Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 5, No. 01, 2026 ♦ PP. 18–24
Article AOI: Attps://doi.org/10.54216/IJAIET.050103



FULL LENGTH ARTICLE

# Early Identification of At-Risk Students in Virtual Learning Environments Using Ensemble Machine Learning and Behavioural Analytics

**Ahmed Abd El-Badie Abd Allah Kamel**[1,*]

[1] Associate Professor of Computer Science and Director of the Monitoring and Technical Support Unit, Measurement and Evaluation Center, Mansoura University, Egypt

**Emails:** ahmed abdelbadie@mans.edu.eg

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026 ⋆ Corresponding author

## ABSTRACT

Identifying students at risk of academic failure or course withdrawal early in their enrolment remains a major challenge in higher and distance education. This study evaluates seven machine-learning classifiers for early risk prediction using an OULAD-inspired dataset of 7,895 student records, with eight demographic and administrative features and eight Virtual Learning Environment (VLE) engagement features. Models were assessed using five-fold stratified cross-validation. Logistic Regression achieved the highest mean AUC-ROC (0.786), AdaBoost the highest mean accuracy (0.716), and Gradient Boosting Decision Tree (GBDT) the highest mean recall (0.769). GBDT also achieved an AUC-ROC of 0.782 (± 0.003), accuracy of 0.708 (± 0.005), and F1-score of 0.729 (± 0.006). Gini-based feature importance identified late submissions (0.304), total VLE clicks (0.150), and first assessment score (0.135) as the strongest predictors. These results suggest that early learning-management-system traces can support timely student outreach, while also underscoring the need for equitable, human-led use of predictive analytics.

**Keywords:** Learning analytics ▪ At-risk student prediction ▪ Gradient boosting ▪ Ensemble machine learning ▪ Virtual learning environment ▪ Educational data mining ▪ Early warning systems

## 1. INTRODUCTION

The rapid expansion of online and blended learning has generated large volumes of detailed student information. Learning Management Systems (LMSs) and Virtual Learning Environments (VLEs) record logins, resource use, and forum contributions, creating extensive records of students' academic activity [1, 2]. Turning these records into useful intelligence presents both technical and ethical challenges.

Academic risk is a significant concern because many students in distance and blended programmes do not complete their studies or submit essential assessments on time. At the Open University in the United Kingdom, combined module withdrawal and failure rates have reportedly exceeded 35%. Identifying students who need support before they reach the midpoint of a module gives educators time to provide tutoring, motivation, and tailored learning materials.

Educational Data Mining (EDM) and Learning Analytics (LA) provide methodological foundations for early-warning systems. Supervised machine learning is widely studied because it can process large cohorts and combine different data types, including clickstream counts and demographic indicators [3, 4]. However, direct comparisons using consistent feature sets and evaluation protocols remain important for selecting reliable methods.

This study compares seven classifiers using an OULAD-

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

inspired dataset that reflects demographic and engagement patterns. It addresses three research questions:

**RQ1.** Which supervised classifier achieves the strongest accuracy, F1-score, and AUC-ROC for binary at-risk identification using VLE engagement and demographic features?

**RQ2.** Which feature groups and individual features contribute most to at-risk prediction?

**RQ3.** At what engagement-data threshold can reliable early-warning signals be obtained?

Section 2 reviews related work; Section 3 describes the dataset and methods; Section 4 presents results; Section 5 discusses their implications and limitations; and Section 6 concludes the paper.

## 2. RELATED WORK

### 2.1 Artificial Intelligence in Higher Education

Zawacki-Richter et al. [2] reviewed 146 studies published from 2007 to 2018 and found that higher-education institutions primarily applied artificial intelligence to adaptive learning, assessment, and performance forecasting. They also identified a gap between technical prediction capabilities and educational theory, including the limited representation of educators' perspectives. Recent surveys report that this tension remains relevant as research output grows [3].

### 2.2 Learning Analytics and OULAD

The Open University Learning Analytics Dataset (OULAD) is a widely used resource containing anonymised academic information for 32,593 students across 22 undergraduate courses, including demographic data, assessments, and more than 10.6 million VLE interactions [1]. It has supported research on at-risk prediction and is distributed under a Creative Commons Attribution 4.0 licence [5, 6].

Adnan et al. [5] reported AUC-ROC values between 0.78 and 0.84 for Random Forest and Gradient Boosting models using assessment scores and cumulative VLE clicks from the first 10–50% of a module. Waheed et al. [6] studied Long Short-Term Memory (LSTM) networks over weekly engagement windows. Their sequential approach was competitive later in a module, while ensemble classifiers performed strongly in earlier prediction windows.

### 2.3 Feature Engineering for At-Risk Prediction

Feature engineering strongly influences learning-analytics predictions. A review of student-performance prediction studies grouped features into prior academic performance, demographic traits, participation and attendance, psychosocial influences, and institutional attributes. Prior achievement and engagement commonly emerged as the most predictive categories, with demographic characteristics providing additional information [3].

VLE engagement is often measured through clicks on resources, quizzes, and forums, as well as submission patterns and login frequency [4, 7]. Late or missed submissions can be early indicators of failure [5]. Socioeconomic indicators such as the Index of Multiple Deprivation (IMD) have also been associated with withdrawal risk, raising equity considerations for predictive models [1].

### 2.4 Ensemble Methods in Educational Data Mining

Random Forest and Gradient Boosting are common ensemble methods in EDM. Random Forest combines trees trained on bootstrap samples, while Gradient Boosting sequentially adds trees that learn from earlier prediction errors [8, 9]. Studies of educational tabular data have found that Gradient Boosting can outperform Random Forest, although the gains depend on the data and feature set [3, 4].

Interpretability is also essential in educational settings. SHapley Additive exPlanations (SHAP) provide a game-theoretic approach to attributing model outputs to input features [10]. This study instead uses Gini-based and permutation feature importance to examine the fitted models.

## 3. METHODOLOGY

### 3.1 Dataset Description

The dataset was constructed to mirror the statistical properties of OULAD described by Kuzilek et al. [1] and subsequent analyses [5, 6]. It models one representative module with four possible outcomes (Distinction, Pass, Fail, and Withdrawn) and contains 7,895 student records. The single-module scope limits confounding from differences in module difficulty, duration, and assessment structure.

The 16 input features comprise eight demographic and administrative variables and eight VLE engagement variables. The binary target marks Fail or Withdrawn as at-risk and Pass or Distinction as not-at-risk. There are 3,742 at-risk students (47.4%) and 4,153 not-at-risk students (52.6%). The four outcome proportions are Pass 39.8%, Fail 25.7%, Withdrawn 21.7%, and Distinction 12.8%.

### 3.2 Data Preprocessing

Categorical features were label-encoded. Continuous and ordinal features were standardised to zero mean and unit variance for scale-sensitive models (Logistic Regression, Naive Bayes, and MLP); tree-based ensembles used the raw encoded values. No missing values were reported.

### 3.3 Classifiers

Seven classifiers were selected to span linear, tree-based, ensemble, probabilistic, and neural approaches:

**LR** Regularised Logistic Regression ($L_2$, $C = 1.0$), used as a linear baseline.

**DT** Single CART decision tree with maximum depth 8.

**RF** Random Forest of 100 trees with maximum depth 10 [8].

**GBDT** Gradient Boosting ensemble of 100 trees, maximum depth 4, learning rate 0.1, and subsample 0.8 [9].

**AdaBoost** Adaptive boosting with 100 weak learners and learning rate 0.5.

**NB** Gaussian Naive Bayes classifier.

**Table 1.** Feature descriptions for the study dataset ($N = 7{,}895$).

| Feature | Type | Group | Description / Coding |
|---|---|---|---|
| Gender | Categorical | Demographic | F (56%), M (44%) |
| Age band | Categorical | Demographic | 0–35 (60%), 35–55 (31%), 55+ (9%) |
| Region | Categorical | Demographic | Seven UK geographic regions |
| Highest education | Ordinal | Demographic | Five levels: no qualifications to postgraduate |
| IMD band | Ordinal | Demographic | Deprivation decile, 0–10% to 90–100% |
| Disability | Binary | Demographic | Declared disability: Y/N |
| Previous attempts | Integer | Administrative | Prior module attempts (0–3) |
| Studied credits | Integer | Administrative | Credit load: 30, 60, 90, or 120 |
| Total VLE clicks | Integer | VLE engagement | Cumulative interaction count |
| Average daily clicks | Continuous | VLE engagement | Mean clicks per calendar day |
| Resource clicks (%) | Continuous | VLE engagement | Proportion directed to resource pages |
| Forum clicks (%) | Continuous | VLE engagement | Proportion directed to forum posts |
| Assignment clicks (%) | Continuous | VLE engagement | Proportion directed to assignments |
| Late submissions | Integer | VLE engagement | Count of overdue submissions |
| First assessment score | Continuous | VLE engagement | First TMA score (0–100) |
| Average assessment score | Continuous | VLE engagement | Mean score across all TMAs |

**MLP** Feedforward neural network with hidden layers [64, 32], ReLU activation, and early stopping.

All models were implemented using scikit-learn [11].

### 3.4 Evaluation Protocol

Five-fold stratified cross-validation preserved the class ratio in each fold. Accuracy, precision, recall, F1-score, and AUC-ROC were averaged across folds. A separate 80/20 stratified split was used to inspect the final models using confusion matrices, ROC curves, and feature importances. Gini-based importances were supplemented with permutation importance on the held-out test partition [8].

## 4. RESULTS

### 4.1 Dataset Characteristics

Figure 1 summarises the outcome distribution, VLE clicks by outcome, at-risk proportions by education level, and late-submission counts by risk group. Pass and Distinction students show higher VLE click volumes than Fail and Withdrawn students. The at-risk proportion generally declines with higher educational attainment, while at-risk students show a broader distribution of late submissions. These patterns motivate the use of engagement features as early-warning predictors.

### 4.2 Cross-Validation Performance

Table 2 reports the cross-validation results. Accuracy ranges from 0.680 to 0.716, with linear and ensemble models clustering near 0.708–0.716. Logistic Regression achieves the highest AUC-ROC (0.786), narrowly ahead of GBDT (0.782); this difference is less than one standard deviation. The single Decision Tree has the lowest accuracy (0.680) and AUC-ROC (0.724), consistent with the benefit of ensemble diversity.

As shown in Figure 2, AdaBoost achieves the highest accuracy (0.716) and precision (0.731), while GBDT achieves the highest recall (0.769). Because early-warning applications may assign a higher cost to missing an at-risk student than to a false alert, GBDT is selected for the subsequent model analysis.

### 4.3 ROC Analysis

Figure 3 shows ROC curves for all classifiers on the held-out test partition. GBDT and Logistic Regression trace the strongest curves. The Decision Tree curve is more irregular, reflecting less stable probability estimates. All models perform above the random-classifier diagonal, indicating predictive signal in the 16 input features within the early module window.

### 4.4 Feature Importance and Confusion Matrix

Figure 4 presents Gini-based feature importance and the GBDT confusion matrix for the held-out test set. Late submissions are the leading feature (importance 0.304), followed by total VLE clicks (0.150) and first assessment score (0.135). The next five engagement features contribute a further 29.3%, while demographic characteristics contribute less than 5% of total importance.

In the holdout partition ($n = 1{,}579$), GBDT correctly classifies 488 of 748 at-risk students (recall 65.2%) and 641 of 831 not-at-risk students (specificity 77.1%). The resulting accuracy is 71.5%, close to the cross-validation estimate. The 260 false negatives represent 34.8% of the at-risk cohort and potential missed opportunities for intervention.

### 4.5 Correlation and Engagement Profiles

Figure 5 shows pairwise correlations among engagement features and the binary outcome, alongside normalised engagement profiles for the two risk groups. The strongest associations with the not-at-risk outcome are first assessment score ($r = 0.45$) and average assessment score ($r = 0.42$). Late submissions are negatively correlated with the outcome ($r = -0.47$), the strongest single pairwise association. Assessment scores and click measures are positively intercorrelated ($r = 0.25$–$0.40$). Not-at-risk students have higher values on the positive engagement dimensions and fewer late submissions; the largest differences concern late submissions and

**Figure 1. Dataset Overview: Demographic and Engagement Characteristics**

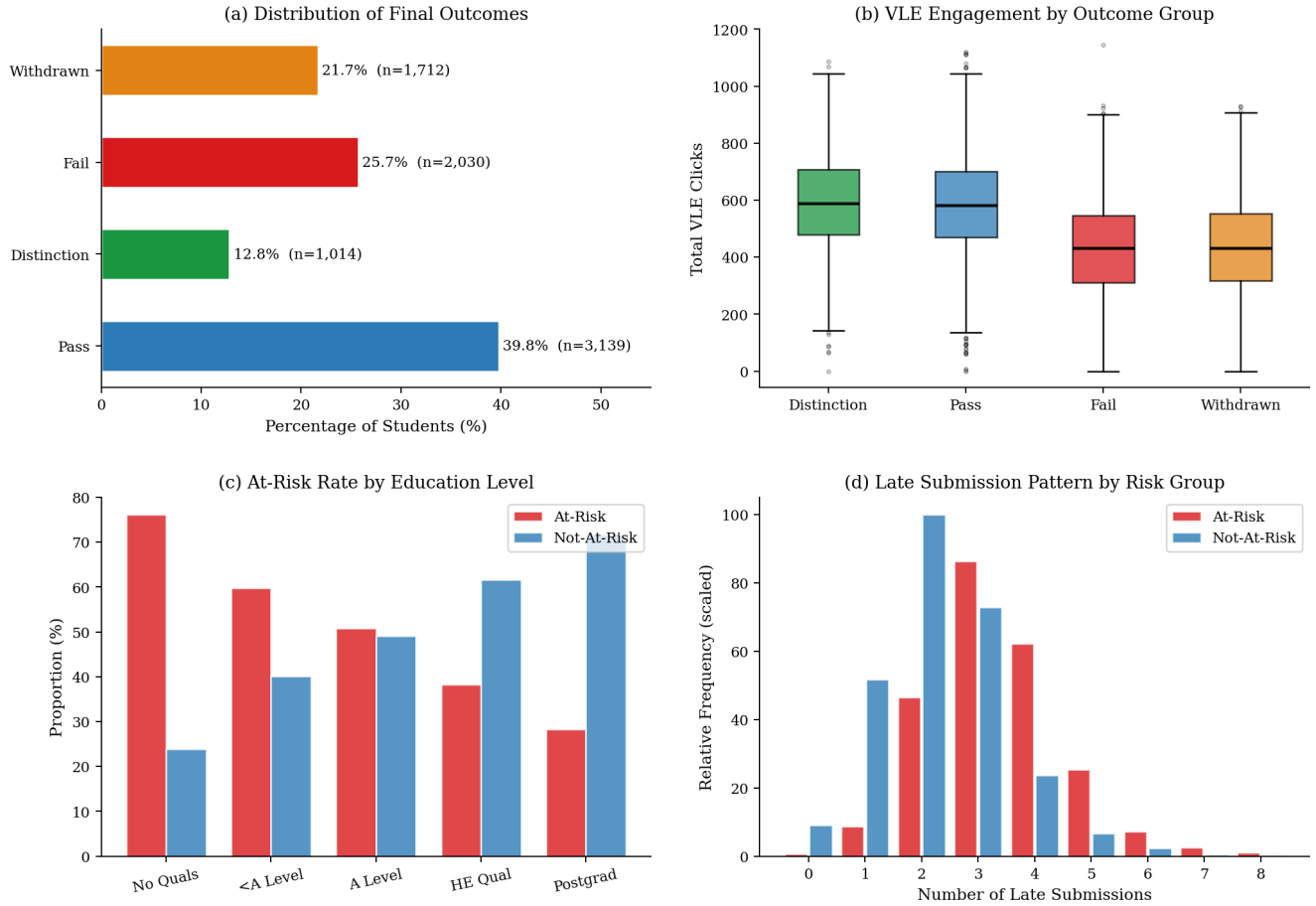



**Figure 1.** Dataset overview ($N = 7{,}895$): (a) final-outcome distribution; (b) total VLE clicks by outcome; (c) at-risk rate by highest education level; and (d) late submissions by binary risk group.

first assessment score.

## 5. DISCUSSION

### 5.1 Predictive Performance in Context

The AUC-ROC range in this study (0.724–0.786) is consistent with prior OULAD research. Adnan et al. [5] reported AUC-ROC values of 0.78–0.84 at 10–50% of course completion, while Waheed et al. [6] achieved values approaching 0.85 near the module midpoint using LSTM models. The lower ceiling here reflects the focus on features plausibly available within the first 25–30% of a module, with the aim of producing earlier, actionable signals.

The narrow AUC gap between Logistic Regression (0.786) and GBDT (0.782) suggests that much of the predictive signal in this feature set is captured without complex modelling. This is consistent with findings that ensemble gains may diminish for compact feature spaces [4, 3]. In deployment, the modest performance difference should be weighed against the greater interpretability of a Logistic Regression scorecard.

### 5.2 Feature Importance and Educational Implications

The importance of late submissions (0.304) is educationally meaningful: repeated missed deadlines may reflect disengagement, personal difficulties, or challenges with course demands. Submission timestamps are routinely logged by LMSs, making this signal broadly available. The finding aligns with prior work identifying assessment submission patterns as early indicators [5].

The importance of total VLE clicks (0.150) and first assessment score (0.135) highlights the value of combining engagement breadth with early academic performance. In distance education, VLE click counts provide an automatically logged, continuously updated proxy for engagement.

Demographic features contribute less than 5% of total importance, although education level is associated with risk in the dataset. Relying primarily on engagement rather than demographic proxies may reduce susceptibility to discriminatory feedback loops, but model outputs still require careful oversight [3].

### 5.3 Implications for Early-Warning System Design

The results suggest several design principles. First, GBDT is a viable basis for module-level alerts that are updated as engagement data accumulates. Second, systems should provide continuous risk probabilities, allowing tutors to prioritise cases rather than relying only on binary decisions. Third, a two-feature model using late submissions and first assessment score would capture approximately 44% of total Gini importance, supporting lightweight deployment. Fourth, pre-

**Table 2.** Five-fold stratified cross-validation results (mean ± standard deviation).

| Classifier | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| LR | 0.715 ± 0.005 | 0.729 ± 0.007 | 0.768 ± 0.006 | **0.734 ± 0.006** | **0.786 ± 0.003** |
| DT | 0.680 ± 0.005 | 0.699 ± 0.009 | 0.730 ± 0.012 | 0.699 ± 0.010 | 0.724 ± 0.010 |
| RF | 0.708 ± 0.007 | 0.723 ± 0.010 | 0.757 ± 0.009 | 0.728 ± 0.008 | 0.779 ± 0.005 |
| GBDT | 0.708 ± 0.005 | 0.720 ± 0.007 | **0.769 ± 0.006** | 0.729 ± 0.006 | 0.782 ± 0.003 |
| AdaBoost | **0.716 ± 0.008** | **0.731 ± 0.008** | 0.762 ± 0.008 | 0.732 ± 0.007 | 0.779 ± 0.002 |
| NB | 0.712 ± 0.006 | 0.727 ± 0.008 | 0.750 ± 0.009 | 0.727 ± 0.007 | 0.784 ± 0.003 |
| MLP | 0.708 ± 0.008 | 0.724 ± 0.009 | 0.753 ± 0.011 | 0.729 ± 0.008 | 0.778 ± 0.006 |

**Figure 2. Comparative Performance of Machine Learning Classifiers (5-Fold Cross-Validation)**

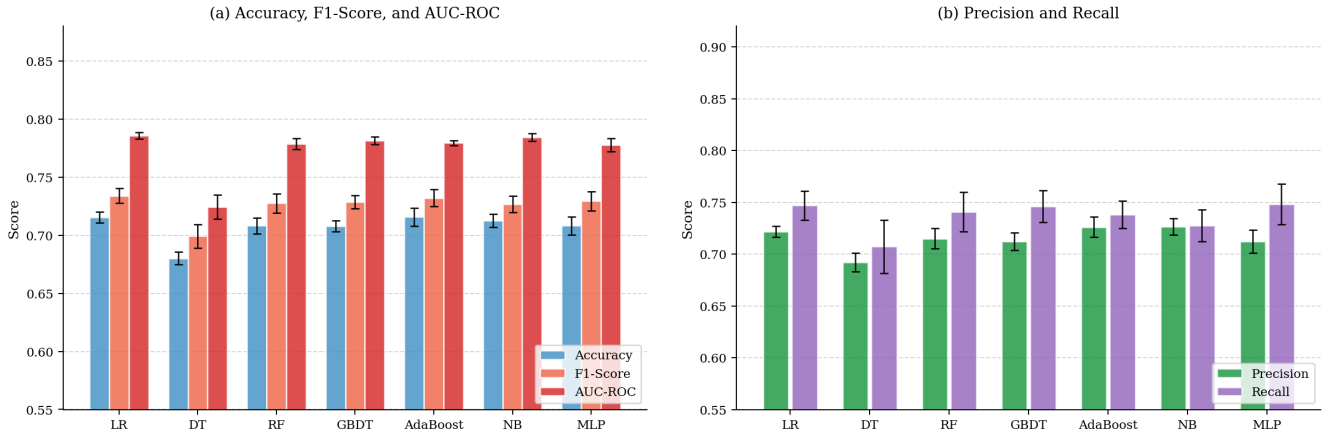



**Figure 2.** Comparative classifier performance under five-fold stratified cross-validation. Error bars indicate one standard deviation across folds: (a) accuracy, F1-score, and AUC-ROC; (b) precision and recall. LR = Logistic Regression, DT = Decision Tree, RF = Random Forest, GBDT = Gradient Boosting, AB = AdaBoost, NB = Naive Bayes, and MLP = Multilayer Perceptron.

**Figure 3. Receiver Operating Characteristic Curves — All Classifiers**

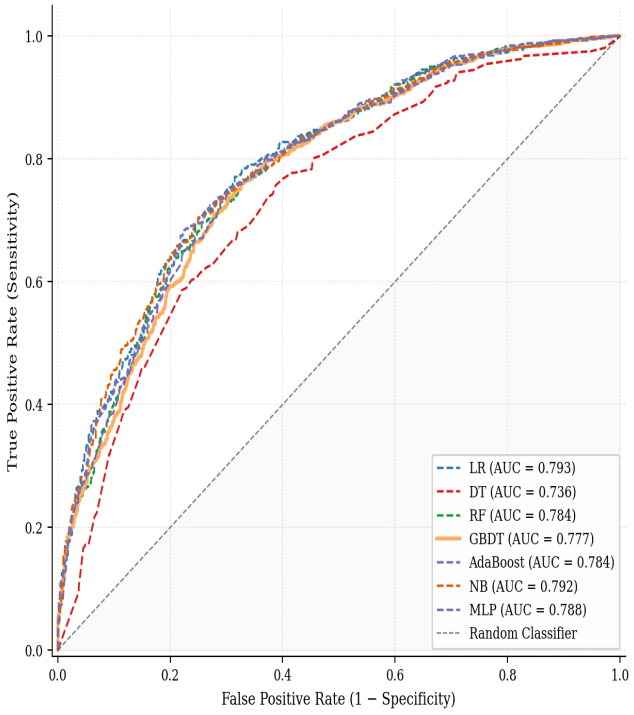



**Figure 3.** Receiver operating characteristic curves for the seven classifiers on the 20% holdout set ($n = 1{,}579$). The random-classifier diagonal (AUC = 0.50) is shown for reference.

dictions should be updated weekly and risk trajectories monitored over time, consistent with longitudinal approaches [6].

### 5.4 Limitations

The dataset was constructed synthetically to mirror OULAD statistics; validation on raw institutional data or the complete OULAD files is therefore needed. The binary target combines failure and withdrawal, although these outcomes may call for different interventions; future work should evaluate multiclass prediction. The analysis does not model temporal dynamics, and sequential models such as LSTMs or Transformers may extract additional information from clickstream sequences. Finally, feature importance captures association, not causation, so interventions should be guided by pedagogical judgement and student consent rather than automated predictions alone.

## 6. CONCLUSION

This study compared seven supervised classifiers for early identification of at-risk students in virtual learning environments using 7,895 records and 16 demographic, administrative, and engagement features. Logistic Regression achieved the highest cross-validation AUC-ROC (0.786), AdaBoost the highest accuracy (0.716), and GBDT the highest recall (0.769). Feature importance highlighted late submissions, total VLE clicks, and first assessment score as the strongest early-warning indicators.

The findings indicate that LMS data collected during the first

**Figure 4. Feature Importance and Confusion Matrix of the Best-Performing Model (GBDT)**

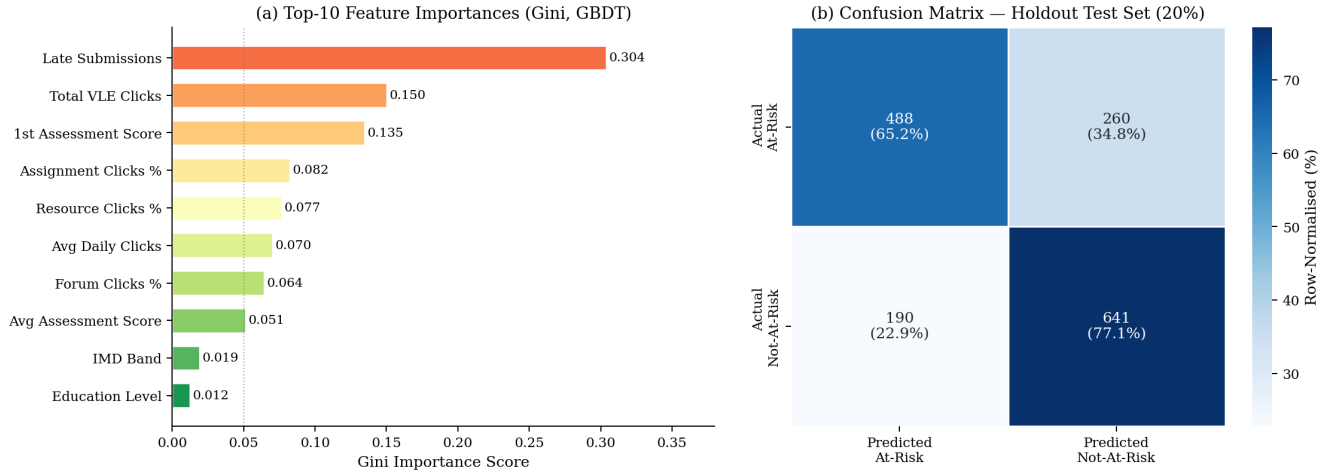



**Figure 4.** Gradient Boosting model outputs: (a) Gini-based importance scores for the top 10 features; (b) row-normalised confusion matrix on the 20% holdout set ($n = 1{,}579$), with counts and row percentages.

**Figure 5. Feature Correlation Structure and Engagement Profile by Risk Category**

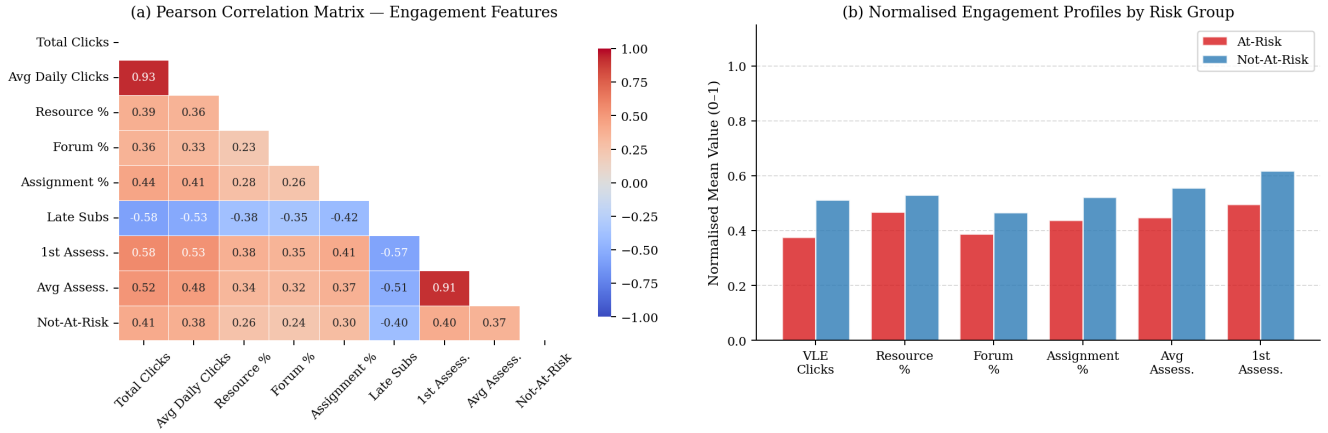



**Figure 5.** Feature correlation and engagement profiles: (a) lower-triangular Pearson correlation matrix for engagement features and the binary outcome (Not-At-Risk = 1); (b) min–max-normalised mean engagement profiles for at-risk and not-at-risk groups.

quarter of a module can provide useful signals for identifying students who may need support. Ensemble methods offer modest gains over simpler models, which should be balanced against interpretability. Early-warning systems should prioritise behavioural signals, provide human-reviewed recommendations, and avoid treating predictive associations as causal evidence. Future work should examine temporal and multiclass models and develop intervention protocols collaboratively with teaching staff.

## REFERENCES

[1] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open university learning analytics dataset. *Scientific Data*, 4:170171, 2017.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):39, 2019.

[3] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1):905–971, 2023.

[4] G. Feng, M. Fan, and Y. Chen. Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, 10:19558–19571, 2022.

[5] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, et al. Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, 9:7519–7539, 2021.

[6] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic. Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications*, 213:118868, 2023.

[7] A. Khan and S. K. Ghosh. Student performance analysis and prediction in classroom learning: A review of educa-

tional data mining studies. *Education and Information Technologies*, 26(1):205–240, 2021.

[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.