INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN ECONOMICS & MANAGEMENT SCIENCES

OPEN ACCESS JOURNAL

Vol 15, Issue 4, (2026) E-ISSN: 2226-3624

Student Dropout Prediction for Quality Education in Higher Education: A Comparison of Feature Selection and Machine Learning Pipelines

Vimala A/P Doraisamy[1], Arif Ghani[2,3], Nur Nadiah Mahbob[2*]

[1]INTI International College Penang Malaysia, [2]UNITAR International University Malaysia, [3]SEGi University Malaysia

*Corresponding Authors email: nur.nadiah@unitar.my

DOI Link: http://dx.doi.org/10.6007/IJAREMS/v15-i4/29145

Published Online: 26 November 2026

Abstract

Improving quality education in higher education requires identifying students who may leave before completing their studies and offering timely, equitable support. This study compares four feature-selection methods—chi-square, analysis of variance (ANOVA), information gain, and Pearson correlation filtering—with four classifiers: logistic regression, decision tree, extreme gradient boosting (XGBoost), and random forest. The public Predict Students' Dropout and Academic Success dataset contained 4,424 records and 35 predictors. After reserving 794 enrolled cases for risk scoring, 3,630 labelled dropout and graduate cases were split 80:20 for model development and testing. Across 16 pipelines, chi-square–XGBoost had the strongest reported overall holdout profile (accuracy = .9187, precision = .9266, F1 = .9126, recall = .9037, AUC–ROC = .9576). Applied to the enrolled cohort, the pipeline classified 287 students (36.1%) as high risk, 114 (14.4%) as medium risk, and 393 (49.5%) as low risk under predefined thresholds. The comparison indicates that feature selection affects performance and that logistic regression remains a competitive benchmark. These results are limited to one holdout split and one institutional dataset; the risk tiers have neither observed outcome validation nor evidence of improved retention. To support higher education quality, we propose an early-warning process with time-appropriate predictors, probability calibration, subgroup auditing, human review, adequate support capacity, and prospective evaluation. This approach may inform supportive interventions, but the current findings do not demonstrate a causal improvement in educational outcomes.

Keywords: Quality Education, Higher Education Quality, Student Dropout Prediction, Feature Selection, Machine Learning, Early Warning Systems

Introduction

Student attrition is simultaneously an educational, institutional, and social problem. Leaving university without completing a qualification can interrupt individual career trajectories, deepen financial strain, and weaken institutions' ability to deliver equitable learning outcomes. Yet dropout is rarely the result of a single event. It emerges through interacting academic, financial, demographic, and institutional conditions that become visible

at different points in the student journey. This complexity makes prediction attractive but also makes careless interpretation dangerous: a highly accurate model can still be too late, poorly calibrated, unfair across groups, or disconnected from services capable of helping the students it flags.

Educational data mining and learning analytics have therefore moved beyond descriptive reporting toward early-warning systems. Prior studies show that logistic regression, decision trees, random forests, boosting methods, neural networks, and ensembles can identify patterns associated with academic risk. Recent evidence also reinforces three recurring lessons. First, academic progress and engagement variables are often highly predictive. Second, no algorithm dominates across every dataset and prediction window. Third, institutional usefulness depends on much more than accuracy: models require explainability, calibration, external validation, and an intervention protocol that respects student autonomy (López-Zambrano et al., 2021; Marcolino et al., 2025; Slade & Prinsloo, 2013).

Feature selection is central to this problem because educational datasets mix categorical enrolment information, prior attainment, socioeconomic indicators, macroeconomic context, and semester-level academic outcomes. A selector determines which signals are made available to a classifier and, indirectly, what kind of student story the model can learn. Filter methods such as chi-square, ANOVA, information gain, and correlation screening are computationally efficient, but they represent different statistical relationships. Their interaction with linear, tree-based, bagging, and boosting classifiers deserves controlled comparison rather than being treated as a minor preprocessing choice.

This article converts a completed project analysis into a concise, publication-oriented study. It evaluates 16 feature-selection-classifier pipelines on a public higher-education dataset and then examines the operational meaning of probability-based risk tiers for students whose final status was not yet observed. Previous studies provide relevant methodological anchors. Baashar et al. (2022) demonstrated the value of systematically comparing artificial-intelligence models for postgraduate academic performance, while Ng et al. (2024) combined educational big-data architecture, imbalance handling, and predictive modelling. Building on this educational data-mining literature, the present study focuses specifically on dropout, explicitly treats feature selection as a pipeline-level decision, and adds deployment safeguards that are usually absent from algorithm comparisons.

The study addresses three questions: (RQ1) How does predictive performance vary across combinations of four feature-selection methods and four machine-learning algorithms? (RQ2) Which pipeline provides the strongest balance of discrimination and aggregate classification metrics on the reported holdout set? (RQ3) How can its probability outputs be translated into an actionable risk-stratification process without overstating validity or automating consequential decisions?

Literature Review and Conceptual Framing

Dropout prediction as an educational data-mining task

Educational data mining applies computational methods to understand learners and learning environments, whereas learning analytics emphasizes the measurement, analysis, and communication of data for improving learning. Both traditions support dropout

prediction, but they differ from purely commercial classification because the predicted subject is a student and the desired outcome is not merely correct classification. The institution must decide when to predict, which data are legitimate to use, what support follows a flag, and how to avoid reinforcing existing disadvantage.

The public dataset used here was introduced to support prediction of dropout, enrolment, and graduation using 35 variables covering application information, demographic and socioeconomic characteristics, macroeconomic indicators, prior qualification, and first- and second-semester academic performance (Realinho et al., 2021). Its breadth is a strength for method comparison. It also creates a timing challenge: variables observed at the end of the second semester can yield high discrimination but cannot support a genuinely pre-entry alert. Any deployment claim must therefore be tied to a defined prediction horizon.

Reviews consistently report broad methodological diversity and limited comparability across studies. López-Zambrano et al. (2021) found that early performance prediction depends strongly on the timing and composition of available data. Issah et al. (2023) similarly identified variation in influential attributes, algorithms, and evaluation practices. Current applied work combines administrative records with learning-management-system logs, class-balancing methods, and optimized ensemble learners; Marcolino et al. (2025), for example, showed how temporal Moodle activity and imbalance handling can improve identification of at-risk learners. These findings caution against treating a leaderboard result as portable evidence.

Why feature selection and model family should be evaluated jointly

Chi-square selection tests dependence between categorical or discretized features and the target. ANOVA F-tests assess mean separation for numeric predictors. Information gain ranks the reduction in class uncertainty associated with a variable, while Pearson correlation filtering captures linear association. Because these criteria reward different structures, a subset that serves logistic regression may not be optimal for a boosted-tree model. Conversely, tree-based methods can model nonlinearities and interactions, but redundant or noisy predictors can still affect generalization, explanation stability, and maintenance.

The four classifiers span distinct inductive assumptions. Logistic regression provides a regularized linear baseline and probabilistic output. A decision tree yields transparent rules but can be unstable. Random forest reduces variance through bootstrap aggregation and random feature sampling (Breiman, 2001). XGBoost constructs additive trees sequentially and uses regularization and shrinkage to control complexity (Chen & Guestrin, 2016). Comparing all four under every selector provides a more informative test than tuning one favored algorithm in isolation.

Evaluation also requires metric pluralism. Accuracy can conceal poor minority-class performance. Precision concerns the correctness of positive predictions; recall concerns how many positive cases are detected; F1 balances the two. The area under the receiver-operating-characteristic curve summarizes ranking discrimination across thresholds (Fawcett, 2006). None of these alone establishes calibration, fairness, or intervention benefit. A model with a high AUC can still produce probabilities that overstate individual risk or impose unmanageable caseloads.

From risk scores to responsible early warning

A useful early-warning system is a socio-technical cycle: define the prediction time, estimate risk, review explanatory evidence, select a proportionate response, document contact, and evaluate outcomes. Ethical learning analytics literature emphasizes transparency, privacy, consent or legitimate institutional purpose, data minimization, and the possibility that predictions stigmatize students (Slade & Prinsloo, 2013). Human oversight is therefore not a cosmetic addition. Advisors must be able to distinguish a genuine request for help from a misleading pattern and must avoid presenting a score as destiny.

This framing changes the research question from 'Which algorithm wins?' to 'Which validated pipeline can support which decision, at what time, with what safeguards?' It also motivates subgroup audits. A difference in the proportion classified high risk across encoded gender categories may reflect real differences in observed inputs, historical inequities, measurement artifacts, or model error. Without labelled prospective outcomes and fairness metrics, it is a monitoring signal—not evidence that one group is inherently more likely to drop out.

Method

Research design and data

The study uses a quantitative comparative modelling design based on the Predict Students' Dropout and Academic Success dataset in the UCI Machine Learning Repository (Realinho et al., 2021). The source contains 4,424 student records, 35 predictor variables, and three status labels: dropout (n = 1,421), graduate (n = 2,209), and enrolled (n = 794). Because enrolled students had not reached a final outcome, supervised model development used only dropout and graduate records (n = 3,630). The enrolled group was retained for a demonstration of risk scoring after model selection.

The labelled data were split 80:20, producing 2,904 development records and 726 holdout records. The development set contained 1,137 dropout and 1,767 graduate cases; the holdout set contained 284 dropout and 442 graduate cases. The analysis therefore evaluated a binary outcome and did not claim that an enrolled status was equivalent to either final class. The public dataset does not contain direct identifiers, and the present article reports aggregate secondary analysis only.

Table 1
Analytical sample and outcome handling

Partition	Dropout	Graduate	Enrolled	Total
Development	1137	1767	—	2904
Holdout	284	442	—	726
Risk-scoring cohort	—	—	794	794
Full dataset	1421	2209	794	4424

Note. The enrolled cohort was excluded from supervised training because its final outcome was not observed.

Preprocessing and Feature Selection

The project workflow included data inspection, treatment of categorical fields, scaling where appropriate, feature selection, model fitting, and holdout evaluation. Categorical values were numerically encoded in the source workflow. This choice enables conventional estimators but can impose artificial order on nominal categories; future replication should compare one-hot, target-safe, or native categorical encoding. All transformations should be learned on development data and applied unchanged to the holdout set to prevent information leakage.

Four filter-based selectors were compared: chi-square, ANOVA F-test, information gain, and Pearson correlation screening. The project report sometimes abbreviated the last method as 'PCA'; in this article it is named Pearson correlation filtering to prevent confusion with principal component analysis. The selectors were applied to the same candidate predictor pool and crossed with all four model families.

Models, Hyperparameters, and Evaluation

The models were implemented using Python's scikit-learn ecosystem (Pedregosa et al., 2011) and XGBoost. Logistic regression used the default L2-regularized configuration. The decision tree used Gini impurity and maximum depth 5. Random forest used 200 estimators, Gini impurity, and maximum depth 10. XGBoost used 200 estimators, maximum depth 5, learning rate .05, subsample .80, and log-loss evaluation. Each of the 16 pipelines was assessed using the same reported holdout metrics: accuracy, aggregate precision, aggregate recall, aggregate F1, and AUC–ROC. Because the source analysis did not provide repeated cross-validation, confidence intervals, calibration error, or paired significance tests, differences are described as observed holdout differences rather than population-level superiority.

Risk-tier Demonstration and Analytic Safeguards

The highest-performing pipeline generated dropout probabilities for the 794 enrolled records. Predefined thresholds classified probabilities of .75 or higher as high risk, .50–.74 as medium risk, and below .50 as low risk. These cut points were not optimized against intervention costs or prospectively validated outcomes. They are therefore treated as operational scenarios, not clinical-style diagnoses. Counts and percentages were recomputed from the reported category totals; this resolves a denominator inconsistency in the project narrative, where one passage stated 793 although the three risk counts and gender-category totals sum to 794.

Results

Comparative performance of the 16 pipelines

Table 2

Holdout performance by feature-selection and classifier pipeline

Selector	Classifier	Accuracy	Precision	F1	Recall	AUC
Chi-square	Logistic regression	.9119	.9168	.9055	.8980	.9471
Chi-square	Decision tree	.8994	.9186	.8896	.8759	.9290
Chi-square	**XGBoost — BEST**	.9187	.9266	.9126	.9037	.9576
Chi-square	Random forest	.9119	.9195	.9051	.8961	.9507
ANOVA	Logistic regression	.9174	.9219	.9116	.9044	.9471
ANOVA	Decision tree	.8884	.8922	.8801	.8725	.9109
ANOVA	XGBoost	.9132	.9187	.9069	.8992	.9563
ANOVA	Random forest	.9132	.9225	.9063	.8966	.9515
Information gain	Logistic regression	.9146	.9235	.9079	.8984	.9501
Information gain	Decision tree	.8912	.9052	.8812	.8691	.9140
Information gain	XGBoost	.9174	.9228	.9114	.9038	.9559
Information gain	Random forest	.9132	.9225	.9063	.8966	.9528
Pearson correlation	Logistic regression	.8857	.8876	.8776	.8709	.9318
Pearson correlation	Decision tree	.8871	.8928	.8782	.8695	.9180
Pearson correlation	XGBoost	.9077	.9095	.9016	.8959	.9425
Pearson correlation	Random forest	.8981	.9042	.8902	.8817	.9401

Note. *Metrics are reproduced from the project's consolidated results table. Precision, recall, and F1 are reported aggregate values. No uncertainty estimates or significance tests were available.*

Chi-square–XGBoost ranked first on accuracy (.9187), precision (.9266), F1 (.9126), and AUC–ROC (.9576). Its recall (.9037) was also strong, although ANOVA–logistic regression (.9044) and information-gain–XGBoost (.9038) were marginally higher. Thus, the selected pipeline offered the best overall metric profile rather than an absolute maximum on every criterion.

Algorithm rankings depended on the selector. XGBoost produced the highest AUC within every feature-selection family, but its accuracy advantage was small in several comparisons. Under ANOVA, logistic regression achieved .9174 accuracy compared with .9132 for XGBoost; under information gain, both XGBoost and logistic regression were close (.9174 and .9146). These results make a strong case for retaining a transparent linear benchmark. The decision tree was consistently weakest or near weakest, while random forest was stable but did not lead any selector family.

Pearson correlation filtering produced the lowest accuracy for all four classifiers, ranging from .8857 to .9077. The best and worst observed accuracies differed by .0317, and the best and worst AUC values differed by .0468. These are operationally meaningful ranges in a

system that may screen hundreds or thousands of students. They show that preprocessing cannot be treated as a neutral preliminary step.

Risk Stratification of Currently Enrolled Students

Table 3

Probability-based risk tiers in the enrolled cohort

Tier	Probability rule	n	%	Suggested service intensity
High	$p \geq .75$	287	36.1	Advisor review and timely outreach
Medium	$.50 \leq p < .75$	114	14.4	Monitor and offer low-friction support
Low	$p < .50$	393	49.5	Universal support; no adverse inference

Note. *Percentages use n = 794. Suggested responses are governance recommendations, not tested interventions.*

The selected pipeline placed 287 enrolled students in the high-risk tier, 114 in the medium-risk tier, and 393 in the low-risk tier. More than half of the cohort (50.5%) would therefore receive at least a medium flag. This is analytically plausible but operationally demanding. A university adopting these thresholds would need capacity for approximately 401 reviews or contacts, and 287 relatively intensive high-risk reviews. Threshold selection should consequently incorporate available staffing, false-positive costs, and the expected benefit of support—not only statistical performance.

The source analysis also reported risk tiers by encoded gender. Among category 0 (reported as female), 140 of 487 records (28.7%) were high risk; among category 1 (reported as male), 147 of 307 (47.9%) were high risk. This 19.1 percentage-point difference is a descriptive model-output disparity, not an observed dropout-rate difference. Without final outcomes by subgroup, it is impossible to determine whether the model is equally calibrated, has equal error rates, or reflects structural differences in the predictors. The result should trigger validation and review rather than group-targeted adverse action.

Table 4

Reported risk tiers by encoded gender category

Encoded category	High n (%)	Medium n (%)	Low n (%)	Total
0 (reported female)	140 (28.7)	68 (14.0)	279 (57.3)	487
1 (reported male)	147 (47.9)	46 (15.0)	114 (37.1)	307
Total	287 (36.1)	114 (14.4)	393 (49.5)	794

Note. *These are predicted risk distributions. They do not measure observed subgroup dropout or establish bias, causation, or inherent group differences.*

Discussion

Main Findings

The comparative experiment yields three main findings. First, XGBoost paired with chi-square selection achieved the strongest overall holdout profile, with AUC–ROC of .9576 and accuracy of .9187. This is consistent with the capacity of gradient boosting to model nonlinear

interactions while controlling complexity through shrinkage and regularization. Second, model family alone did not determine performance. Changing the feature-selection method altered every classifier's result, and correlation screening was systematically weaker. Third, simpler models remained competitive. ANOVA–logistic regression nearly matched the selected pipeline on accuracy and produced the highest reported recall, underscoring the value of transparency-sensitive model selection.

The findings are consistent with related educational data-mining research while extending it in a specific direction. Baashar et al. (2022) showed that artificial-intelligence models can support academic-performance evaluation and that comparisons across models matter. Ng et al. (2024) demonstrated how large educational datasets, imbalance treatment, and algorithmic comparison can be integrated into predictive modelling. The present results add a factorial pipeline perspective: the selector and classifier should be selected jointly, and the winning pipeline should be assessed within a decision system rather than presented as a standalone technical artifact.

The high AUC indicates strong ranking ability on the particular holdout set, but it does not prove that probabilities are accurate, that performance will transfer to another university, or that intervention will reduce dropout. This distinction is crucial because the public dataset originates from one higher-education context. Institutional policies, grading regimes, tuition arrangements, labor markets, and student populations can alter both base rates and predictor relationships.

Timing, leakage, and the meaning of 'early'

A central interpretive issue is the inclusion of first- and second-semester curricular outcomes. These variables may legitimately support an end-of-year continuation model, but they cannot support an admission-stage or first-week early warning. If the target status is determined after these outcomes, late-semester performance can dominate the model and create an impression of exceptional predictive strength while reducing the time available to help. A publication-quality deployment study should therefore define multiple landmark models: pre-entry, early semester, end of first semester, and end of second semester. Each model should use only data available at its landmark.

Temporal landmarking also clarifies the fairness and usefulness trade-off. Early models may be less accurate but create more time for support. Later models may be more accurate but merely restate academic difficulty already visible to instructors. The preferred model should maximize expected intervention benefit, not retrospective classification alone.

A prediction-to-action governance framework

To convert the current prototype into an institutional early-warning system, six gates are recommended. Each gate can stop deployment if its conditions are not met. This turns governance into an operational control rather than a statement of principles.

Table 5
Required gates from model evidence to student support

Gate	Required evidence	Decision rule
1. Purpose and timing	Defined decision, eligible population, prediction landmark, and permissible data	Do not deploy an undefined "early" model.
2. Validation	Repeated nested cross-validation, untouched temporal test, and external or local validation	Reject a pipeline whose advantage is unstable.
3. Probability quality	Calibration plot, Brier score, threshold sensitivity, and prevalence monitoring	Do not treat uncalibrated scores as individual probabilities.
4. Fairness and privacy	Subgroup sensitivity, specificity, precision, calibration; data minimization and access controls	Investigate disparities before use; prohibit adverse automated action.
5. Service readiness	Named owner, review queue, outreach protocol, opt-out path, and capacity at each tier	Set thresholds to match useful support capacity.
6. Impact evaluation	Prospective stepped-wedge, randomized, or strong quasi-experimental evaluation	Scale only when the combined model-plus-support process improves outcomes.

Note. *A model may be technically accurate yet fail one or more deployment gates.*

At the validation gate, repeated stratified cross-validation should be nested around feature selection and hyperparameter tuning. An untouched temporal cohort should then assess performance under changing enrolment conditions. External validation at another institution is preferable because it tests transportability rather than merely resampling the same population. Reporting should include confidence intervals and paired comparisons, while calibration should be evaluated independently from discrimination.

At the fairness gate, the institution should examine error rates and calibration across groups only after obtaining observed outcomes. A higher flag rate may be justified by different prevalence, may reflect inequitable conditions requiring additional resources, or may signal model distortion. The appropriate response is diagnostic: compare false-negative and false-positive rates, inspect feature contributions, test alternative thresholds, and involve affected stakeholders. Removing a protected attribute alone does not prevent proxy discrimination.

At the service-readiness gate, high-risk status should create a human review, not an automatic sanction. Outreach should be supportive and non-stigmatizing—for example, an invitation to discuss academic planning, finance, wellbeing, accessibility, or study skills. Advisors should see concise reasons and data freshness, record contextual corrections, and be able to override the model. Students should not lose enrolment opportunities, funding, or access based solely on a prediction.

Finally, the impact gate must test the full intervention. A model can increase contact without improving retention, or support may benefit students regardless of predicted tier. A stepped-wedge rollout, randomized encouragement design, or well-controlled quasi-

experiment can estimate whether the prediction-plus-support system improves continuation and whether effects differ across groups.

Practical Implications

For researchers, the study shows why tables should report the full selector–classifier matrix rather than only the winning score. Full reporting reveals that logistic regression is a credible alternative and that Pearson screening underperforms consistently. Reproducible code should preserve the train–test boundary, record selected features in every fold, and publish random seeds and package versions.

For institutional leaders, the risk-tier counts translate threshold choice into workload. Under the current thresholds, 36.1% of enrolled students would enter the most intensive queue. Before implementation, leaders should model advisor capacity, estimated contact rates, and the harms of false alarms. A lower caseload may require a higher threshold; a universal low-cost intervention may justify a broader tier. These are policy choices supported by model evidence, not outputs the algorithm can decide.

For practitioners, explanations should focus on changeable or support-relevant conditions. A demographic characteristic may help audit the system but should not become a reason for deterministic treatment. The most defensible interface would show the score range, prediction date, recent academic indicators, uncertainty, and available support, together with a clear statement that the prediction is not a final judgment.

Limitations and Future Research

Several limitations constrain the conclusions. The evidence comes from one public dataset and one reported 80:20 split. Without repeated resampling, confidence intervals, or paired tests, the small differences among leading pipelines may be sampling noise. No independent external or temporal cohort was used. The enrolled cohort lacked final outcomes, so its tier distribution cannot establish predictive validity, calibration, or intervention efficacy.

The predictor set includes semester-level academic variables, and the article cannot reconstruct a truly early prediction horizon from the consolidated results. Numerical encoding of nominal variables may introduce artificial ordinality. The source analysis did not report class-specific confusion matrices for every pipeline, probability calibration, missing-data sensitivity, feature stability, or detailed subgroup error metrics. The probability thresholds were administratively defined rather than optimized against costs and benefits.

Future research should pre-register prediction landmarks; compare native categorical encoders and leakage-safe preprocessing; use nested cross-validation and temporal testing; assess calibration and decision curves; and validate across institutions. Explainability should be evaluated for stability and usefulness rather than presented only as feature rankings. Most importantly, a prospective study should test whether differentiated support triggered by the model improves retention, student experience, and equity relative to existing practice.

Conclusion

A controlled comparison of 16 feature-selection and machine-learning pipelines found that chi-square–XGBoost provided the strongest overall reported performance for classifying

dropout versus graduation. The result is technically promising: accuracy reached .9187 and AUC–ROC .9576. Equally important, the analysis showed that feature-selection choice affected every classifier and that logistic regression remained close to the leading ensemble under several selectors.

The article's broader conclusion is that predictive quality is necessary but insufficient. Risk tiers for 794 enrolled students demonstrate how scores can organize support, yet they also expose unresolved questions about capacity, calibration, subgroup performance, and timing. A responsible early-warning system must therefore connect a temporally valid model to human review, proportionate support, privacy and fairness controls, and prospective impact evaluation. The proper endpoint is not a more accurate label; it is earlier, more equitable, and demonstrably effective assistance for students.

Acknowledgement
The publication fees for this research article were fully funded by UNITAR International University Malaysia.

References

Baashar, Y., Hamed, Y., Alkawsi, G., Capretz, L. F., Alhussian, H., Alwadain, A., & Al-amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. Alexandria Engineering Journal, 61(12), 9867–9878. https://doi.org/10.1016/j.aej.2022.03.021

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision Analytics Journal, 7, 100204. https://doi.org/10.1016/j.dajour.2023.100204

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0

Ng, Y. L., Ting, T. T., Teoh, C. K., Khattak, U. F., & Almaiah, M. A. (2024). Educational big data analytics: Machine learning based academic performance predictive modelling. Pakistan Journal of Life and Social Sciences, 22(2), 7442-7477. https://doi.org/10.57239/PJLSS-2024-22.2.00562

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. Psicothema, 33(3), 456–465. https://doi.org/10.7334/psicothema2021.62

Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from Moodle log data. Scientific Reports, 15, 9840. https://doi.org/10.1038/s41598-025-93918-1

Melo, E., Silva, I., Costa, D. G., Viegas, C. M. D., & Barros, T. M. (2022). On the use of explainable artificial intelligence to evaluate school dropout. Education Sciences, 12(12), 845. https://doi.org/10.3390/educsci12120845

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9, 11. https://doi.org/10.1186/s40561-022-00192-z

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. Frontiers in Psychology, 12, 698490. https://doi.org/10.3389/fpsyg.2021.698490