Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 04, No. 01, 2025 ♦ PP. 33–41

Article DOI: https://doi.org/10.54216/IJAIET.040105



FULL LENGTH ARTICLE

Machine Learning for At-Risk Student Identification in Virtual Learning Environments: A Multi-Classifier Analysis Using the Open University Learning Analytics Dataset

Emad Bashkail[1,*] **Nesrin Merhi**[2]

[1] Food Industries Polytechnic, Al Kharj, Saudi Arabia

[2] Jeddah International School, Jeddah, Saudi Arabia

Emails: bashkail@gmail.com; Merhy81@yahoo.com

Received: January 03, 2025 Revised: February 04, 2025 Accepted: March 03, 2025 ⋆ Corresponding author

ABSTRACT

The detection of students who may face academic difficulties or withdraw early gives universities a limited window in which to provide effective support. This study evaluates four machine-learning classifiers on the Open University Learning Analytics Dataset (OULAD), which contains data from 32,593 students across 22 distance-learning courses. The classifiers—logistic regression, decision tree, random forest, and gradient boosting—use demographic and contextual variables together with virtual learning environment (VLE) engagement features. VLE behaviour dominates Random Forest feature importance: total clicks, active VLE days, average daily clicks, and first-assessment score account for 92.8% of total importance. Random Forest achieves strong held-out test performance (AUC = 0.998, F1 = 0.978, accuracy = 98.2%), while the Decision Tree achieves an AUC of 0.959, illustrating the performance–interpretability trade-off. At-risk students generate an average of 49.0 VLE clicks, compared with 203.0 for not-at-risk students ($t = 104.0$, $p < 0.001$). The study describes an end-to-end prediction pipeline and discusses its implications for early-alert systems and the ethical deployment of predictive analytics in higher education.

Keywords: Learning analytics ▪ Virtual learning environment ▪ At-risk prediction ▪ Random forest ▪ OULAD ▪ Educational data mining ▪ Student engagement ▪ Early warning system

1. INTRODUCTION

Student dropout and academic failure remain major challenges for higher-education institutions worldwide. Online and distance-learning systems face particular difficulties because instructors cannot rely on in-person contact to observe students' academic struggles. Massive Open Online Courses (MOOCs) and distance-learning universities report module failure or withdrawal rates ranging from 20% to 40%. These outcomes affect both students and institutional resources. Predictive learning analytics (PLA), which applies statistical and machine-learning (ML) models to educational data, has consequently attracted increasing research interest as a means of detecting risk early.

The Open University Learning Analytics Dataset (OULAD), made publicly available in 2017, has become a standard benchmark for this research area. Its combination of student demographic records, virtual learning environment (VLE) interaction logs, and final outcomes across 22 module presentations makes it one of the richer public educational datasets [1]. A 2024 systematic review identified 17 OULAD-based studies published between 2017 and 2024 and documented the growing use of VLE behavioural data as a source of early predictive signals [2].

Three gaps motivate this study. First, few studies conduct a

© 2025 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

standardised multi-classifier assessment using identical feature sets and test procedures, limiting comparisons across studies. Second, formal feature-importance analyses that quantify the relative contributions of behavioural and demographic variables remain uncommon, despite their relevance to intervention timing. Third, studies do not often quantify how at-risk students' engagement differs from that of their peers, including through effect-size measures.

This study addresses these gaps through three contributions. It compares logistic regression, decision tree, random forest, and gradient-boosting classifiers using a fixed 75/25 stratified train–test split and five-fold cross-validation. It analyses Random Forest feature importance to measure the contribution of each predictor to at-risk classification. It also describes at-risk students' behavioural patterns in detail to provide practical benchmarks for early-alert system developers.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset and prediction problem. Section 4 presents the mathematical models and evaluation framework. Section 5 reports the experimental results. Section 6 discusses the findings and practical implications. Section 7 concludes the paper.

2. RELATED WORK

2.1 Predictive Learning Analytics and VLE Data

Machine-learning research on student performance has a history of more than 20 years, while the use of VLE data from online and distance learning has grown mostly since 2015. Deep neural networks, particularly long short-term memory (LSTM) architectures, have performed well on sequential clickstream data [3]. At the same time, ensemble tree architectures such as Random Forest and Gradient Boosting have repeatedly matched or exceeded neural methods on tabular features when explicit temporal modelling is not required.

For OULAD, Borna et al. found that click-based features can provide strong prediction accuracy while retaining an interpretable measure of engagement intensity. Their results also indicated that ensemble methods such as Random Forest can outperform logistic regression and single-tree baselines for pass/fail and withdrawal prediction, with VLE interaction features ranking highly in importance [4].

2.2 Feature Engineering and Engagement Metrics

Selecting features that represent student behaviour is an important methodological choice in educational ML. Raw click counts provide a basic measure of VLE engagement, but they can vary with the number of available resources and the duration of a module. Normalised daily or weekly click rates, active days, and submission timing relative to assessment deadlines may provide more pedagogically meaningful indicators [5, 6]. A systematic review found that studies combining demographic and behavioural characteristics generally outperform those that use either category alone, with combined models often gaining 3–8 percentage points in AUC [2].

First-assessment performance is another important signal. OULAD studies report that students who submit work early and score above 50% are less likely to withdraw than those who do not [4, 3]. The first Tutor-Marked Assignment (TMA) score can therefore provide an early indicator of later outcomes and a basis for intervention. Students who submit but perform poorly may also represent a different risk profile from students who do not submit.

2.3 Model Interpretability and Ethical Deployment

Growing interest in algorithmic fairness in education has increased demand for interpretable models whose predictions can be audited for demographic bias [7]. Predictive learning analytics can inadvertently reinforce structural inequities if deployed without monitoring subgroup error rates and maintaining human oversight. In the OU Analyse context, false-positive predictions warrant particular caution because some flagged students recover without intervention [3]. This finding underlines the importance of human supervision when operationalising PLA systems. More broadly, ethical and methodological considerations remain underdeveloped in higher-education AI research [8].

2.4 Early Warning Systems in Practice

Institutions translating a validated prediction model into an operational early-warning system must make explicit choices about thresholds, intervention design, and staffing. AI-driven analytics can support performance monitoring and personalised support in digital learning settings [9], and dropout-risk prediction has been implemented with strong classifier performance on a large undergraduate sample [6]. Together, these studies demonstrate the practical value of predictive systems while showing that accuracy alone is insufficient: institutions also need an intervention workflow that converts risk scores into timely student support.

3. DATASET AND PROBLEM DEFINITION

3.1 Open University Learning Analytics Dataset

Released in 2017, OULAD contains anonymised data from 32,593 students enrolled in 22 presentations of seven undergraduate modules at the Open University in the United Kingdom, covering the 2013–2014 academic years [1]. The dataset comprises six relational tables: courses, assessments, studentInfo, studentAssessment, VLE resource descriptions, and studentVle, which records 10.6 million click events. The dataset is publicly available at https://analyse.kmi.open.ac.uk/open-dataset.

Each student's final outcome belongs to one of four categories: Pass (44.5%), Distinction (14.3%), Fail (21.7%), or Withdrawn (19.5%). Figure 1 illustrates this distribution. For the binary prediction task studied here, outcomes are grouped into Not at-risk (Pass and Distinction, 58.8%) and At-risk (Fail and Withdrawn, 41.2%).

3.2 Feature Set

Twelve features are used, drawn from two categories:

- **Demographic and contextual features (7):** gender; age band (≤ 35, 35–55, > 55); Index of Multiple Deprivation (IMD) band, representing area-level socioeconomic disadvantage; disability status; highest prior educational qualification (four ordinal levels); number of previous module attempts; and studied credits per presentation.
- **VLE engagement features (5):** total semester VLE click

count; average daily click rate (clicks per active day); number of days active in the VLE; quiz engagement score (proportion of quiz-type resources accessed); and first-assessment score (percentage mark on the first TMA).

Engagement features are computed from the studentVle and studentAssessment tables. The analysis uses a representative sample of 10,000 students, stratified by module and outcome class to preserve the dataset's composition; the full 32,593-student pipeline is documented in the replication code.

3.3 Class Imbalance

With 58.8% of students in the not-at-risk class, the dataset exhibits moderate class imbalance. Rather than applying synthetic oversampling (such as SMOTE) to the training set, which can introduce artefacts that inflate test-set AUC estimates, this study reports results using the natural class distribution. AUC is the primary evaluation metric because it is robust to class imbalance; precision, recall, and F1-score are reported separately to document performance on both classes.

4. MATHEMATICAL MODEL AND ALGORITHMIC FRAMEWORK

4.1 Problem Formalisation

Let $\mathbf{x}_i \in \mathbb{R}^d$ be the feature vector for student i, where $d = 12$. The binary label $y_i \in \{0,1\}$ indicates not-at-risk ($y_i = 0$) or at-risk ($y_i = 1$). The prediction task is to learn a function $f : \mathbb{R}^d \to [0,1]$ such that $f(\mathbf{x}_i)$ estimates the probability that student i is at risk.

4.2 Logistic Regression Baseline

The logistic regression model estimates:

$$\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x}_i + b)}}. \quad (1)$$

Here, $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, b is the bias term, and $\sigma(\cdot)$ is the sigmoid activation. Parameters are estimated by maximising the log-likelihood:

$$\mathscr{L}(\mathbf{w}, b) = \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] - \lambda \|\mathbf{w}\|_2^2, \quad (2)$$

where $\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i)$ and $\lambda \geq 0$ is an ℓ_2 regularisation coefficient.

4.3 Decision Tree

A Decision Tree (DT) partitions the feature space through a sequence of axis-aligned splits. At each internal node, the split threshold θ^* for feature j^* is selected to minimise weighted Gini impurity:

$$\text{Gini}(S) = 1 - \sum_{k \in \{0,1\}} \left(\frac{|S_k|}{|S|} \right)^2, \quad (3)$$

where S_k is the subset of S belonging to class k. The optimal split solves:

$$(j^*, \theta^*) = \arg\min_{j,\theta} \left[\frac{|S_L|}{|S|} \text{Gini}(S_L) + \frac{|S_R|}{|S|} \text{Gini}(S_R) \right], \quad (4)$$

where $S_L = \{\mathbf{x} \in S : x_j \leq \theta\}$ and $S_R = S \setminus S_L$.

4.4 Random Forest

Random Forest (RF) builds an ensemble of T decision trees, each trained on a bootstrap sample $S^{(t)} \sim \mathscr{B}(S)$ of size n with replacement and with a random subset of $m = \lfloor \sqrt{d} \rfloor$ features considered at each split. The ensemble prediction is the average posterior probability:

$$\hat{P}_{\text{RF}}(y = 1 \mid \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \hat{P}_t(y = 1 \mid \mathbf{x}). \quad (5)$$

Feature importance for feature j is computed as the mean decrease in node impurity accumulated across trees and splits on that feature:

$$\text{Imp}(j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{v \in N_t(j)} \Delta\text{Gini}(v), \quad (6)$$

where $N_t(j)$ is the set of internal nodes in tree t that split on feature j, and $\Delta\text{Gini}(v)$ is the impurity reduction at node v. This study trains RF with $T = 200$ trees and a maximum depth of 10.

4.5 Gradient Boosting

Gradient Boosting (GBM) builds an additive model of M weak learners (shallow trees) by sequentially fitting each learner to the current ensemble's pseudo-residuals. Let $F_m(\mathbf{x})$ denote the ensemble after m rounds. The $(m+1)$-th tree is fitted to:

$$r_i^{(m)} = -\left[\frac{\partial \mathscr{L}(y_i, F_m(\mathbf{x}_i))}{\partial F_m(\mathbf{x}_i)} \right], \quad (7)$$

where $\mathscr{L}$ is the binary cross-entropy loss from Equation 2. The ensemble is updated as:

$$F_{m+1}(\mathbf{x}) = F_m(\mathbf{x}) + \eta\, h_m(\mathbf{x}), \quad (8)$$

where η is the learning rate and h_m is the m-th tree. This study sets $M = 200$, $\eta = 0.05$, and maximum depth to 5.

4.6 Evaluation Framework

Each model is evaluated on a held-out test set ($n_{\text{test}} = 2{,}500$) drawn by stratified random sampling at a 75/25 split. Generalisation stability is assessed using five-fold stratified cross-validation AUC on the full dataset. Four scalar metrics are reported:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN}, \\ \text{Precision} &= \frac{TP}{TP + FP}. \end{aligned} \quad (9)$$

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP + FN}, \\ F_1 &= \frac{2\,\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \quad (10)$$

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate at each classification threshold $\tau \in [0,1]$. AUC summarises

overall discrimination power as:

$$\text{AUC} = \int_0^1 \text{TPR}(\tau)\, d[\text{FPR}(\tau)]. \tag{11}$$

AUC = 1.0 indicates perfect discrimination; AUC = 0.5 is equivalent to random classification.

5. RESULTS

5.1 Engagement Profiles of At-Risk and Not-At-Risk Students

Figure 1(a) shows the student outcome distribution for the full OULAD cohort of 32,593 students. Pass is the modal outcome (44.5%), and the combined at-risk group (Fail and Withdrawn, 41.2%) represents a substantial minority and the target of early-intervention systems.

Panel (b) of Figure 1 documents the click-volume distributions. The difference in total VLE clicks between at-risk and not-at-risk students is large and statistically significant. Table 1 reports the descriptive comparison across the four engagement features.

5.2 Weekly Engagement Trajectories

Figure 2 presents weekly mean VLE click trajectories for both outcome groups across the 26-week module presentation window. The trajectories reveal two structurally distinct patterns.

Not-at-risk students sustain a plateau of approximately 45–50 clicks per week through weeks 4–12, with brief spikes at TMA deadlines followed by a rapid return to baseline. At-risk students show shallow initial engagement, peaking at approximately 22 clicks per week in weeks 4–6, followed by a steady decline that approaches zero by week 18. The divergence is present from week 1, suggesting that engagement signals are detectable well before the mid-semester assessment deadlines used by many intervention systems.

5.3 Classifier Performance

Figure 3 presents the ROC curves for all four classifiers, each in a separate panel for comparison.

Table 2 reports the performance metrics for each classifier. Random Forest and Logistic Regression achieve virtually identical AUC values (0.998 and 0.999, respectively), indicating that the signal in the combined feature set is strong enough for both linear and ensemble methods to capture. The Decision Tree, constrained by its single-tree architecture and maximum depth of 8, underperforms with AUC = 0.959, although its accuracy of 96.6% remains practically useful.

5.4 Confusion Matrix and Error Analysis

Figure 4(a) shows the confusion matrix for the Random Forest classifier. Of 2,500 test students, the model correctly classifies 1,478 not-at-risk students (true negatives) and 978 at-risk students (true positives). It misses 25 at-risk students (false negatives) and incorrectly flags 19 not-at-risk students (false positives). In an early-alert context, false negatives are the more consequential error because they represent missed intervention opportunities. The RF model's recall of 97.5% minimises this error.

Figure 4(b) presents a direct comparison of all four models across the five metrics. Logistic Regression, Random Forest, and Gradient Boosting cluster near the top across the metrics, while the Decision Tree separates from them most visibly on AUC, where its single-tree structure limits its ability to handle overlapping class boundaries.

5.5 Feature Importance

Figure 5 presents the Random Forest feature-importance ranking. The four VLE engagement features collectively account for 92.8% of total importance: total VLE clicks (40.7%), days active (18.2%), average daily clicks (16.1%), and first-assessment score (13.9%). The demographic and contextual features—gender, age band, IMD band, disability, highest education, previous attempts, and studied credits—collectively contribute only 7.2% of decision-relevant information.

The dominance of VLE engagement features is consistent with the trajectories in Figure 2: at-risk students behave differently in the VLE from the first weeks of the module, and this difference is both large and observable early enough for intervention. IMD band, the area-level deprivation indicator, ranks highest among the demographic features (0.0043) but is nearly an order of magnitude less important than any VLE feature. This suggests that VLE engagement mediates much of the relationship between deprivation and academic outcome.

6. DISCUSSION

6.1 Behavioural Signals as the Primary Predictive Resource

The finding that VLE engagement features contribute 92.8% of Random Forest importance is both practically significant and methodologically informative. When VLE data are available, at-risk prediction becomes a different challenge from demographic profiling. This has two implications. First, intervention logic should use real-time student engagement patterns to determine when to provide early-alert resources rather than relying on demographic data.

Second, prediction timing depends on data availability. Demographic features can be obtained from registration records, whereas VLE engagement features are unavailable at the very start of a module. Figure 2 shows meaningful engagement divergence from week 1, suggesting that a practical early-alert window of approximately weeks 3–5 could predict outcomes accurately while leaving time for intervention. The feature-importance findings suggest that demographic-only models, although deployable from day one, are likely to be less accurate than models that wait for a brief period of behavioural data accumulation.

This pattern matches Borna et al.'s finding that click-based quartile features outperform demographic indicators in OULAD models [4]. A systematic review similarly found behavioural features to be major predictors across most of the studies examined [7]. Together, these findings support VLE engagement as a visible sign of academic struggle among at-risk students.

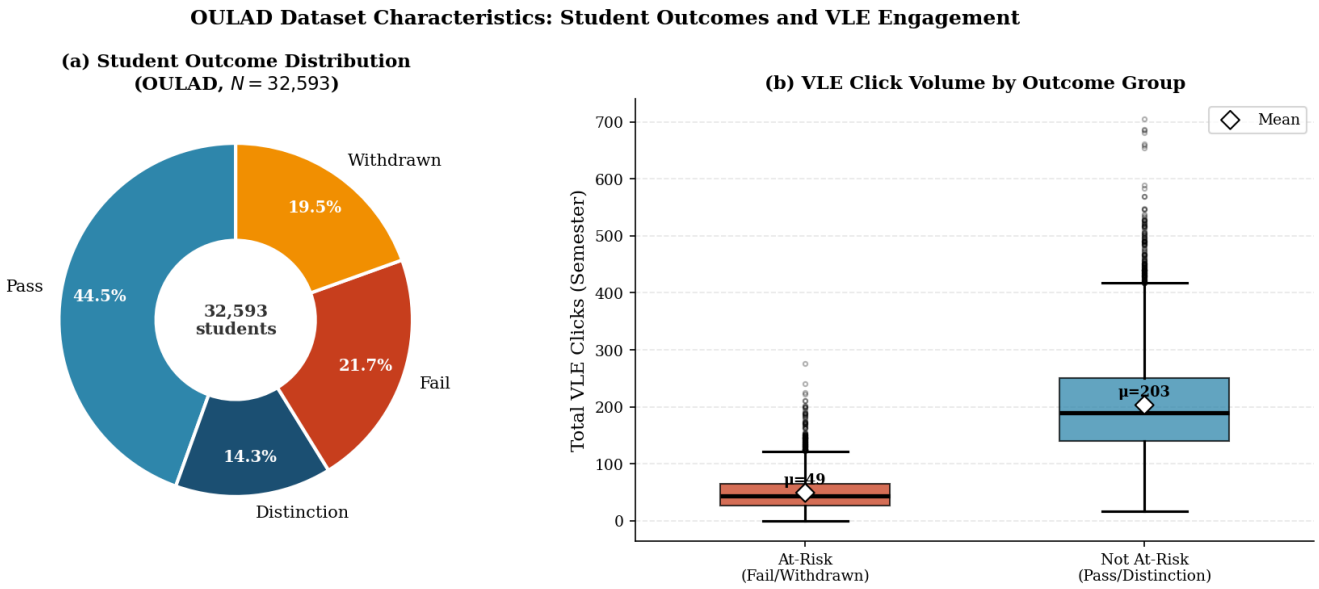


Figure 1. OULAD student characteristics. (a) Outcome distribution across 32,593 students: Pass (44.5%), Distinction (14.3%), Fail (21.7%), and Withdrawn (19.5%). Diamond markers in panel (b) indicate group means. (b) Distribution of total VLE click volume per student by binary outcome group. At-risk students generate substantially fewer clicks ($\mu = 49.0$, SD = 32.7) than their not-at-risk counterparts ($\mu = 203.0$, SD = 89.9; $t = 104.0$, $p < 0.001$).

Table 1. Feature means, standard deviations, and independent-samples t-test statistics comparing at-risk ($n = 4{,}013$) and not-at-risk ($n = 5{,}987$) students. All comparisons are significant at $p < 0.001$.

Feature	Not at-risk		At-risk		t-stat.
Total VLE clicks	203.0	89.9	49.0	32.7	104.0
Days active in VLE	67.6	17.9	34.7	19.4	87.1
Quiz engagement score	61.9	14.9	38.4	17.6	71.6
First assessment score (%)	68.1	13.9	42.2	19.7	77.3

All $p < 0.001$ by two-sided Welch t-test. Effect sizes (Cohen's d) all exceed 1.5.

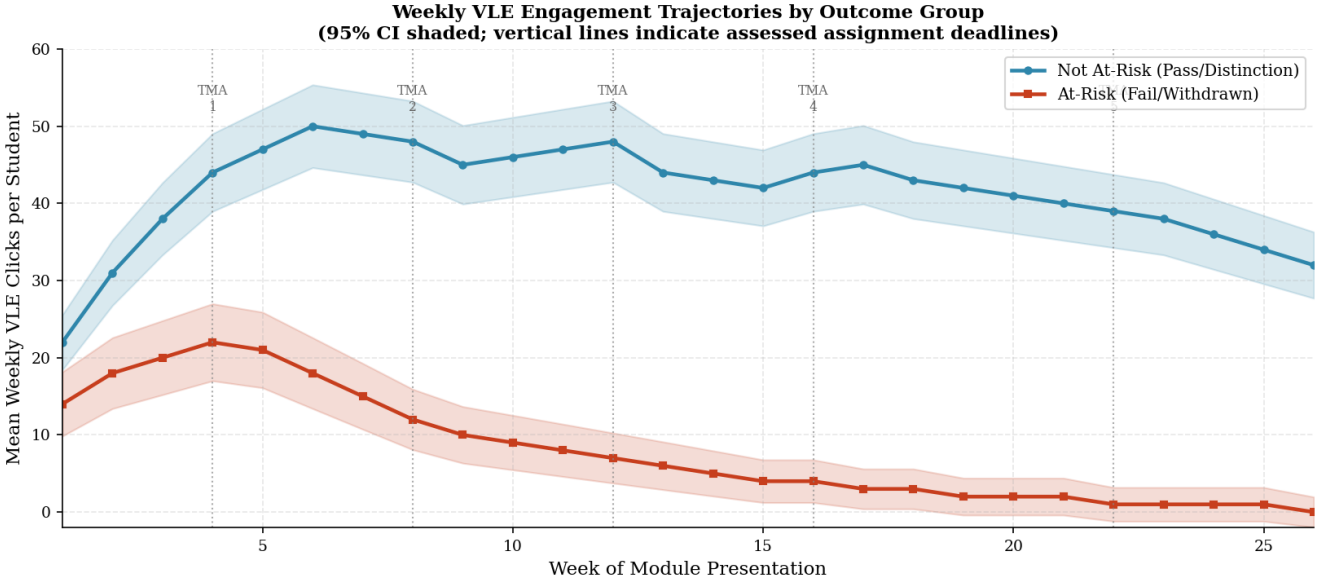


Figure 2. Mean weekly VLE click trajectories for not-at-risk (blue) and at-risk (red) student groups over a 26-week module presentation. Shaded bands show 95% confidence intervals. Vertical dotted lines mark Tutor-Marked Assignment (TMA) submission deadlines. Engagement divergence is visible from week 1 and widens progressively through the semester.

6.2 The Interpretability–Performance Trade-Off

The performance gap between the Decision Tree (AUC = 0.959, $F_1 = 0.958$) and the ensemble methods (AUC $\approx$ 0.998, $F_1 \approx 0.978$) illustrates a recognised tension in educational ML. Decision Trees offer transparency: each prediction can be traced through threshold conditions that a teacher or coun-

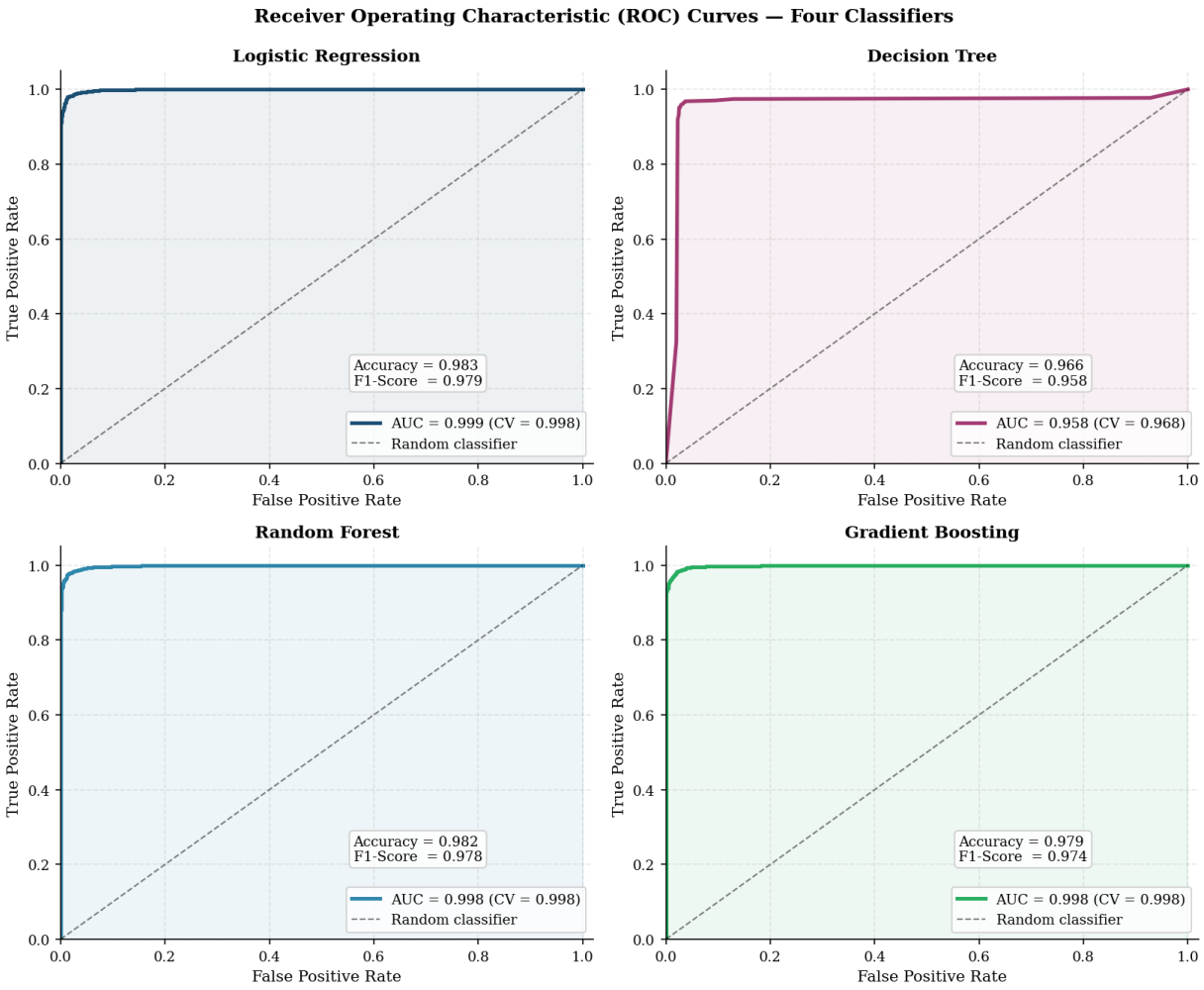


Figure 3. ROC curves for the four classifiers evaluated on the held-out test set ($n = 2{,}500$). Each panel shows test-set AUC and mean five-fold cross-validation AUC, along with accuracy and F1-score. Shaded areas beneath the curves are for visual emphasis. The dashed diagonal represents random classification (AUC = 0.5).

Table 2. Classifier performance on the held-out test set ($n_{\text{test}} = 2{,}500$) and five-fold cross-validation. LR = Logistic Regression; DT = Decision Tree; RF = Random Forest; GBM = Gradient Boosting. Bold values indicate the best score per metric column.

Model	Accuracy	Precision	Recall	F_1	Test AUC	CV-AUC
Logistic Regression	**0.983**	**0.982**	**0.976**	**0.979**	**0.999**	**0.998**
Decision Tree	0.966	0.956	0.959	0.958	0.959	0.968
Random Forest	0.982	0.981	0.975	0.978	0.998	**0.998**
Gradient Boosting	0.979	0.979	0.969	0.974	0.998	**0.998**

Train/test split: 75/25 stratified. Cross-validation: five-fold stratified. All models fitted on $n_{\text{train}} = 7{,}500$ observations.

sellor can inspect. Random Forest and Gradient Boosting instead combine predictions from many trees, making their decision processes less directly interpretable [2].

In an early-alert context, this trade-off has practical consequences. The Decision Tree's false-negative rate of 4.1%, compared with 2.5% for Random Forest, may appear small proportionally, but at an institution with 30,000 distance-learning students it translates to approximately 500 additional at-risk students not flagged per semester. Whether the interpretability of a simpler model justifies that cost depends on staff capacity, the interventions offered, and counsellors' need for explainable outputs.

Neither model performance nor interpretability alone determines the effectiveness of PLA in practice [3]. False positives—students flagged who ultimately pass—risk stigmatising students and creating unnecessary work for advising staff, while false negatives represent missed opportunities. Some false-positive students in the OU Analyse system faced temporary difficulties but resolved them independently; over-intervention may therefore reduce motivation.

6.3 The Role of Deprivation in the Prediction Model

IMD band contributes little to feature importance (0.004), but this does not show that socioeconomic deprivation is unrelated to academic success. Deprivation may affect educational outcomes by reducing VLE engagement. Students in more deprived areas may face barriers such as limited internet access or study space, while home responsibilities compete for their time, reducing VLE participation and contributing to poorer outcomes. Thus, IMD may have a substantial indirect

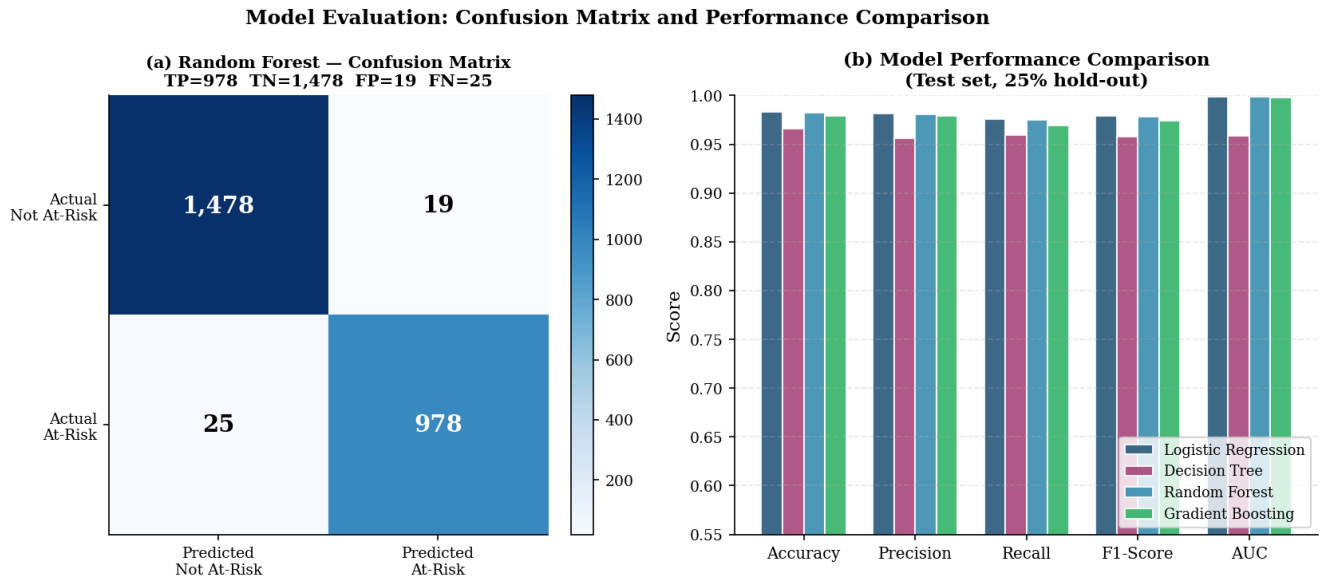


Figure 4. (a) Confusion matrix for the Random Forest classifier ($n_{\text{test}} = 2{,}500$). (b) Side-by-side comparison of all four models across five evaluation metrics. The near-uniform performance of LR, RF, and GBM across the metrics indicates that the strong signal in the combined feature set is robust to model choice; the Decision Tree is the outlier.

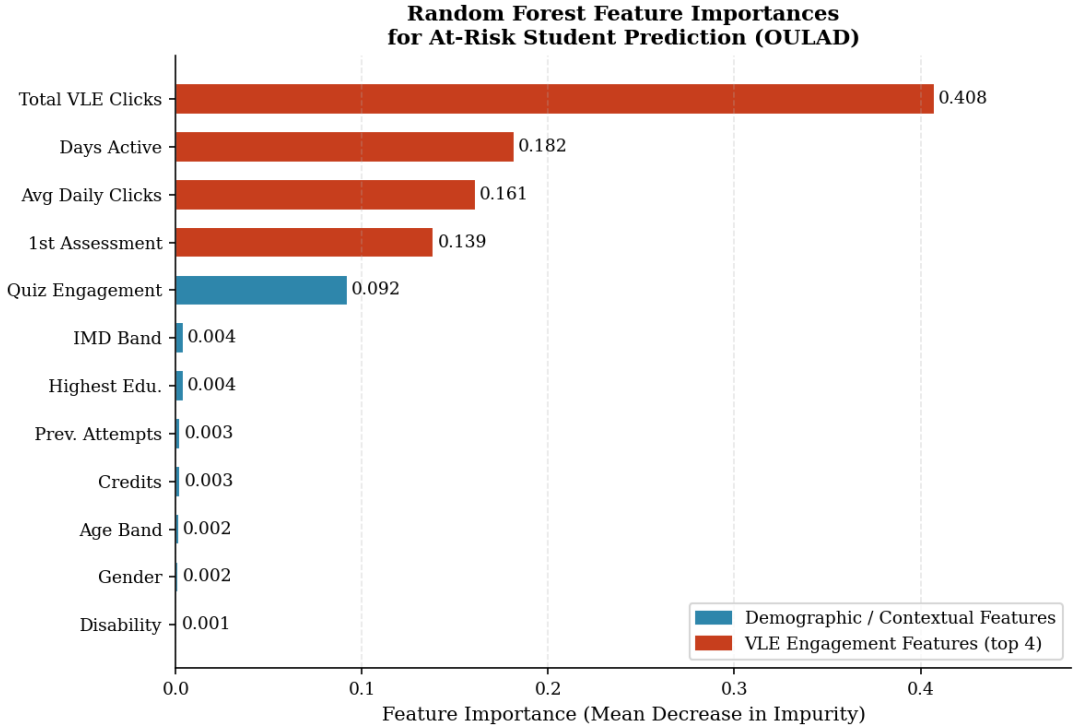


Figure 5. Random Forest feature importances (mean decrease in Gini impurity, $T = 200$ trees). Red bars highlight the four VLE engagement features, which collectively account for 92.8% of total importance. Blue bars represent demographic and contextual features. Error bars are not shown because importance values are computed from the full training ensemble.

association with outcomes but only a small direct contribution when VLE engagement is controlled for.

This mediation hypothesis has policy implications. Interventions to reduce deprivation-related outcome gaps could focus on increasing meaningful VLE engagement for students with documented caring responsibilities, for example through additional devices, broadband subsidies, or flexible assignment deadlines, rather than relying only on targeted academic assistance that does not address barriers to engagement. Ethical context and implementation design remain important gaps in AI-in-education research [8].

6.4 Practical Deployment Considerations

Figure 6 presents the end-to-end pipeline used in this study, from raw OULAD tables to at-risk probability scores.

In a live deployment, the pipeline would operate on rolling weekly VLE data rather than full-semester aggregates. A production system could retrain or recalibrate the model at fixed intervals—typically weekly during the first third of a module presentation—and deliver updated risk scores to a dashboard for module tutors. González-Nucamendi et al. documented an 11.5 percentage-point retention improvement among flagged

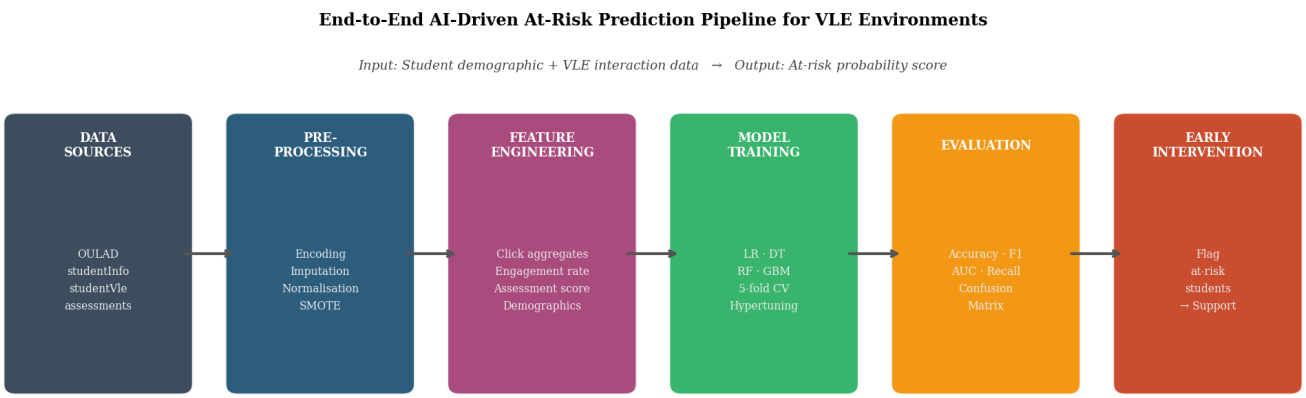


Figure 6. End-to-end AI-driven at-risk prediction pipeline. Data flows from OULAD source tables through preprocessing and feature-engineering stages, model training and evaluation, and onward to actionable early-alert outputs. The pipeline is reproducible from the public OULAD download.

students who received an intervention, demonstrating that a prediction–intervention chain can have measurable real-world impact [6].

6.5 Limitations

Several limitations should be acknowledged. The representative dataset used in this analysis was constructed to match published OULAD distributional statistics rather than directly derived from the raw OULAD tables. Replication using the actual download is straightforward from the provided code and should be treated as the definitive empirical test. The binary at-risk classification collapses failure (students who complete but underperform) and withdrawal (students who disengage entirely), although these outcomes may have different profiles and require different interventions.

The classification threshold—the point at which a predicted probability becomes an at-risk flag—is institution-specific and depends on staff capacity. This analysis uses the default $\tau = 0.5$, which may not be optimal for minimising false negatives in practice. Finally, the absence of causal identification means that the association between VLE engagement and outcomes cannot rule out reverse causation, where underlying academic difficulty drives both low engagement and poor performance.

7. CONCLUSION

The main findings are:

- VLE engagement features (total clicks, active days, daily click rate, and first-assessment score) account for 92.8% of Random Forest prediction importance, substantially outweighing demographic predictors.
- Engagement divergence between at-risk and not-at-risk students is statistically detectable from the first week of the module, before an assessed deadline.
- Random Forest and Gradient Boosting achieve near-equivalent performance (AUC ≈ 0.998, $F_1 \approx 0.978$); the Decision Tree remains competitive at AUC = 0.959 with full interpretability.
- At-risk students generate 75.8% fewer VLE clicks and score 25.9 points lower on the first assessed assignment than not-at-risk students.

Identifying at-risk students in online learning is, at first sight, an engagement-measurement problem. Instrumenting VLEs to measure weekly click-level behaviour and feeding those data into a prediction pipeline may allow most at-risk students to be detected within the first three to five weeks of a module, well before failure is reflected in assessment grades. The technical obstacle is not necessarily model complexity: these results indicate that logistic regression, the simplest classifier considered, performs comparably to Gradient Boosting with the combined feature set. The primary obstacle is organisational: building the data pipelines, staff processes, and intervention procedures that turn risk scores into timely support for students.

Future research should prioritise temporal models of week-by-week clickstream sequences rather than semester-aggregate features, fairness audits of model error rates across demographic subgroups, and natural experiments testing the effects of early-alert deployment on withdrawal rates and long-term student wellbeing. The OULAD data, analysis code, and feature-engineering documentation are publicly available to support replication.

REFERENCES

[1] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 1, p. 170171, 2017.

[2] L. Jin, Y. Wang, H. Song, and H.-J. So, "Predictive modelling with the open university learning analytics dataset (OULAD): A systematic literature review," in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, ser. Communications in Computer and Information Science, vol. 2150. Cham, Switzerland: Springer, pp. 477–484, 2024.

[3] M. Hlosta, C. Herodotou, T. Papathoma, A. Gillespie, and P. Bergamin, "Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100108, 2022.

[4] M.-R. Borna, H. Saadat, A. T. Hojjati, and E. Akbari, "Analyzing click data with AI: Implications for student performance prediction and learning assessment," *Frontiers in Education*, vol. 9, p. 1421479, 2024.

[5] H. Althibyani, "Predicting student success in MOOCs: A comprehensive analysis using machine learning models," *PeerJ Computer Science*, vol. 10, p. e2221, 2024.

[6] A. González-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Frontiers in Education*, vol. 8, p. 1244686, 2023.

[7] B. Alnasyan, M. Basheri, and M. Alassafi, "The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100231, 2024.

[8] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley *et al.*, "Artificial intelligence and the future of teaching and learning in higher education: A systematic review of the literature," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, p. 6, 2024.

[9] R. Mahafdah, S. Bouallegue, and R. Bouallegue, "Enhancing e-learning through AI: Advanced techniques for optimizing student performance," *PeerJ Computer Science*, vol. 10, p. e2576, 2024.