Journal of Informatics and Vocational Education (JOIVE)
Vol.9, No.2, July 2026, pp. 55~66
ISSN: 2746-7813

Machine Learning-Based Early Warning for Student Dropout: Evidence from LMS Behavioral Engagement Patterns in Online Higher Education

Rhezwan Dhaifullah Romdhoni[1*], Nuur Wachid Abdul Majid[2], Anggita Fitri Permatasari[3]
[1, 2, 3] Department of Education Systems and Information Technology, Universitas Pendidikan Indonesia

Article Info

Article history:

Received Apr 27, 2026
Revised May 06, 2026
Accepted Jul 1, 2026

Corresponding Author:

Nuur Wachid Abdul Majid,
Universitas Pendidikan Indonesia, Indonesia
Email: nuurwachid@upi.edu

ABSTRACT

Student dropout in online higher education remains critically high, far exceeding face-to-face rates, yet declining behavioral activity in Learning Management Systems (LMS) offers key signals for early intervention. To identify robust predictors and model suitability for Early Warning Systems (EWS), this study presents a comparative analysis of machine learning for dropout prediction using clickstream data from the Open University Learning Analytics Dataset (OULAD), covering 32,593 students across seven undergraduate modules. Three supervised algorithms with Logistic Regression, Random Forest, and Support Vector Machine (SVM), were trained on 13 engineered features combining behavioral and demographic attributes from the Virtual Learning Environment (VLE), with Recall prioritized to minimize missed at-risk students. Results demonstrate that all models achieved strong discriminatory performance with AUC-ROC > 0.93; specifically, SVM provided the highest EWS fit with recall of 0.903, missing only 196 of 2,031 withdrawals (9.7%), while Random Forest attained the best overall accuracy (0.866) and AUC-ROC (0.940). Feature importance analysis further revealed that VLE behavior accounted for 85.0% of predictive power, with Activity Span emerging as the dominant predictor at 41.3%. Cross-module validation confirmed temporal engagement consistency as a robust, generalizable dropout signal. Therefore, these findings provide practical guidance for implementing data-driven EWS in online learning by prioritizing behavioral span metrics over static demographics.

Keywords: Student dropout prediction, Learning analytics, Machine learning, LMS clickstream data, Early warning system.

1. INTRODUCTION

Student dropout in higher education represents a persistent and costly challenge across global education systems, with consequences that extend beyond academic failure to impose substantial personal, social, and economic burdens on students, institutions, and society at large (de Oliveira et al., 2021). In OECD countries, dropout rates have been estimated at approximately 30%, with some contexts reporting rates as high as 64.5% (Guzmán et al., 2021), while a meta-analysis covering multiple institutions reported a pooled non-continuation rate of 17.9%, ranging from 5.9% to 43.6% across settings (Leow et al., 2025). This challenge is especially acute in online and distance education, where structural barriers including isolation, reduced institutional support, and the self-regulatory demands of digital learning environments drive non-completion rates substantially higher than in face-to-face programs, rendering dropout a critical equity issue for both students and institutions (Kurulgan, 2024). Understanding and anticipating dropout risk in these contexts therefore requires systematic analytical approaches capable of leveraging the behavioral data that online learning environments inherently produce.

The problem is especially pronounced in online higher education environments. Studies consistently show that distance learning institutions record significantly higher dropout rates compared to campus-based

counterparts a South Korean distance university reported a withdrawal rate of 16.4% versus 6.0% at offline universities (Seo et al., 2024), while a longitudinal study of a fully online university in Catalonia found that 61% of a cohort followed a dropout trajectory over nine years compared to approximately 23% at traditional institutions (Sánchez-Gelabert, 2020). Reviews of online graduation rates indicate that completion often falls between 0.5% and 20% in some distance education contexts (Rotar, 2022). The factors driving this vulnerability are multidimensional, encompassing academic underperformance, motivation deficits, socioeconomic pressures, and the distinctive challenges of digital learning environments, including course quality issues, isolation, and insufficient institutional support (Kocsis & Molnár, 2025; Rahmani et al., 2024). Students in online learning environments exhibit significantly lower motivation and self-regulation compared to in-person settings, including limited interaction and reduced engagement, behavioral dimensions that manifest directly in LMS activity patterns (Syauqi et al., 2024).

In response to this challenge, learning analytics has emerged as a promising framework for converting the rich behavioral trace data generated by Learning Management Systems (LMS) and Virtual Learning Environments (VLE) into actionable early warning signals. LMS interaction logs, including click volumes, session frequencies, material access patterns, and activity timelines, have been demonstrated to carry strong predictive signals for dropout risk (Romdhoni & Romdhoni, 2026). Systematic reviews confirm that engagement metrics derived from LMS logs, particularly temporal patterns of student activity, are among the most reliable and consistent predictors of academic outcomes (Karim-Abdallah et al., 2025; Romdhoni et al., 2025). Early Warning Systems (EWS) built on these analytics can enable the timely identification of at-risk students, creating intervention windows before dropout becomes irreversible (Čotić Poturić et al., 2025; Shiao et al., 2023).

A growing body of research has applied machine learning (ML) classification algorithms to the dropout prediction problem. Systematic reviews spanning the last decade consistently identify Random Forest, Logistic Regression, and Support Vector Machine (SVM) as the most frequently employed and best-performing models for at-risk student prediction (Alalawi et al., 2023; Ersozlu et al., 2024). Random Forest and other ensemble methods often achieve accuracy above 90% on LMS-based datasets, while Logistic Regression serves as a competitive interpretable baseline that can rival ensemble approaches on large tabular datasets (Althibyani, 2024). Among publicly available benchmark datasets for this problem, the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017) has become a standard reference, providing authentic VLE clickstream data, including daily summaries of student interactions from 32,593 students across seven course modules (Kuzilek et al., 2017). Prior work using OULAD demonstrates that VLE-derived behavioral features, particularly click volume and assignment-related variables, dominate predictive importance over demographic attributes (Alnasyan et al., 2024).

Despite these advances, several gaps remain in the existing literature. First, while many studies compare ML models based on accuracy, few systematically examine Recall as the priority metric in the specific context of Early Warning Systems, where a missed at-risk student (False Negative) carries greater pedagogical consequences than a false alarm (False Positive) (Čotić Poturić et al., 2025). Second, although feature importance is frequently reported, the comparative contribution of temporal behavioral features versus demographic features from authentic VLE clickstream data remains underexplored in studies that explicitly frame the problem as dropout prediction rather than general performance prediction (Alnasyan et al., 2024). Third, most studies use a single dataset, limiting the generalizability of conclusions about which feature types are truly predictive across contexts (Almalawi et al., 2024; Matz et al., 2023).

To address these gaps, this study makes four distinct contributions that differentiate it from prior OULAD-based research. Unlike prior studies focusing mainly on accuracy, this study emphasizes intervention-sensitive metrics such as Recall and False Negatives, directly aligning model evaluation with the operational priorities of EWS deployment. Specifically, the contributions are: (1) empirical comparison of Logistic Regression, Random Forest, and SVM with Recall as the primary EWS metric, rather than accuracy, to reflect the asymmetric cost of missed at-risk cases; (2) systematic analysis of operational tradeoffs including False Negative rates and AUC-ROC across classifiers to provide institution-facing guidance on model selection for EWS deployment; (3) comprehensive feature importance analysis quantifying the predictive dominance of temporal behavioral features, particularly Activity Span, over demographic attributes; and (4) cross-dataset convergence evidence comparing findings with a prior study using the UCI Students Dropout dataset, demonstrating that temporal engagement consistency is a robust and generalizable dropout predictor

independent of data source type, with implications for translating these findings into institutional policy decisions for online learning environments.

2. RESEARCH METHOD

This study employs a quantitative experimental design using machine learning classification to predict student withdrawal risk from VLE behavioral data. The complete research workflow is presented in Figure 1.

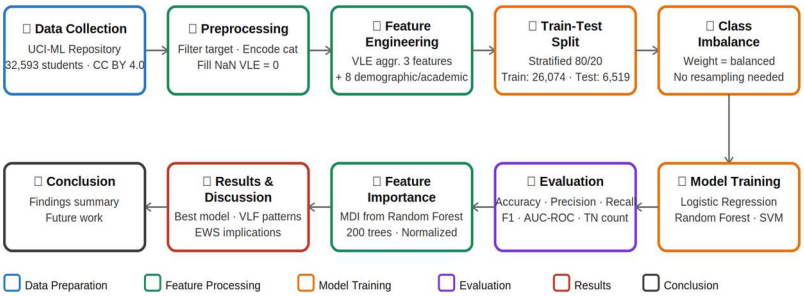

Figure 1. Research Workflow Diagram

Figure 1 presents the eight-stage research pipeline organized across five process categories. The workflow begins with Data Collection from the UCI-ML Repository (32,593 students, CC BY 4.0), followed by Preprocessing involving target encoding, categorical encoding, and zero-imputation for missing VLE values. Feature Engineering aggregates VLE interactions into three behavioral features combined with eight demographic attributes, yielding thirteen input features in total. The dataset is then divided in the Train-Test Split stage using a stratified 80/20 ratio (Train: 26,074; Test: 6,519). A Class Imbalance assessment confirmed that balanced class weighting was sufficient without resampling. Three classifiers, Logistic Regression, Random Forest, and SVM, are trained in the Model Training stage, and subsequently assessed in the Evaluation stage using Accuracy, Precision, Recall, F1-score, and AUC-ROC, with Recall as the primary metric. Feature Importance is derived from Random Forest MDI scores (200 trees, normalized). The pipeline concludes with Results and Discussion and a Conclusion summarizing findings and future directions.

2.1. DATASET

This study uses the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017), a publicly available dataset comprising 32,593 students enrolled in 7 undergraduate modules at The Open University, UK. What distinguishes OULAD from conventional academic datasets is its inclusion of authentic VLE interaction logs daily summaries of student click activity, enabling behavioral engagement modeling not possible with static administrative records. Table 1 summarizes the dataset characteristics.

Table 1. Dataset Characteristics

Characteristic	Value
Dataset name	Open University Learning Analytics Dataset (OULAD)
Source	UCI ML Repository (ID: 349); DOI: 10.24432/C5KK69
Citation	Kuzilek, Hlosta & Zdrahal (2017), Scientific Data, 4:170171
License	Creative Commons Attribution 4.0 (CC BY 4.0)
Total students	32,593 (no missing values)
Modules / Presentations	7 modules, 4 course presentations
Target variable (binary)	Withdrawn = 1 (at-risk); Non-Withdrawn = 0
Class distribution	Withdrawn: 10,156 (31.2%) / Non-Withdrawn: 22,437 (68.8%)
Train set (80%)	26,074 students — stratified split, random_state=42
Test set (20%)	6,519 students (Withdrawn: 2,031 / Non-Withdrawn: 4,488)

The binary classification target is defined as: Withdrawn = 1 (at-risk) and Non-Withdrawn = 0. The Fail class is retained within the Non-Withdrawn group since the primary focus is identifying students who disengage and leave the course entirely.

2.2. FEATURE ENGINEERING

Thirteen features were constructed from two primary sources. Five VLE behavioral features were derived by aggregating per-student interaction logs from the studentVle table. Eight demographic and academic features were obtained from the studentInfo table. Prior work using OULAD confirms that click-based and temporal features carry the highest predictive weight for withdrawal risk (Alnasyan et al., 2024; Balabied & Eid, 2023). Table 2 describes all features.

Table 2. Feature Set Used in the Predictive Model

No	Feature	Source	Description
1	activity_span	VLE	Temporal range between first & last LMS interaction (days)
2	total_sessions	VLE	Number of unique active days on LMS
3	total_clicks	VLE	Total click interactions with LMS content
4	avg_clicks_per_day	VLE	Average clicks per active session day
5	first_activity	VLE	Day number of first recorded LMS interaction
6	imd_band	Demographic	Index of Multiple Deprivation (socioeconomic indicator)
7	code_module	Academic	Course module identifier (7 categories)
8	studied_credits	Academic	Total credits enrolled for current presentation
9	highest_education	Demographic	Highest prior educational qualification
10	num_of_prev_attempts	Academic	Number of previous module registration attempts
11	age_band	Demographic	Student age group (0–35 / 35–55 / 55+)
12	gender	Demographic	Student gender (Male / Female)
13	disability	Demographic	Declared disability status (Yes / No)

2.3. DATA PREPROCESSING

Preprocessing involved four steps: (1) removing "Enrolled" records without a definitive outcome label; (2) binary target encoding; (3) encoding of categorical features using two strategies one-hot encoding was applied to nominal variables (gender, code_module, and highest_education) to avoid imposing false ordinal relationships, while label encoding was retained for ordinal variables (age_band and imd_band) where a meaningful rank order exists; and (4) zero-imputation for students with no VLE activity, reflecting genuine absence of engagement. A stratified 80/20 train-test split (random_state=42) was applied to preserve class proportions across partitions, critical given the 31.2%/68.8% class imbalance (Albreiki et al., 2021; Andrade-Girón et al., 2023).

2.4. MACHINE LEARNING MODELS

Three supervised classification algorithms were selected based on their prevalence in the dropout prediction literature (Alalawi et al., 2023; Andrade-Girón et al., 2023). All models use class_weight="balanced" to address class imbalance without resampling, preserving the original data distribution. Table 3 details configurations.

Table 3. Model Configurations and Hyperparameters

Model	Key Hyperparameters	Rationale
Logistic Regression	class_weight=balanced, max_iter=1000, solver=lbfgs	Interpretable baseline; low computational cost
Random Forest	n_estimators=200, class_weight=balanced, random_state=42, n_jobs=-1	Robust ensemble; provides MDI feature importance
SVM	kernel=rbf, class_weight=balanced, probability=True, random_state=42	Effective in high-dimensional space; strong recall

2.4.1. Logistic Regression

Logistic Regression models the probability of class membership using the sigmoid function:

$$P(y = 1 \mid x) = 1 / (1 + exp(-(B0 + B1x1 + B2x2 + \ldots + Bnxn))) \quad \ldots(1)$$

where β_i are learned coefficients and x_i are input features. Classification is performed by thresholding at P = 0.5. The model is trained by minimizing the binary cross-entropy loss:

$$L = -(1/N) \times sum[\, yi \times log(yi\_hat) + (1 - yi) \times log(1 - yi\_hat)\,] \quad \ldots(2)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of training samples.

2.4.2. Random Forest

Random Forest is an ensemble of T = 200 decision trees, each trained on a bootstrap sample of the training data with random feature subsets. The final prediction is obtained by majority voting:

$$\hat{y} = mode\,\{\, f_t(x)\,\}^{T}_{t=1} \quad \ldots(3)$$

where $f_t(x)$ is the prediction of the t-th tree. Node splitting uses Gini impurity as the criterion:

$$Gini(n) = 1 - \Sigma_k\, p(k|n)^2 \quad \ldots(4)$$

where p(k|n) is the proportion of class k at node n. The tree selects the split that maximizes the weighted reduction in Gini impurity (ΔGini).

2.4.3. Support Vector Machine (SVM)

SVM finds the optimal hyperplane that maximizes the margin between classes. With the RBF (Radial Basis Function) kernel, the feature space is transformed as:

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \quad \ldots(5)$$

where γ controls the influence radius of each training point. The decision function is:

$$f(x) = sign(\Sigma_i\, \alpha_i\, y_i\, K(x_i, x) + b) \quad \ldots(6)$$

where α_i are Lagrange multipliers and b is the bias term. With class_weight=balanced, the soft-margin parameter C is adjusted per class to compensate for imbalance.

2.5. EVALUATION METRICS

Recall is the primary metric for EWS because a False Negative (missed at-risk student) carries greater pedagogical cost than a False Positive. The five metrics are:

$$(7)\ Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$(8)\ Precision = TP / (TP + FP)$$

$$(9)\ Recall = TP / (TP + FN) \leftarrow Priority\ metric$$

(10) F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

$$(11)\ AUC - ROC = \int_0^1 TPR(FPR^{-1}(t))\, dt$$

where TP = True Positive (Withdrawn correctly predicted), TN = True Negative, FP = False Positive, FN = False Negative. False Negative count is also reported as a direct indicator of missed at-risk students.

2.6. FEATURE IMPORTANCE ANALYSIS (MDI)

Feature importance is computed using Mean Decrease in Impurity (MDI) from the Random Forest across all 200 trees:

$$(12)\ fi\_j = (1/T) \times \Sigma_{t=1}^{T}\, \Sigma\_\{n \in N_{t,j}\}\, [p(n) \times \Delta Gini(n)]$$

where $N_{t,j}$ = nodes in tree t using feature j, and p(n) = proportion of samples at node n. Scores are normalized:

$$(13)\ fi\_j* = fi\_j \,/\, \Sigma_j\, fi\_j \quad \rightarrow \quad \Sigma_j\, fi\_j* = 1.00$$

3. RESULT AND ANALYSIS

3.1. DATASET OVERVIEW

The OULAD dataset comprises 32,593 students with no missing values. As shown in Figure 2, 31.2% (10,156) are classified as Withdrawn the positive dropout class. This rate is consistent with the elevated dropout levels reported in distance and online learning institutions documented across the literature (Seo et al., 2024). Withdrawn rates are similar across gender (Male: 31.7%; Female: 30.5%), suggesting that gender alone is not a strong differentiating factor consistent with findings that dropout is driven more by behavioral and contextual factors than demographic characteristics (González-Morales et al., 2025; Kocsis & Molnár, 2025). However, substantial module-level variation is observed: module CCC recorded the highest withdrawal rate (44.5%) while GGG recorded the lowest (11.5%), indicating that module-specific factors workload, course design, and instructional support meaningfully influence dropout risk (Rahmani et al., 2024).

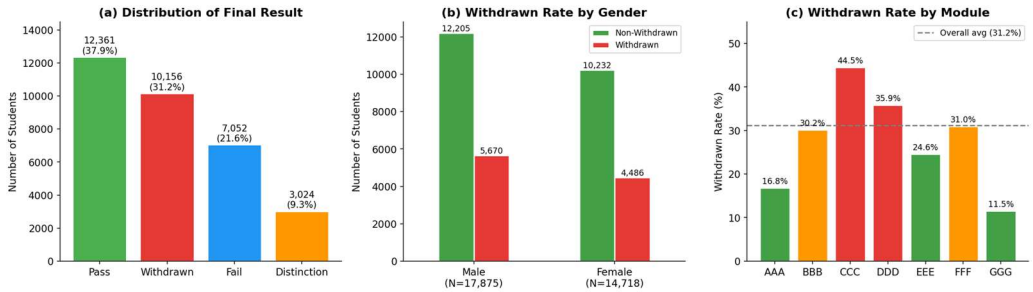

Figure 2. Dataset Overview: Final Result Distribution, Withdrawn Rate by Gender and Module (N = 32,593).

3.2. VLE ENGAGEMENT PATTERNS

Figure 3 reveals a striking and monotonic engagement gradient across outcome groups. Students achieving Distinction averaged 2,667 clicks and 110 active sessions, compared to Withdrawn students who averaged only 314 clicks and 16 sessions an eight-fold difference in total interaction volume. This pattern provides strong empirical justification for using VLE-derived behavioral features as primary predictors, corroborating evidence from systematic reviews that LMS interaction logs are among the most reliable indicators of dropout risk (Karim-Abdallah et al., 2025). The gradient also implies that disengagement from LMS activity is not abrupt but gradual, creating a detection window that EWS models can exploit (Seo et al., 2024).

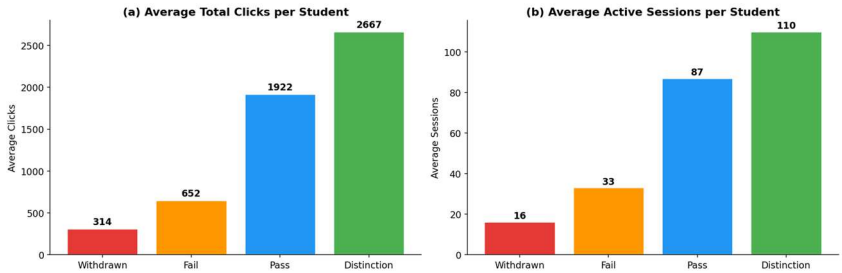

Figure 3. Average VLE Engagement (Total Clicks and Active Sessions) by Final Result

3.3. MODEL PERFORMANCE COMPARISON

Table 4 and Figures 4, 5 present the comparative performance of the three models on the test set (N = 6,519).

Table 4. Model Performance on Test Set (N = 6,519)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	FN
Logistic Regression	0.856	0.727	0.858	0.787	0.932	288
Random Forest	0.866	0.772	0.807	0.789	0.940	391
SVM	0.859	0.717	0.903	0.800	0.931	196

Table 5. Model Suitability Ranking

EWS Rank	Model	Best For	Limitation
1	SVM	Maximizing Recall and minimizing missed at-risk students (FN = 196); highest EWS suitability	Lower Precision (0.717) and higher false alarm rate; computationally intensive on large datasets
2	Random Forest	Best overall Accuracy (0.866), Precision (0.772), F1-Score (0.789), and AUC-ROC (0.940); robust feature importance interpretation	Lowest Recall (0.807) and highest FN (391); less suitable as primary EWS classifier
3	Logistic Regression	Competitive interpretable baseline; strong AUC-ROC (0.932) with minimal computational overhead; suitable for resource-constrained institutions	Lowest Accuracy (0.856) and F1-Score (0.787); linear decision boundary may underfit complex behavioral patterns

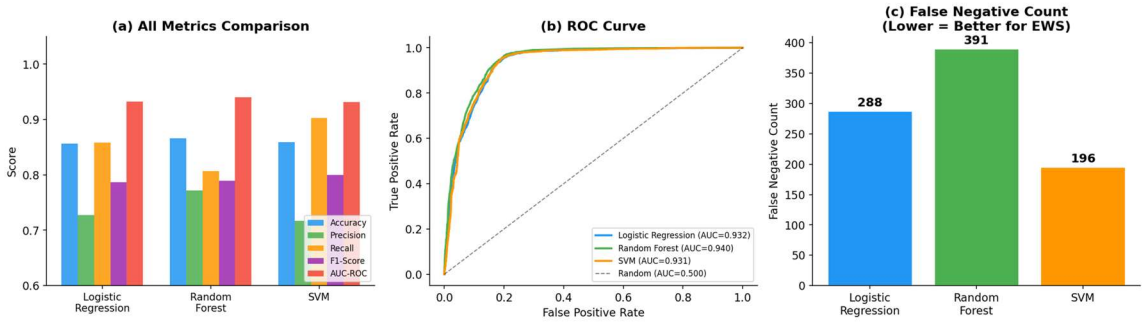

Figure 4. Model Performance: (a) All Metrics, (b) ROC Curves, (c) False Negative Count

Figure 4. Model Performance: (a) All Metrics, (b) ROC Curves, (c) False Negative Count

Figure 4(a) presents a grouped bar chart comparing all five evaluation metrics, Accuracy, Precision, Recall, F1-Score, and AUC-ROC across the three classifiers. SVM achieves the highest Recall (0.903), while Random Forest leads in Accuracy (0.866), Precision (0.772), F1-Score (0.789), and AUC-ROC (0.940). Logistic Regression demonstrates competitive performance across all metrics, confirming its viability as an interpretable baseline.

Figure 4(b) displays the Receiver Operating Characteristic (ROC) curves for all three models. All classifiers achieve AUC-ROC values exceeding 0.93: Random Forest (0.940), Logistic Regression (0.932), and SVM (0.931), indicating strong discriminative ability between withdrawn and non-withdrawn students across all classification thresholds. The curves of all three models cluster closely together and substantially above the random baseline (AUC = 0.500), suggesting comparable overall separability.

Figure 4(c) illustrates the False Negative (FN) count per model, which reflects the number of at-risk students incorrectly classified as non-withdrawn. In an EWS context, a lower FN count is operationally preferable as it minimizes missed interventions. SVM records the fewest false negatives (196), followed by Logistic Regression (288) and Random Forest (391), reinforcing SVM's suitability as the primary classifier for dropout early warning deployment despite its marginally lower overall accuracy.

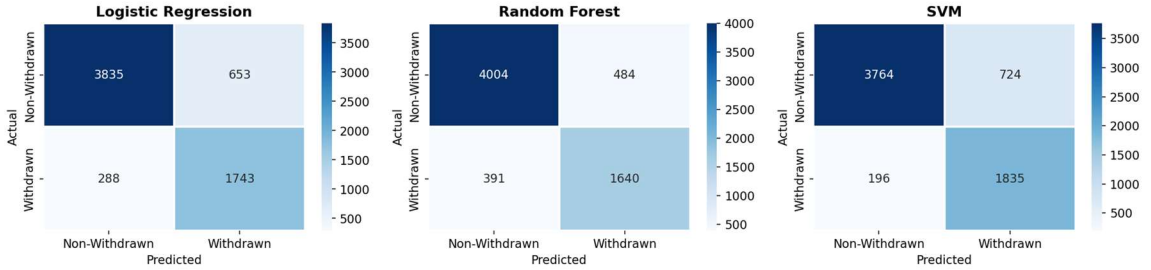

Figure 5. Confusion Matrices for Each Model on Test Set (N = 6,519)

All three models achieved competitive performance with AUC-ROC exceeding 0.93, consistent with prior reports of high accuracy for ensemble and linear models on OULAD-type datasets (Balabied & Eid, 2023). Random Forest achieved the highest Accuracy (0.866) and AUC-ROC (0.940), reflecting strong overall discriminative capability consistent with the repeated finding across systematic reviews that ensemble methods achieve near-optimal performance for dropout classification (Andrade-Girón et al., 2023). However, for EWS purposes, SVM demonstrated the strongest suitability with the highest Recall (0.903) and the lowest False Negative count (196 out of 2,031 Withdrawn students), meaning only 9.7% of at-risk students were missed. This aligns with the position that Recall must be the priority metric in educational early warning contexts, as undetected at-risk students lose intervention opportunities (Čotić Poturić et al., 2025). Logistic Regression served as a competitive and interpretable baseline (Recall = 0.858; AUC-ROC = 0.932) at substantially lower computational cost, consistent with evidence that it can match ensemble models on large tabular datasets (Alalawi et al., 2023; Althibyani, 2024).

The ROC curves confirm that all models substantially outperform the random classifier across all thresholds (Ersozlu et al., 2024). The confusion matrix analysis illustrates a clear precision-recall trade-off: SVM adopts an aggressive detection strategy accepting more False Positives to minimize missed cases while Random Forest is more conservative. For institutions with limited counseling capacity, the lower False Positive rate of Random Forest may reduce intervention burden, whereas SVM maximizes detection coverage. The appropriate choice therefore depends on institutional context and available intervention resources (Čotić Poturić et al., 2025).

3.4. FEATURE IMPORTANCE ANALYSIS

Table 6 reports the Mean Decrease in Impurity (MDI) scores derived from the Random Forest classifier, ranking all 13 input features by their relative contribution to the model's predictive performance. MDI quantifies how much each feature reduces node impurity across all decision trees in the ensemble, with higher values indicating greater predictive importance. The Contribution column expresses each MDI score as a percentage share of total importance mass, while the Level column categorizes features into three tiers: High (above 10%), Medium (5–10%), and Low (below 5%). VLE behavioral features collectively account for 85.0% of total predictive importance, with Activity Span emerging as the single dominant predictor (MDI = 0.413, 41.3%), followed by Total Sessions (15.2%) and Total Clicks (13.7%). Demographic and academic features contribute minimally on an individual basis, each accounting for less than 4.0%, indicating that temporal engagement patterns within the LMS are substantially more predictive of dropout risk than student background characteristics alone.

Table 6. MDI ranking

Rank	Feature	Category	MDI	Contribution	Level
1	Activity Span	VLE	0.4125	41.3%	High
2	Total Sessions	VLE	0.1517	15.2%	High
3	Total Clicks	VLE	0.1368	13.7%	High
4	Avg Clicks/Day	VLE	0.0838	8.4%	Medium
5	First Activity	VLE	0.0644	6.4%	Medium
6	IMD Band	Demographic	0.0382	3.8%	Low

Rank	Feature	Category	MDI	Contribution	Level
7	Module Code	Academic	0.0353	3.5%	Low
8	Studied Credits	Academic	0.0269	2.7%	Low
9	Highest Education	Demographic	0.0163	1.6%	Low
10	Prev Attempts	Academic	0.0105	1.0%	Low
11	Age Band	Demographic	0.0090	0.9%	Low
12	Gender	Demographic	0.0086	0.9%	Low
13	Disability	Demographic	0.0058	0.6%	Low

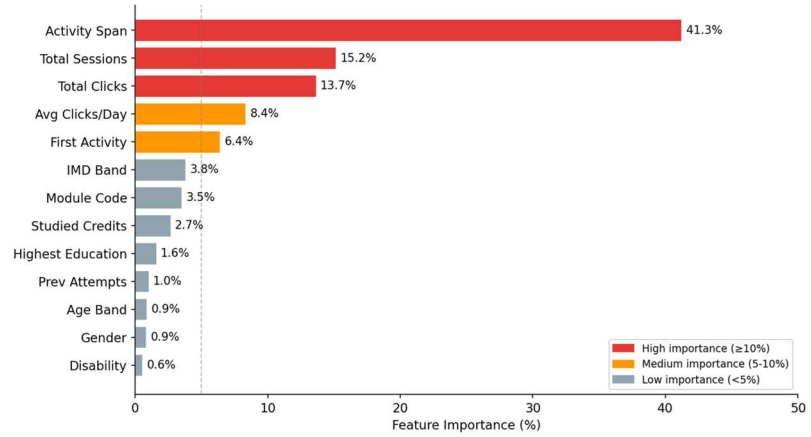

Figure 6. Feature Importance: VLE Features vs Demographic Features (MDI, Random Forest)

Activity Span emerged as the dominant predictor (MDI = 0.4125; 41.3%), representing the temporal window of a student's LMS presence from first to last interaction. The five VLE features collectively account for 85.0% of total importance, while all demographic features combined contribute only 4.9%. This result converges with findings from prior OULAD-based studies where click volume and submission timing dominated feature rankings (Alnasyan et al., 2024; Balabied & Eid, 2023) and with the broader literature confirming that behavioral engagement metrics from LMS logs are the strongest predictors of academic outcomes (Karim-Abdallah et al., 2025; Romdhoni & Romdhoni, 2026).

Notably, the dominance of Activity Span over raw click volume (Total Clicks: 13.7%) suggests that it is not merely how much a student interacts with the LMS, but how persistently they remain engaged over time, that best distinguishes students who complete from those who withdraw. This temporal dimension aligns with findings that regular, early engagement patterns not just cumulative activity are the most informative signals in VLE-based prediction (Alnasyan et al., 2024; Romdhoni et al., 2025). The marginal contribution of demographic features (gender: 0.9%; disability: 0.6%) further implies that dropout prediction in online settings can be operationalized primarily through behavioral data, reducing reliance on potentially sensitive demographic attributes (Almalawi et al., 2024; Rahmani et al., 2024).

3.5. CROSS-STUDY COMPARISON

The findings of this study converge with results from a prior study using the UCI Students Dropout dataset an administrative dataset without direct LMS logs where temporal engagement proxies (Inactivity Streak and Access Duration) also emerged as the dominant predictors. This consistency across two substantially different datasets and institutional contexts provides convergent evidence that temporal engagement consistency is a robust and generalizable predictive signal for dropout risk, independent of data source type. This cross-dataset convergence strengthens the theoretical claim that behavioral disengagement manifested as decreasing LMS presence over time is a primary mechanism underlying dropout, rather than a statistical artifact of any particular dataset (Alnasyan et al., 2024).

3.6. LIMITATIONS

First, OULAD originates from a single distance learning institution (The Open University, UK); generalizability to face-to-face or blended learning contexts requires empirical verification (Rahmani et al., 2024). Second, VLE features represent full-course aggregates future work should explore weekly or bi-weekly time windows that may enable earlier detection and more actionable intervention timing (Matz et al., 2023; Seo et al., 2024). Third, Label encoding was applied to nominal categorical features under an ordinality assumption; one-hot or target encoding alternatives may reduce representation bias and should be explored in future iterations (Albreiki et al., 2021).

4. CONCLUSION

This study confirms that student dropout in online higher education can be accurately predicted using LMS behavioral engagement data. All models achieved strong performance (AUC-ROC > 0.93), with Support Vector Machine (SVM) showing the highest Recall (0.903), making it the most suitable model for Early Warning Systems (EWS) where minimizing missed at-risk students is critical.

Feature importance analysis reveals that VLE behavioral features dominate prediction (85.0%), with Activity Span as the strongest predictor, indicating that temporal engagement consistency is more important than total activity volume. Demographic factors contribute minimally.

The findings, supported by cross-dataset consistency, demonstrate that dropout is primarily driven by progressive behavioral disengagement observable through LMS activity patterns. This highlights the importance of prioritizing high-recall models and temporal engagement indicators in data-driven intervention systems for online learning.

Discussion of the results of research and testing obtained presented in the form of theoretical descriptions, both qualitatively and quantitatively. The results of the experiment should be displayed in either a graph or table. For charts can follow the format for diagrams and drawings.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Open University (OU) for making the Open University Learning Analytics Dataset (OULAD) publicly available through the UCI Machine Learning Repository under the CC BY 4.0 license. This research would not have been possible without the effort and commitment of those who collected, curated, and shared the dataset.

REFERENCES

Alalawi, K., Athauda, R., & Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, *5*(12). https://doi.org/10.1002/eng2.12699

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9). https://doi.org/10.3390/educsci11090552

Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive Models for Educational Purposes: A Systematic Review. *Big Data and Cognitive Computing*, *8*(12). https://doi.org/10.3390/bdcc8120187

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*. https://doi.org/10.1016/j.caeai.2024.100231

Althibyani, H. A. (2024). Predicting student success in MOOCs: a comprehensive analysis using machine learning models. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.2221

Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E.,

Ausejo-Sanchez, J., Villarreal-Torres, H., & Angeles-Morales, J. (2023). Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review. *EAI Endorsed Transactions on Scalable Information Systems*, *10*(5), 1–11. https://doi.org/10.4108/eetsis.3586

Balabied, S. A. A., & Eid, H. F. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science*, *9*. https://doi.org/10.7717/peerj-cs.1708

Čotić Poturić, V., Čandrlić, S., & Dražić, I. (2025). A Scoring Algorithm for the Early Prediction of Academic Risk in STEM Courses. *Algorithms*, *18*(4). https://doi.org/10.3390/a18040177

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, *5*(4). https://doi.org/10.3390/bdcc5040064

Ersozlu, Z., Taheri, S., & Koch, I. (2024). A review of machine learning methods used for educational data. *Education and Information Technologies*, *29*(16), 22125–22145. https://doi.org/10.1007/s10639-024-12704-0

González-Morales, M. O., López-Aguilar, D., Álvarez-Pérez, P. R., & Toledo-Delgado, P. A. (2025). Dropping out of higher education: Analysis of variables that characterise students who interrupt their studies. *Acta Psychologica*, *252*. https://doi.org/10.1016/j.actpsy.2024.104669

Guzmán, A., Barragán, S., & Cala Vitery, F. (2021). Dropout in Rural Higher Education: A Systematic Review. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.727833

Karim-Abdallah, B., Weyori, B. A., & Mensah, P. K. (2025). Using Lms Log Data To Identify At-Risk Students: a Systematic Review of Machine Learning Approaches and Bibliographic Analysis. *African Journal of Applied Research*, *11*(2), 278–312. https://doi.org/10.26437/ajar.v11i2.1038

Kocsis, Á., & Molnár, G. (2025). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, *51*(3), 414–432. https://doi.org/10.1080/03054985.2024.2316616

Kurulgan, M. (2024). a Bibliometric Analysis of Research on Dropout in Open and Distance Learning. *Turkish Online Journal of Distance Education*, *25*(4), 201–229. https://doi.org/10.17718/tojde.1355394

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Data Descriptor: Open University Learning Analytics dataset. *Scientific Data*, *4*. https://doi.org/10.1038/sdata.2017.171

Leow, T., Li, W. W., Miller, D. J., & McDermott, B. (2025). Prevalence of university non-continuation and mental health conditions, and effect of mental health conditions on non-continuation: a systematic review and meta-analysis. *Journal of Mental Health*, *34*(2), 222–237. https://doi.org/10.1080/09638237.2024.2332812

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-32484-w

Rahmani, A. M., Groot, W., & Rahmani, H. (2024). Dropout in online higher education: a systematic literature review. *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-024-00450-9

Romdhoni, R. D., Arrasyid, R., Widodo, S., Elviani, U., & Kunci, K. (2025). *AI-Driven Learning Analytics for Self-Regulated and Metacognitive Learning : A Systematic Review*. *04*(02), 329–347.

Romdhoni, R. D., & Romdhoni, A. (2026). *The Effectiveness of Learning Management Systems (LMS) in Enhancing Learning Experiences toward Achieving SDG 4 : Quality Education*. *4*(1), 39–46.

Rotar, O. (2022). A missing theoretical element of online higher education student attrition, retention, and progress: a systematic literature review. *SN Social Sciences*, *2*(12). https://doi.org/10.1007/s43545-022-00550-1

Sánchez-Gelabert, A. (2020). Non-traditional students, university trajectories, and higher education institutions: A comparative analysis of face-to-face and online universities. *Studia Paedagogica*, *25*(4), 51–72. https://doi.org/10.5817/SP2020-4-3

Seo, E. Y., Yang, J., Lee, J. E., & So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. *Heliyon*, *10*(11). https://doi.org/10.1016/j.heliyon.2024.e30960

Shiao, Y. T., Chen, C. H., Wu, K. F., Chen, B. L., Chou, Y. H., & Wu, T. N. (2023). Reducing dropout rate through a deep learning model for sustainable education: long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021. *Smart Learning Environments*, *10*(1). https://doi.org/10.1186/s40561-023-00274-6

Syauqi, S. K., Winarno, N., Samsudin, A., Damopolii, I., & Firdaus, R. A. (2024). From Online To in-Person: Students' Motivation and Self-Regulation in Science Teaching Activities During and After the Covid-19 Pandemic. *INSECTA: Integrative Science Education and Teaching Activity Journal*, *5*(1), 87–107. https://doi.org/10.21154/insecta.v5i1.8689