ISSN: 2581-3404 (Online) IF: 5.86 (SJIF)
IJIRTM, Volume-10, Issue-6, November-2026

Leakage-Free Student Dropout Prediction in Open University Distance Learning via AUC-Weighted Ensemble Classification

Yogesh Joshi[1], Deepak Negi[2]
Research Scholar[1], Associate Professor[2]
Faculty of Technology and Computer Applications, Amrapali University, India[1,2]
yogesh1070@amrapali.ac.in[1], dsingh.negi23@gmail.com[2]

Abstract. *Online and distance learning environments continue to face substantial student dropout, creating academic, institutional, and student-support challenges. Reliable dropout modelling therefore requires not only strong predictive performance but also careful control of temporal leakage and evaluation bias. This study presents a leakage-controlled framework for student dropout classification using the Open University Learning Analytics Dataset (OULAD). The dataset contains 32,593 enrolments across seven modules and four presentations from 2013–2014. After excluding 3,510 enrolments without recorded VLE activity, 29,083 enrolments were retained, with a dropout rate of 24.55%. A total of 76 temporally valid features were constructed from demographic characteristics, registration information, VLE activity, assessment behaviour, and interaction features, while post-outcome variables were excluded. The modelling framework used a stratified 80/20 hold-out split, with inner 10-fold cross-validation for threshold optimisation and AUC-based ensemble weighting. RUSBoost, Random Forest, LogitBoost, GentleBoost, and Logistic Regression were evaluated. On the held-out test set, the AUC-weighted ensemble achieved an AUC of 0.9296, F1-score of 0.7498, and accuracy of 0.8540. McNemar and DeLong tests indicated statistically significant differences against several individual models, while the difference from LogitBoost was not statistically significant. Feature-importance analysis highlighted assessment engagement and performance-related variables as important predictors. The findings demonstrate the value of leakage-controlled feature construction, nested evaluation, and ensemble modelling for reproducible student dropout analysis.*

Keywords: Student dropout prediction; Learning analytics; Educational data mining; Ensemble learning; Temporal leakage; OULAD; Statistical validation.

Introduction

The emergence of learning via the internet and distance education has considerably changed the field of higher education by providing access to academic programmes for students that are not able to attend classes in person due to geographical, economic, or social limitations (Alnasyan et al., 2024). Despite these learning opportunities being convenient and accessible for many, there are certain issues connected to student retention and learning progress (Oyedotun et al., 2025). The issue of student withdrawal or drop-out is especially relevant in this context as it may be caused by a number of factors, including academic problems, motivation, learning strategies, personal issues, or inability to adjust to new learning methods. For institutions, a high percentage of students who withdraw from academic programmes can influence resource management, educational planning, provision of support, and the result of studies as well (Deeva et al., 2022). Therefore, detection of potential students who have problems with

staying in academic programmes became an interesting topic for researchers in the fields of learning analytics and educational data mining. With the growing digitalization of learning, institutions have accumulated large amounts of student data which provide an opportunity to understand learning processes better. Demographics, registration data, assessment data, and virtual environment interactions might be combined to identify potential patterns related to student retention and attrition (Niyogisubizo et al., 2022). Traditionally, identification of at-risk students was based on their academic performance retrospectively and manually identified indicators. While this method is still applicable, it may not be sufficient to capture multivariate relations present in educational big data (Niyogisubizo et al., 2022). With the use of machine learning, it is possible to develop models that will identify certain patterns in multiple variables and calculate the probability of specific outcomes (Fernandez-Garcia et al., 2021). In this case, learning analytics and machine learning can be used for student dropout prediction to make interventions more evidence-based (Fauszt, 2026).

The Open University Learning Analytics Dataset (OULAD) is one of the key benchmark datasets being used for studying issues such as engagement, academic success, and student retention. The data in the OULAD consists of various features related to the students' demographic background, registration, course delivery, assessments, and interaction with the Virtual Learning Environment (VLE) in different modules/presentations (Vo et al., 2026). The multi-dimensional nature of the dataset makes it quite useful for exploring the association between the individual-level features, academic behavior, and online engagement with educational outcomes (Krüger et al., 2023). As a result, OULAD offers a suitable empirical setting for investigating the performance of machine learning algorithms in drop-out classification and investigating the impact of using various data sources on predictive performance (Krüger et al., 2023). At the same time, there is a number of methodological issues associated with creating a model for predicting the dropout status of students and the role of different data sources in this process that should not be underestimated. One of the most important issues related to the construction of a prediction model is whether the information that will be provided to the model will be really available at the time of prediction (Bettahi et al., 2025). It means that in educational datasets there could be variables containing information created after a student withdraws or becomes aware of the final outcome. If such variables are used for model creation, then the model receives information about the target variable indirectly (M et al., 2026). The phenomenon called label leakage or information leakage can cause significant overfitting and results that are far from being real prediction. Therefore, a model subjected to leakage might show good performance in evaluation but much lower performance on real future or unknown.

Thus, temporal validity becomes an indispensable factor in student dropout forecasting. The features must be designed in accordance with the required phase of prediction so that no information produced afterwards is included in (Rabelo & Zárate, 2025). Also, the process of developing models and evaluating their final performance must be distinguished. Using the same observations for model selection, parameter tuning, choice of threshold, and/or weighting of ensembles can lead to a positively biased estimation of their performance. Thus, the use of an independent test set is needed to assess the performance of the developed models on observations not used during the optimization (Bettahi et al., 2025). This aspect becomes especially crucial when comparing multiple learning algorithms, classification thresholds, and ensemble combinations. The potential of machine-learning techniques in forecasting students associated with withdrawal has been shown previously using the OULAD dataset (Raj & V G, 2022). However, cross-study comparisons may be rather difficult due to the difference in the feature construction procedure, definition of the outcome, splitting of the data, optimization process, and evaluation metrics used by researchers.

Specifically, studies that do not explicitly control for post-outcome information can be misleading with regard to the actual performance from a predictive point of view (Vaarma & Li, 2024). Likewise, when the process of model selection and optimisation takes place without complete isolation of the test data used in the final assessment of the

performance, the reported performance cannot be an accurate representation of the generalisation to unseen observations (Vaarma & Li, 2024).

The current study seeks to overcome this problem by providing a framework for student dropout classification through OULAD in which leakage is controlled. The study creates a full suite of 76 features consisting of demographics, registration, assessment, VLE activities and interaction based metrics. Special consideration is given to excluding those variables that may provide information that was produced after the outcome (Paz, 2025). Instead of relying on a single classification technique, the study examines different machine learning methods and also considers if combination of different models could increase the performance of classification (Patel & Amin, 2024).Logistic regression is chosen as a basic model while RUSBoost, Random Forest, LogitBoost, and GentleBoost are used in a multiple models framework (Meeraa Shahibo & Ahangama, 2025). Another important feature of the methodology used in this study is the distinction between model creation and evaluation process. The data is split using stratified 80/20 hold-out method, keeping the test part as unseen test set (M et al., 2026). All model tuning efforts, including finding the best threshold and ensembling weight calculation, are conducted only using training data through inner 10-fold cross-validation. In particular, the threshold used in classifier is chosen based on F1 score criterion, not taking automatically the usual probability threshold, which allows incorporating this trade-off into modelling. Analogously, the ensembling assigns weights to individual models based on cross-validated ROC-AUC performance in training data. This approach helps to avoid the situation when the information about the final test set may somehow influence the modelling process.

Literature Review

2.1 Leakage Prevention and Temporal Feature Engineering

One major problem with the methodology for analytics for distance learning is target and temporal leakage, which involves information from after the intended predictive period influencing the predictor variables (Meeraa Shahibo & Ahangama, 2025). Data leakage is brought about by including data on post-enrolment administrative occurrences, unregistration dates, or final term results into the data used, thus generating unrealistic performance levels (Alzabidi et al., 2026). It is recommended in the methodology that only after the closing of the observation period should the ground truth be obtained, and that features should be created using only data up to the point of prediction (Davis et al., 2023).

In OULAD and other virtual learning environment contexts, predictive capacity itself changes with each subsequent time horizon. Initial models based on demographic information and enrollment data tend to perform moderately in terms of discrimination capacity, according to Sunil P. Patel et al. (2026). By mid-term, the discrimination performance has improved significantly due to VLE interaction history, access of materials, and formative assessments (Tan et al., 2022). Longitudinal studies demonstrate that a limitation of features to pre-study variables and early behavior on board allows generating useful warning signals in addition to temporal independence (Algarni & Sheldon, 2025). Moreover, dynamic extraction of features in multi-week intervals helps predict disengagement within the first weeks of a presentation regardless of downstream test scores and final grades (Algarni & Sheldon, 2025). It should be noted that temporal validity is not just a technical requirement but an indispensable condition for being operationally ready.

2.2 AUC-Weighted Ensemble Classification Frameworks

In addition to the issue of leakage prevention, there is the matter of designing classification architecture that would enable exploiting the heterogeneity, sparseness, and non-linearity of educational data (Park & Yoo, 2021). Ensembles have always demonstrated higher performance compared to single learners due to combining multiple inductive biases (Singh, 2025). In bagging approaches like Random Forest, variance is lowered by means of bootstrap aggregation, whereas boosting algorithms deal with difficulties iteratively by re-weighting samples. Modern tabular ensembles include Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost and are often used as base learners in high-dimensional logs of VLE (Khoushehgir & Sulaimany, 2023). An emerging trend goes further than simple averaging or voting and employs weighting of the

ensemble members based on the discriminative power of each model, usually AUC (Sitam & Anusha, 2026). Given the class imbalance of distance-education datasets, threshold-independent measures like AUC serve as more stable optimisation objectives than accuracy (Hutt et al., 2022). Weighting techniques which allow for the retraining of base learners using cross-validation performance metrics have proved to be beneficial in improving the recall and global AUC of the model. Meta-learning methods which involve stacking of the probability estimates of the learners using methods such as logistic regression or soft voting have been found to be instrumental in ensuring the detection of true positives without causing too many false alarms (Mubarak et al., 2021). Evolutionary and metaheuristic search methods, such as particle-swarm optimization, have also been used to determine the most appropriate weights of the learner ensembles using synthetic oversampling to produce an AUC of over 0.95 on selected institutional datasets (Guamán-Hidalgo & Enciso, 2026).

2.3 Class Imbalance, Interpretability, and Institutional Deployment

Imbalanced class representation, model explainability and the transfer of model predictions to practice form another cluster of problems (Khoushehgir & Sulaimany, 2023). In the vast majority of distance education cohorts, those who complete their studies vastly outnumber those who drop out, so uncalibrated predictive algorithms based on institutional databases will exhibit majority class bias. The result is an overoptimistic accuracy score but poor ability of such models to identify the smaller class of attritees (Sitam & Anusha, 2026). Data-based approaches like Synthetic Minority Over-sampling Technique (SMOTE), random under-sampling or SMOTE-Tomek methods have become common in order to address this problem (Mubarak et al., 2021). Algorithmic solutions, including cost-sensitive loss functions and optimized decision boundaries, can enhance the recall of attritees as well. In conjunction with these solutions, the rising call for explainable artificial intelligence should be noted (Dass et al., 2021). Probability scores generated by black box algorithms provide little help to the academic advisors, hence post hoc explanation techniques such as SHapley Additive exPlanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been utilized in order to determine the relevant behavioral attributes (Yin et al., 2020). All global feature attributions invariably find interaction intensity, submission counts, and grade trends in the early stages to be the decisive influencers for predicting continuation. Stratifying continuous risk scores into discrete strata allows the institution to distribute its limited outreach interventions based on not only the probability but also the contributing factors to the said probability (B. A. Alnasyan et al., 2025). In all, the trio of leakage-free feature engineering, ensemble models built in an imbalance-aware manner, and feature attributions serves as a coherent framework for early warning systems (Yu et al., 2026). In all, the literature has shown that all of temporally appropriate feature construction, performance-weighted ensemble, and imbalance-aware interpretable modelling is essential for effective dropout prediction, but these three elements have rarely come together in one framework in a single evaluation process using the OULAD benchmark (Dewi et al., 2026). This study seeks to fill this gap by creating seventy-six leakage-free features, performing tuning of decision thresholds and weights of the ensemble model through nested cross-validation and statistically comparing the AUC-weighted model in a strict test set (Dewi et al., 2026).

Methodology

3.1 Dataset and Participant Selection

The validity of student dropout prediction models not only depends on the used learning algorithms but also on the quality, structure, and time validity of the used dataset. This paper uses Open University Learning Analytics Dataset (OULAD), which is freely available benchmarking dataset provided by UK Open University. OULAD dataset consists of seven linked CSV files: studentInfo.csv (student demographic characteristics and the final result), studentRegistration.csv (student enrolment and registration data), studentAssessment.csv (student assessment submission and marks), assessments.csv (assessment details), studentVle.csv (student VLE interactions), vle.csv (VLE site and activities metadata), and courses.csv (module and presentation details). The primary linking variables are id_student, code_module, and code_presentation. The full dataset includes 32,593 enrollments in seven modules (AAA – GGG), four presentations (2013B, 2013J, 2014B, 2014J) for a total of 22 module-presentation combinations. The presentations' duration varies from 240 to 269 days (average 259 days, 37 weeks). Final results

are categorized in four classes: Pass, Fail, Withdrawn, and Distinction. In the current study, the dropout is defined as final_result = Withdrawn. The original dropout rate (all enrollments) is approximately 26-27 %. To calculate the behavioral features based on VLE interaction logs, students have to have at least one activity.
Thus, the inner join operation between the two data tables leaves out 3,510 enrolments that have no VLE activities. Thus, the modelling dataset consists of 29,083 enrolments. It can be noted that the group of the left-out students has an elevated dropout rate of 85.54 % compared to 24.55 % of the remaining VLE-active enrolments. Thus, all further modelling activities will be done in a VLE-active subpopulation of enrolments. The results of modelling will be true for the selected subpopulation and will not be representative of the whole enrolments' population. Thus, the elevated withdrawal rate of left-out students is a type of selection bias that is considered to be a limitation of the present study. As in the case of the train-test split, the stratified sampling keeps the class distribution proportion of 24.55 %. Thus, the training sample has 23,267 enrolments and the test set has 5,816 enrolments. The OULAD dataset is released under an open licence for research purposes and includes only anonymized IDs. All modelling activities were carried out offline without any efforts to identify individuals or add sensitive attributes.

3.2 Leakage-Free Feature Engineering

In total, 76 predictors were developed over six conceptually different groups of features, being strictly consistent with temporal validity rules. Demographics data (33 one-hot-encoded variables representing gender, region, highest qualification, IMD category, age category, and disability) is known at enrolment stage and contains no information derived from the outcomes. Features related to registration time (two variables) were calculated only according to the registration date in relation to course beginning; all the features related to unregistration events (date_unregistration, did_unreg flags, or early withdrawal rates) were deliberately excluded in order to avoid temporal leakages. Weekly VLE engagement features (20 variables) were limited by the first eight weeks. The records with week > 8 were discarded, providing the weekly click aggregates (week1_clicks up to week8_clicks) along with summary measures: early_clicks (first two weeks), mid_clicks (three-six weeks), late_clicks (seven-eight weeks), total_clicks, clicks trend, clicks acceleration, clicks standard deviation, coefficient of variation, peak week, log-transformed total clicks, and zero-clicking flag. Since the mean module duration equals 259 days, eight-week period represents about 31% of the average course period.
VLE activity-type composition features (seven variables) code the raw activity types into general behavioural categories: forum activities, quiz activities, content/resource page views and number of distinct activity types, along with respective ratios. Assessment performance and behaviour features (nine variables) consist of average score, average weighted score, score standard deviation, number of submissions, mean tutor-marked assessment score, late submission rate and missing rate. While the VLE features are restricted to weeks 1–8, the assessment aggregates cover the entire length of the module and are constructed solely on the basis of the submission logs without any withdrawal times encoded in them. Interaction features (five variables) include engagement and performance information in the form of interaction between different variables (e.g., score x clicks, weighted score x clicks, missing x zero-activity, late x missing, registration x clicks) and share the same leakage restrictions of their base features. Following concatenation, zero-variance columns were dropped, missing numeric values were filled in with zeros after aggregation, and all numeric features were z-scored. Thus, the resulting feature matrix only consists of information available in a valid prediction setting. None of the features contains withdrawal time or final status information either directly or indirectly.

3.3 Experimental Design and Data Partitioning

The entire modelling pipeline, starting from the raw CSV input to the evaluation of the held-out test data, is contained in one MATLAB script (R2023b), driven by a fixed random seed (rng(42)). This pipeline contains eight successive steps: (1) loading of the seven source CSV files; (2) feature engineering without leakage; (3) merging of tables and missing value imputation; (4) generation of the 76 features matrix; (5) stratified 80/20 train/test split; (6) inner 10-fold cross-validation performed only on the train part of the dataset for the fitting of the models, threshold

optimisation, and the calculation of AUC-based ensemble weights; (7) refitting of all models on the whole training data; and (8) one evaluation run on the untouched test dataset. A clear distinction is made between the training and test datasets. None of the information from the held-out test dataset is used for the selection of features, thresholding, model weights, or any hyperparameters tuning decisions.

3.4 Classification Models and AUC-Weighted Ensemble

Five models were compared: four ensemble-based approaches and a simple linear method. To avoid overfitting due to high dimensionality, logistic regression was trained using ridge regularisation. In the case of RUSBoost, a combination of AdaBoost with random under-sampling of the majority class at each iteration (500 iterations, learning rate 0.05) makes it very well suited for dealing with the class imbalance problem. Random Forest is an aggregation of 400 decision trees, built on bootstrap samples with $\sqrt{p}$ feature sampling at each split (eight features on average). LogitBoost is a method for additive logistic regression with 300 cycles (learning rate 0.05). GentleBoost is also a method for additive logistic regression with a Newton step approximation with the same parameters. Cross-validation was done on the training set using inner 10-fold cross-validation and out-of-fold AUC scores were calculated for each model. Let AUC_i be the AUC of model i computed using cross-validation. Weights for ensembles were calculated as normalised AUC (sum of all AUCs equals to 1). An ensemble score for sample x is the weighted average of model probabilities. There is no test-set leakage because the weights depend only on the training folds. The resulting normalised weights were approximately 25.4 % (RUSBoost), 25.2 % (Random Forest), 25.5 % (LogitBoost), and 25.1 % (GentleBoost), indicating that no single learner dominated the combination.

3.5 Threshold Optimisation, Evaluation Metrics and Statistical Testing

Decision thresholds were optimized solely using predictions of the inner cross-validation by means of a grid search procedure. For each algorithm the threshold value which gave the highest F1-score was determined and then used as a constant decision point in the held-out test. Thus, threshold determination is separated from the final performance evaluation and optimistic bias is prevented. The held-out test set performance is evaluated according to accuracy, precision, recall, F1-score and the area under the ROC curve (AUC). As the positive class represents only 24.55 % of the dataset, F1-score and AUC are more informative than the accuracy metric alone. The statistical significance of difference between the AUC-weighted ensemble and each single model was established using two methods: McNemar's test for agreement between paired classifications and DeLong's test for difference in the AUC values. Both tests were performed on the predictions of the held-out test set at the significance level $\alpha = 0.05$. This design meets three criteria at once: temporal validity of each and every feature used, complete isolation of the held-out test from all model-building procedures and the statistical analysis of classifiers produced.

Results

4.1. Inner Cross-Validation Performance

Both the models' choice and the ensemble weights calculation were carried out using the training set with the use of stratified 10-fold cross-validation. The obtained results of AUC values of the cross-validation, optimised decision thresholds and normalised ensemble weights are presented in Table 1. Logit Boost yielded the largest value of the inner CV AUC (0.9330), then follow RUS Boost (0.9283), Logistic Regression (0.9243), Random Forest (0.9219), and Gentle Boost (0.9186). The normalised AUC weights of the four components of the ensemble were approximately equal to 25.4%, 25.2%, 25.5%, and 25.1% respectively for RUS Boost, Random Forest, Logit Boost, and Gentle Boost. The almost equal weight distribution suggests that none of the four learners dominated the ensemble since all the learners contributed to the final prediction. The optimised thresholds for F1 scores were equal to 0.205 for Logistic Regression, 0.935 for RUS Boost, 0.250 for Random Forest, 0.050 for Logit Boost, 0.050 for Gentle Boost and 0.235 for the ensemble. All AUC values, the thresholds and the ensemble weights were calculated only based on the inner CV in the training set. The held-out test set was not used at any stage of the modelling shown in Table 1.

Model	CV AUC	Tuned Threshold (F1-maximised)	Normalized Ensemble Weight (%)
Logistic Regression	0.9243	0.205	Nil
RUS Boost	0.9283	0.935	25.4
Random Forest	0.9219	0.250	25.2
Logit Boost	0.9330	0.050	25.5
Gentle Boost	0.9186	0.050	25.1

Table 1. Inner 10-Fold Cross-Validation Performance, Tuned Decision Thresholds, and Normalized Ensemble Weights

4.2 Held-Out Test Performance

Generalisation ability of the model was assessed only once on the strictly isolated test set of 20 %. Table 2 indicates that the best AUC (0.9296) and F1-score (0.7498) have been shown by the AUC-weighted ensemble. LogitBoost exhibited the best accuracy (0.8592) and precision (0.6989), while the RUSBoost delivered the best recall (0.9524). Such discrepancies arise due to classic trade-offs between precision and recall for a 24.55 % prevalence of the positive class, where RUSBoost maximizes sensitivity at the expense of specificity, while LogitBoost takes a more conservative approach. Ensemble holds an intermediary position, maintaining high recall (0.8908), while increasing precision compared to RUSBoost. Despite the linear decision boundary, logistic regression has demonstrated relatively good AUC (0.9181), which proves that the leakage-free feature set is very informative. Shown in Table 2.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.8477	0.6310	0.9139	0.7466	0.9181
RUS Boost	0.8346	0.6034	0.9524	0.7387	0.9244
Random Forest	0.8382	0.6169	0.8999	0.7320	0.9149
Logit Boost	**0.8592**	**0.6989**	0.7493	0.7232	0.9295
Gentle Boost	0.8404	0.7033	0.6057	0.6509	0.9159
AUC-Weighted Ensemble	0.8540	0.6473	0.8908	**0.7498**	**0.9296**

Table 2. Final Model Performance on the Held-Out Test Set

AUC-Weighted Ensemble attained the best value for AUC (0.9296) and F1 score (0.7498) out of all the models being assessed. Logit Boost obtained the best values for accuracy (0.8592) and precision (0.6989) while RUS Boost obtained the best value for recall (0.9524). These differences are due to the trade-off in the metrics. RUS Boost obtained a high value for recall while obtaining a low value for precision, whereas Logit Boost obtained high values for accuracy and precision while obtaining a low value for recall. The ensemble resulted in a better F1 score while having a high recall (0.8908). Baseline Logistic Regression obtained an AUC of 0.9181 and an F1 score of 0.7466. All metrics were obtained once after test set completion. No test set information was used in these steps.

4.3 Statistical Significance

McNemar and DeLong tests were carried out for the predictions on the holdout dataset with $\alpha = 0.05$. McNemar test showed statistically significant differences in the performance of the ensemble and other algorithms: $p = 0.0128$ for logistic regression, $p < 0.001$ for RUSBoost, Random Forest, and GentleBoost, and $p = 0.0054$ for GentleBoost. The McNemar test for comparing the performance of the ensemble and LogitBoost returned a non-significant result, with $p = 0.1639$. DeLong test gave an equivalent result: $p < 0.001$ for comparing the AUC of the ensemble and other four

algorithms and p = 0.8105 for the ensemble and LogitBoost shown in Table 3 and 4. All these tests prove that the ensemble algorithm is significantly better in performance than four out of five of the other algorithms.

Comparison	p-value	Significant (α = 0.05)
Ensemble vs Logistic Regression	0.0128	Yes
Ensemble vs RUS Boost	<0.001	Yes
Ensemble vs Random Forest	<0.001	Yes
Ensemble vs Logit Boost	0.1639	No
Ensemble vs Gentle Boost	0.0054	Yes

Table 3 McNemar Test Results

The McNemar test indicated statistically significant differences in classification outcomes between the ensemble and Logistic Regression, RUS Boost, Random Forest, and Gentle Boost. The difference between the ensemble and Logit Boost was not statistically significant (p = 0.1639).

Comparison	p-value	Significant (α = 0.05)
Ensemble vs Logistic Regression	<0.001	Yes
Ensemble vs RUS Boost	<0.001	Yes
Ensemble vs Random Forest	<0.001	Yes
Ensemble vs Logit Boost	0.8105	No
Ensemble vs Gentle Boost	<0.001	Yes

Table 4 DeLong AUC Comparison

The DeLong test revealed statistically significant differences in AUC values for the ensemble and Logistic Regression, RUS Boost, Random Forest, and Gentle Boost models. There were no statistically significant differences between AUC values for the ensemble and Logit Boost models (p = 0.8105). The results of both McNemar and DeLong tests indicate that there are statistically significant differences between AUC-Weighted Ensemble and Logistic Regression, RUS Boost, Random Forest, and Gentle Boost models.

4.4. ROC Curve Analysis

Ranking performance for all six classifiers was excellent, with test AUCs exceeding 0.91 (Figure 1). Curves for the ensemble (0.9296) and LogitBoost (0.9295) are nearly overlapping over the whole range of false positive rates, which confirms the DeLong test result. At the false positive rate of about 0.10, the ensemble achieves a true positive rate of about 0.80, implying that even at a relatively conservative operating point, a majority of the dropouts is identified. Logistic regression is still quite competitive (AUC 0.9181), but the curve is distinctly separated from the two best-performing classifiers. Thus, the ROC analysis supports both the superiority of the ensemble in terms of numbers and the equivalence of the ranking performance.

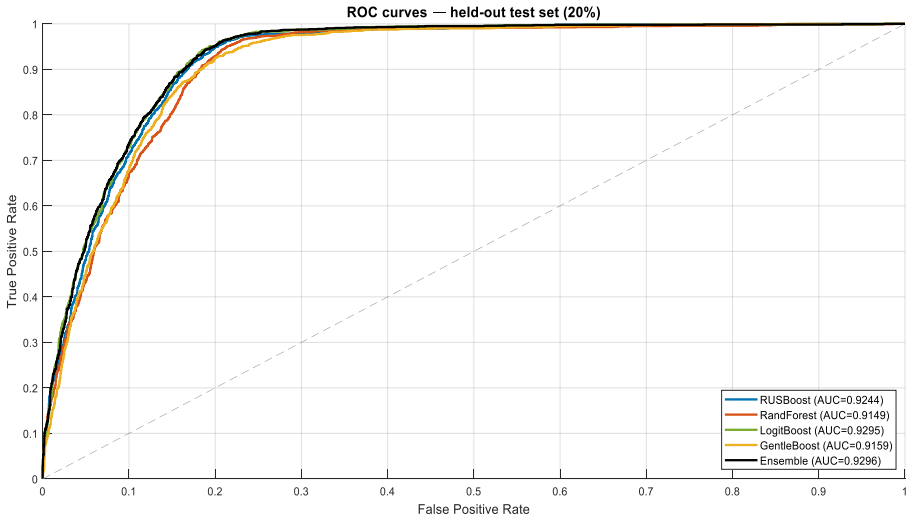


Figure 1 – ROC Curves

4.5. Precision–Recall Analysis

In view of the observed 24.55% drop-out occurrence rate, precision-recall curves provide more useful data, which is not affected by the huge amount of true negatives (Figure 2). The ensemble demonstrates average precision of 0.7729, being better than that of the logistic regression model (0.7581) and GentleBoost (0.7244). The precision level remains fairly high for quite a wide range of recall levels; therefore, the model identifies positive examples effectively and does not introduce too many false positives.

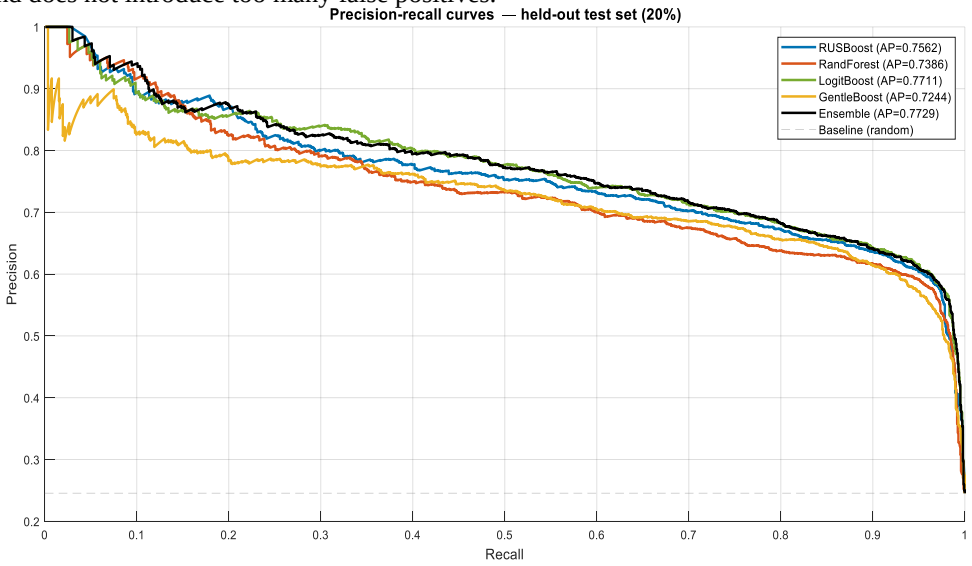


Figure 2 – PR Curves

4.5 Feature Importance

Variance reduction-based random forest predictor importance was applied to measure the individual contribution of the features. The most important predictors were mostly the assessment-related features, namely num_assessments, score_std, avg_score, tma_avg_score, avg_weighted_score, score_x_clicks, and weighted_x_clicks. The weekly

VLE-based features like week7_clicks and week8_clicks showed up to be somewhat less important, starting roughly from rank 12 (Figure 3 and 4). Neither of the registration-related features, namely reg_days_before and reg_norm, made it into the top 20 predictors, but some interaction features provided some meaningful contributions to the feature importance ranking. In terms of the feature group importance, the assessment features turned out to be the most important, followed by weekly VLE behaviour, whereas demographic and registration-related features were less important.

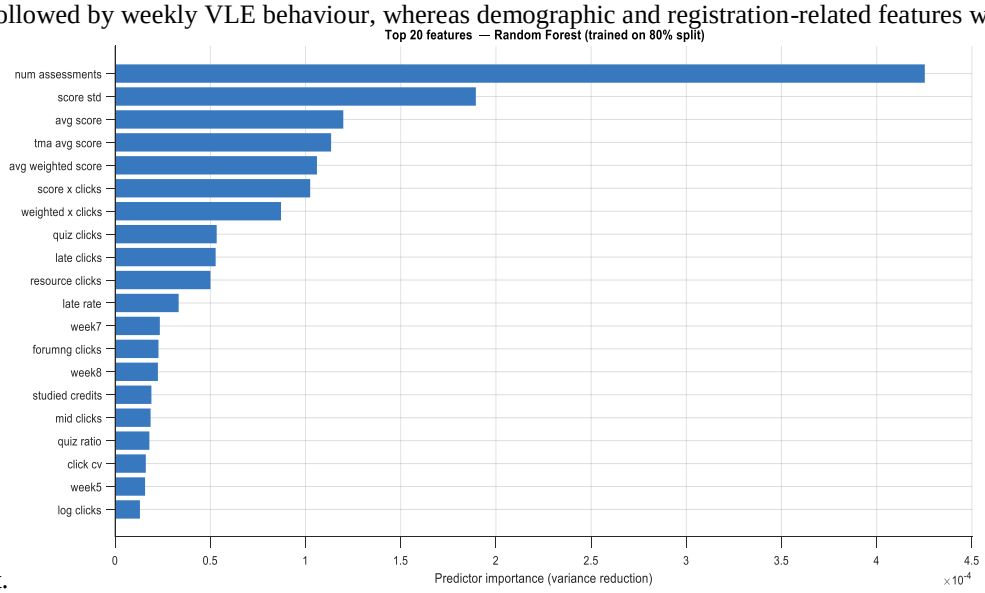


Figure 3 – Top 20 Feature Importance

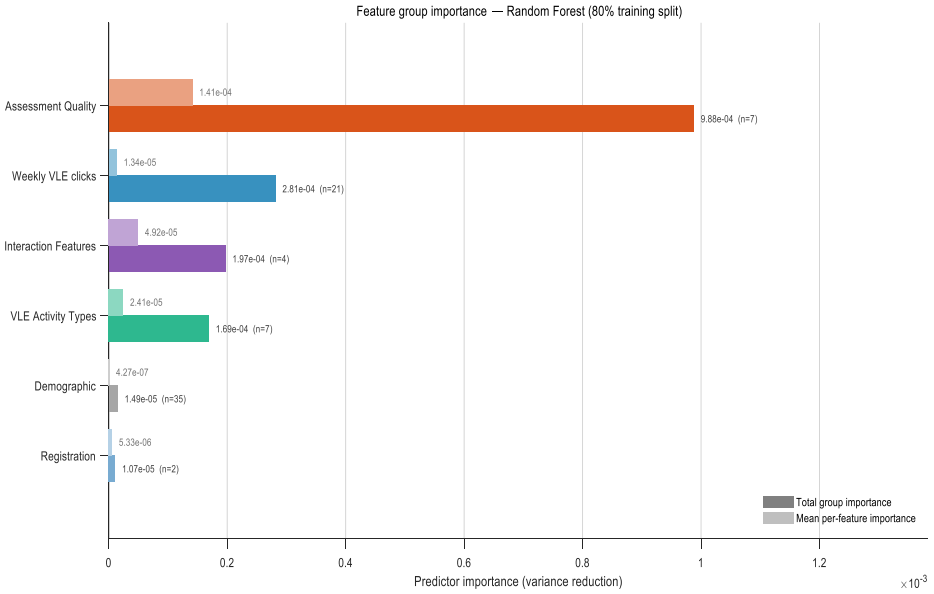


Figure 4 – Feature Group Importance

Confusion Matrix Analysis

The row normalised confusion matrices of the five tested models on the hold-out test set are displayed in Figure 5 below. They show the variation in classification of the non-dropout and dropout samples among the different models. Recall of dropout samples was the highest among RUS Boost models at 95.2%, with only 4.8% being classified as non-dropout. AUC-Weighted Ensemble model classified 89.1% of dropout samples, whereas Logistic Regression, Random Forest, Logit Boost, and Gentle Boost classified 91.4%, 90.0%, 74.9%, and 60.6%, respectively.

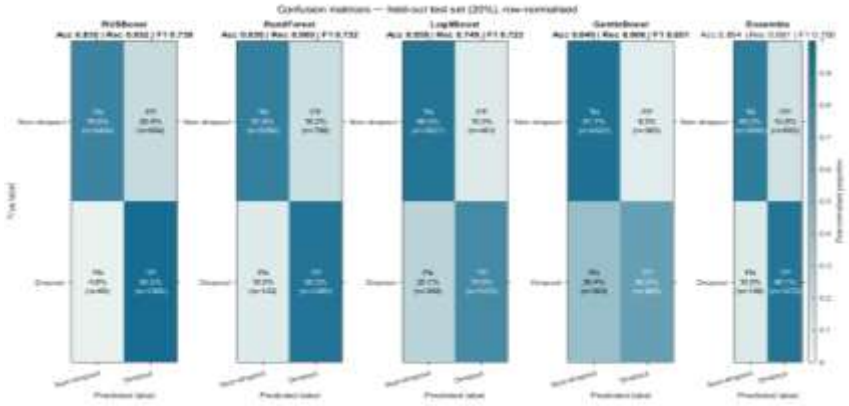


Figure 5 – Confusion Matrix Grid

Threshold Analysis

In Figure 6, the graph between the decision threshold and the F1 score for the hold-out test set is provided, along with the thresholds selected using cross-validation represented through filled points. From the plots, it can be seen that there is significant variation in F1 scores depending on the choice of the decision threshold, especially for boosting algorithms. The chosen thresholds are based on Table 6 inner cross-validation results.

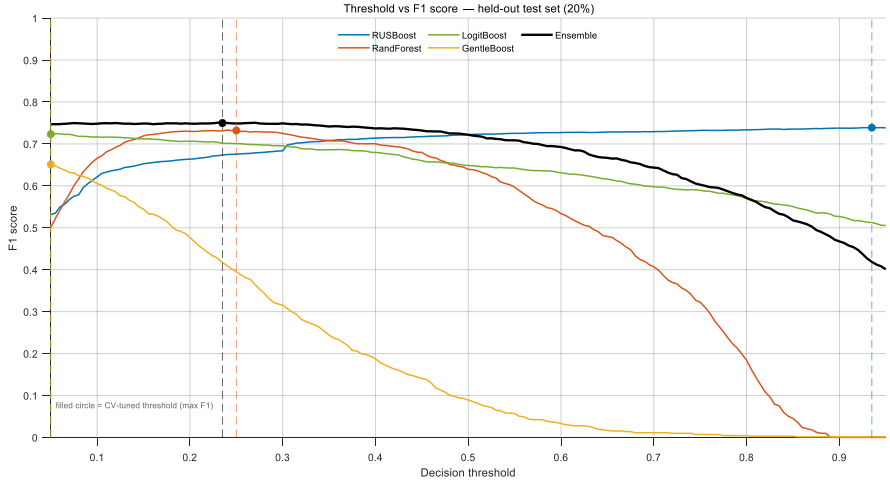


Figure 6 – Threshold vs F1

Module-Level Analysis

Figure 7 displays the actual dropout rates versus the ensemble-predicted dropout rates across various presentations of the module. There is a very strong relationship between the predicted dropout rates and the observed dropout rates with a Pearson correlation coefficient of r = 0.962. The module-wise analysis indicates variability in the dropout rates depending on the module and presentation of the course. For instance, the actual dropout rate for CCC was 45% in 2014B and 38% in 2014J, but the ensemble predictions were 60% and 47% respectively. The actual dropout rates for DDD were 31%, 26%, 33%, and 33% across the four presentations, whereas the predicted dropout rates were 40%, 38%, 37%, and 41% respectively.

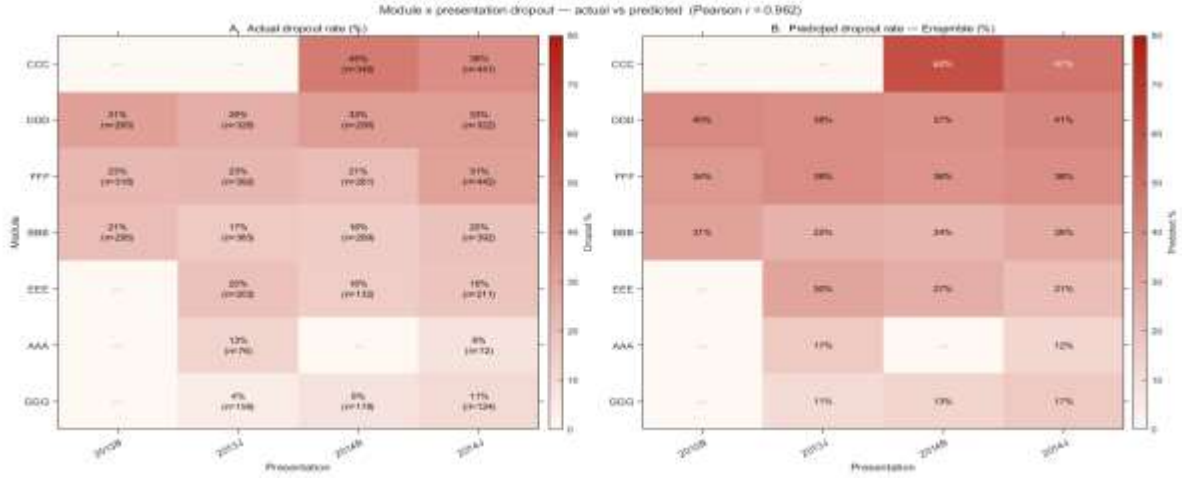


Figure 7 – Module Heatmap

Discussion

The findings indicate that a leakage-controlled modeling approach is able to successfully discriminate between dropouts and non-dropout enrollments with good discrimination power in the sub-population of active participants in the VLE of OULAD. All the tested algorithms gave an AUC score above 0.91, which means that the 76 temporal features have significant predictive power. The AUC-based ensemble had the best AUC (0.9296) and F1-score (0.7498), while LogitBoost had the best accuracy and RUSBoost the best recall score. These results show the well-known trade-off between precision and recall given the imbalance of 24.55 % in favor of positive class. While the ensemble and LogitBoost models were statistically not significantly different based on the McNemar and DeLong test, the former gave better recall score (0.8908 compared to 0.7493) and harmonic mean.

Further clarification of the origins of prediction is provided by feature-importance analysis. The assessment-related variables – number of assessments made, score standard deviation, mean score, and tutor-assessed assessment performance – dominated the top rankings, while the pure weekly clicks on the VLE scored low, as did demographic and timing variables. Feature importance for interaction terms between assessment performance and engagement volume were also high, which shows that interactions between assessment behavior and VLE use contain more information than pure numbers. Taken together, this shows that the early engagement with academics and consistent performance is a better predictor of dropout than the pre-admission factors and the clicks on the VLE. From a practical perspective, institutions developing predictors need to include both of these elements into their calculations. The explicit leakage audit process represents the key methodological contribution of this paper. Through the exclusion of all post-withdrawal variables, as well as the exclusion of VLE variable availability after the first eight weeks, the approach guarantees that every single input would be available at a realistic prediction

horizon. At the same time, the exclusion of 3,510 VLE-inactive enrollments, characterized by an 85.54 % dropout rate, implies that the results obtained would refer only to the subpopulation of active users. In this case, the design of the cohort represents not only an advantage since it allows the authors to avoid learning from those cases that do not provide any behavioral signal but also the limitation since the model cannot be used for predicting behavior of non-active users.

A number of limitations limit the generalizability of the results. The first limitation is the selection bias that results from excluding the inactive learners in VLEs. The second limitation is that while assessment measures do not have any timestamps for withdrawals, they cover the entire period and thus cannot be considered an early warning measure. Thirdly, OULAD data is based on a single provider and two-year period (2013 to 2014). Generalizability to other providers or more recent data sets has not been considered. Fourthly, the evaluation uses random stratified hold-out instead of year-based or presentation-based evaluation that makes the true temporal shift possible. Lastly, the global ranking of Random Forest importance gives limited information; individual level explanation was not attempted. The above restrictions directly suggest potential ways of extending the research. Module-specific or hierarchical model variants may reflect the heterogeneous pedagogical approach within the OULAD seven modules. Current gradient-boosting algorithms like XGBoost and LightGBM could be tested against our model under the same leakage restrictions. Local interpretation techniques could help convert global importance findings into practical diagnostic tools for academic advisors. Forward chaining validation across different presentations or academic years in combination with external datasets of other institutions would be necessary to test the robustness of the results through time and institutions. Above all, a good performance of the model is not enough to ensure its utility; controlled trials are required to check whether the interventions suggested by the scores actually increase retention rates. In conclusion, the paper demonstrates that paying careful attention to temporal validity, nested validation, and multi-metric testing allows to develop a discriminative educational benchmark in terms of statistically significant results. At the same time, the need for transparency in cohort selection and reporting multiple metrics of performance and validation is emphasized.

Conclusion

This paper described a framework to predict student dropouts from the Open University Learning Analytics Dataset using machine learning with leakage control. In this study, after removal of the VLE inactive enrollments, a total of 29,083 observations and 76 temporally valid features were selected. An AUC weighted average of RUSBoost, Random Forest, LogitBoost and GentleBoost with tuning fully nested within cross-validation gave an AUC of 0.9296, F1 score of 0.7498 and accuracy of 0.8540 on the held out test set. Assessment engagement and assessment variability proved to be the best predictors whereas the number of clicks was less significant. These results prove statistically better than most of the models. This study shows the power of considering temporality of data, threshold optimization and evaluation based on multiple metrics to obtain reliable discriminative predictions. On the other hand, the exclusion of the VLE inactive observations, one single institutional cohort and no forward validation reduce the generalizability of the results. Future studies could include module-wise prediction, gradient boosting, local explanation of predictions, external/forward validation, and ultimately testing if triggered interventions improve retention.

References

1. Algarni, S., & Sheldon, F. T. (2025). Personalized course recommendations leveraging machine and transfer learning toward improved student outcomes. *Machine Learning and Knowledge Extraction*, *7*(4), 138. https://doi.org/10.3390/make7040138
2. Alnasyan, B. A., Basheri, M., & Alassafi, M. O. (2025). A comprehensive comparative analysis of deep learning models for student performance prediction in virtual learning environments: Leveraging the OULA dataset and advanced resampling techniques. *IEEE Access*, *13*, 75953–75972. https://doi.org/10.1109/access.2025.3564719

3. Alnasyan, B., Basheri, M., & Alassafi, M. (2024). The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *6*, 100231. https://doi.org/10.1016/j.caeai.2024.100231
4. Alzabidi, E., Yakut, Ö., Saad, S., & Fındık, O. (2026). SOD-FE: A supervised outlier detection and feature engineering approach for student dropout prediction. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-56591-6
5. Bettahi, A., Belouadha, F.-Z., & Harroud, H. (2025). A modular and explainable machine learning pipeline for student dropout prediction in higher education. *Algorithms*, *18*(10), 662. https://doi.org/10.3390/a18100662
6. Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC course using random forest model. *Information*, *12*(11), 476. https://doi.org/10.3390/info12110476
7. Davis, S. E., Matheny, M. E., Balu, S., & Sendak, M. P. (2023). A framework for understanding label leakage in machine learning for health care. *Journal of the American Medical Informatics Association*, *31*(1), 274–280. https://doi.org/10.1093/jamia/ocad178
8. Deeva, G., De Smedt, J., & De Weerdt, J. (2022). Educational sequence mining for dropout prediction in MOOCs: Model building, evaluation, and benchmarking. *IEEE Transactions on Learning Technologies*, *15*(6), 720–735. https://doi.org/10.1109/tlt.2022.3215598
9. Dewi, L. A. P., Anggraini, N., & Senubekti, M. A. (2026). Predicting customer dropout from early behavioral data in a vocational EdTech platform. *International Journal of Economics Social and Technology*, *5*(3), 626–637. https://doi.org/10.59086/ijest.v5i3.2127
10. Fauszt, T. (2026). Time-Consistent prediction in higher education: A framework for preventing data leakage in longitudinal models. *Information*, *17*(6), 581. https://doi.org/10.3390/info17060581
11. Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, *9*, 133076–133090. https://doi.org/10.1109/access.2021.3115851
12. Guamán-Hidalgo, E., & Enciso, L. (2026). Early prediction of student dropout through machine learning in educational data mining: A comparative analysis of algorithms and data balancing techniques. *Applied Sciences*, *16*(18), 8915. https://doi.org/10.3390/app16188915
13. Hutt, S., Baker, R. S., Ashenafi, M. M., Andres-Bray, J. M., & Brooks, C. (2022). Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology*, *53*(4), 756–775. https://doi.org/10.1111/bjet.13231
14. Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, *11*(7), 3130. https://doi.org/10.3390/app11073130
15. Khoushehgir, F., & Sulaimany, S. (2023). Negative link prediction to reduce dropout in Massive Open Online Courses. *Education and Information Technologies*, *28*(8), 10385–10404. https://doi.org/10.1007/s10639-023-11597-9
16. Krüger, J. G. C., Britto, A. de S., Jr., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, *233*, 120933. https://doi.org/10.1016/j.eswa.2023.120933
17. M, A. C., S, H., & K, V. K. (2026). An Intelligent Early Warning System for Student Dropout Prediction using Deep Learning on the OULAD Dataset. *2026 7th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 1798–1803. https://doi.org/10.1109/icicv68925.2026.11554644
18. Meeraa Shahibo, F. F., & Ahangama, S. (2025). Student performance prediction using virtual learning environment (VLE) interactions. *European Conference on E-Learning*, *24*(1), 104–112. https://doi.org/10.34190/ecel.24.1.3986

19. Mincache, R. B., Catharin, L. G., Teixeira, L. de O., Colanzi, T. E., da Costa, Y. M. e. G., & Feltrim, V. D. (2026). Academic dropout prediction using large-scale static institutional data: A multi-scenario machine learning study. *Technologies*, *14*(9), 556. https://doi.org/10.3390/technologies14090556
20. Moresi, M., J. Gómez, M., & Benotti, L. (2021). Predicting students' difficulties from a piece of code. *IEEE Transactions on Learning Technologies*, *14*(3), 386–399. https://doi.org/10.1109/tlt.2021.3092998
21. Mubarak, A. A., Cao, H., & Hezam, I. M. (2021). Deep analytic model for student dropout prediction in massive open online courses. *Computers &Amp; Electrical Engineering*, *93*, 107271. https://doi.org/10.1016/j.compeleceng.2021.107271
22. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. https://doi.org/10.1016/j.caeai.2022.100066
23. Oğul, V. G., & Balcıoğlu, Y. S. (2026). Early identification of at-risk students in online learning environments: A learning analytics approach using machine learning models. *OPUS Journal of Society Research*, *23*(2026), 1–17. https://doi.org/10.26466/opusjsr.1948539
24. Oyedotun, S. A., Ejenarhome, O. P., & Oise, G. P. (2025). Learning analytics and predictive modeling: Enhancing student success through data-driven insights. *Journal of Science Research and Reviews*, *2*(3), 42–51. https://doi.org/10.70882/josrar.2025.v2i3.77
25. Park, H. S., & Yoo, S. J. (2021). Early Dropout Prediction in Online Learning of University using Machine Learning. *JOIV : International Journal on Informatics Visualization*, *5*(4), 347. https://doi.org/10.30630/joiv.5.4.732
26. Patel, K. K., & Amin, K. (2024). Predictive modeling of dropout in MOOCs using machine learning techniques. *The Scientific Temper*, *15*(02), 2199–2206. https://doi.org/10.58414/scientifictemper.2024.15.2.32
27. Paz, H. R. (2025, November 14). *A leakage-aware data layer for student analytics: The capire framework for multilevel trajectory modeling*. arXiv.org. https://arxiv.org/abs/2511.11866
28. Rabelo, A. M., & Zárate, L. E. (2025). A model for predicting dropout of higher education students. *Data Science and Management*, *8*(1), 72–85. https://doi.org/10.1016/j.dsm.2024.07.001
29. Raj, N. S., & V G, R. (2022). Early prediction of student engagement in virtual learning environments using machine learning techniques. *E-Learning and Digital Media*, *19*(6), 537–554. https://doi.org/10.1177/20427530221108027
30. Singh, A. (2025). *Predictive analytics for student success: Investigating how machine learning can predict student success and identify early warning signs for intervention*. Elsevier BV. https://doi.org/10.2139/ssrn.5202086
31. Sitam, G., & Anusha, K. C. (2026). When should educational AI stay silent? Abstention-Aware student risk prediction using learning analytics. *Journal of Advanced College of Engineering and Management*, *12*, 215–238. https://doi.org/10.3126/jacem.v12i01.93932
32. Sunil P. Patel, Hemant N. Patel, & Jigar Patel. (2026). Modeling student retention and success using explainable machine learning techniques. *International Journal of Computer Information Systems and Industrial Management Applications*, *18*(9s), 383–406. https://doi.org/10.70917/ijcisim-2026-3446
33. Tan, C. J., Lim, T. Y., Liew, T. K., & Lim, C. P. (2022). An intelligent tool for early drop-out prediction of distance learning students. *Soft Computing*, *26*(12), 5901–5917. https://doi.org/10.1007/s00500-021-06604-5
34. Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*, 102474. https://doi.org/10.1016/j.techsoc.2024.102474

35. Vo, N. T., Le, Q. N., & Le, H. (2026). Artificial intelligence-driven early prediction of student dropout and academic outcomes in higher education: A comparative study of advanced machine learning approaches. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, *13*(1). https://doi.org/10.4108/eetinis.131.11758
36. Wijiyanto, W. (2026). Course-Disjoint evaluation and capacity-aware triage for student dropout risk prediction. *Journal of Applied Data Sciences*, *7*(3), 1771–1789. https://doi.org/10.47738/jads.v7i3.1419
37. Yin, S., Lei, L., Wang, H., & Chen, W. (2020). Power of attention in MOOC dropout prediction. *IEEE Access*, *8*, 202993–203002. https://doi.org/10.1109/access.2020.3035687
38. Yu, J., Xu, C., & Wang, H. (2026). A lightweight multi-scale deep separable network for early dropout prediction in MOOCs: Balancing accuracy and educational deployability. *Frontiers in Education*, *11*. https://doi.org/10.3389/feduc.2026.1861182