vs DT for the corrected models, ranked by Random Forest importance. Series are distinguished by both colour and fill pattern