Quantile-based e-Learning Student Engagement Classification

Aditya Galih Sulaksono[1,2], Syaad Patmanthara[1] and Harits Ar Rosyid[1]
[1]State University of Malang, Indonesia
[2]Universitas Merdeka Malang, Indonesia

aditya.galih.2205349@students.um.ac.id
syaad.ft@um.ac.id
harits.ar.ft@um.ac.id

https://doi.org/10.34190/ejel.24.3.4678

Abstract: Classifying student engagement accurately is critical for timely academic intervention; however, most existing approaches rely on arbitrarily defined thresholds that lack statistical grounding and are difficult to transfer across institutional contexts. This limitation reduces the practical applicability of engagement analytics in diverse educational settings. This study evaluates a quantile-based engagement classification framework across two contrasting datasets to assess its validity, transferability, and consistency of predictive features. Unlike threshold-based approaches, the proposed framework derives engagement categories directly from dataset-specific interaction distributions. The Open University Learning Analytics Dataset (OULAD) represents large-scale fully online learning, while the Unistudium dataset reflects a smaller blended learning context. The two datasets differ substantially in size and delivery mode, with a student ratio of approximately 17.4 to 1. This contrast provides a rigorous basis for assessing method transferability. Engagement categories (passive, moderate, and active) are derived using dataset-specific quartile thresholds (Q1 and Q3). This strategy adapts automatically to local interaction distributions and avoids manual parameter tuning. Five temporal behavioural features were extracted, including active days, unique actions, and learning consistency. Random Forest was employed as the proposed model, while a Decision Tree classifier was included as a baseline for comparative evaluation. The results indicate that the proposed framework remains effective across different educational contexts. In the OULAD dataset, the model achieved an accuracy of 92.04% with a Cohen's κ of 0.87. In the Unistudium dataset, accuracy reached 72.50% with a Cohen's κ of 0.59. Although performance differed between datasets, variance remained low. Feature importance analysis further revealed strong consistency across contexts, with a Spearman correlation of 0.90. Active days and unique actions were the most influential predictors in both cases. The baseline comparison further confirmed the superiority of Random Forest over the Decision Tree baseline across both datasets. These findings support e-learning practice by offering institutions a statistically grounded and automated method for engagement classification. The approach removes the need for arbitrary thresholds and reduces operational overhead in analytics deployment. From a research perspective, the study establishes realistic performance benchmarks for engagement analytics at different institutional scales, demonstrates the applicability of quantile-based engagement classification across heterogeneous datasets, and confirms that key behavioural engagement indicators transfer reliably across online and blended learning environments.

Keywords: Learning analytics, Student engagement, Quantile classification, Cross-dataset validation, Random forest, Educational data mining

1. Introduction

When engagement classification fails, the cost is borne by students. Misclassifying a disengaged student as moderately active delays academic intervention, while misclassifying an active student as struggling wastes limited tutoring resources. At institutional scale, even modest error rates translate into hundreds of missed support opportunities each semester. Digital learning platforms generate extensive interaction data that can inform student support, intervention, and learning analytics applications (Xu *et al.*, 2025). However, converting these logs into meaningful engagement categories remains methodologically problematic. A central issue lies in how engagement thresholds are defined and validated. Many studies continue to rely on fixed interaction counts that lack clear statistical grounding (Conijn *et al.*, 2017). As a result, engagement definitions vary widely across the literature, making comparison and replication difficult (Howard, Meehan and Parnell, 2018). These rigid thresholds are particularly fragile in cross-institutional settings. A boundary that performs well in a large research-intensive university may be inappropriate for a smaller, teaching-oriented institution.

Alternative analytical approaches introduce different limitations. Unsupervised clustering methods often produce results that are difficult to interpret and insufficiently stable for operational use (Howard, Meehan and Parnell, 2018). Supervised classification methods, by contrast, require extensive manual labelling and tend to perform poorly when applied beyond their original context (Brinton and Chiang, 2015). Both approaches therefore present challenges for institutions seeking scalable and transferable engagement analytics.

Cite this article: Sulaksono, A.G., Patmanthara, S. and Ar Rosyid, H. 2026. "Quantile-based e-Learning Student Engagement Classification", *Electronic Journal of e-Learning*, 24(3), pp 106-122, https://doi.org/10.34190/ejel.24.3.4678

Quantile-based classification offers a promising alternative. By defining engagement categories using distribution-derived thresholds calculated automatically from each dataset, this approach provides a statistically grounded basis for classification (Xu *et al.*, 2025). Thresholds are computed directly from local interaction patterns, allowing the method to adapt automatically to different datasets. This property reduces the need for manual tuning and supports consistent interpretation across contexts.

Despite these advantages, evidence for the cross-context validity of quantile-based engagement classification remains limited. Most existing studies evaluate performance within a single dataset or institutional setting (Rizvi *et al.*, 2022). As a result, it is unclear whether reported performance levels can be expected to transfer to institutions with different scales, delivery modes, or student populations.

To address this gap, the present study evaluates quantile-based engagement classification using two contrasting datasets. The Open University Learning Analytics Dataset (OULAD) represents large-scale distance education, while the Unistudium dataset reflects a smaller blended learning environment. The two contexts differ markedly in institutional setting, delivery model, and scale, with a student ratio of approximately 17.4 to 1. This contrast provides a rigorous basis for examining method generalisability.

The quantile-based approach is particularly suited to this validation task. Quartile thresholds are calculated independently for each dataset, eliminating the need to transfer thresholds between contexts. This automatic adaptation is a central methodological strength. If comparable classification performance and feature importance patterns emerge despite large differences in interaction volume and institutional scale, this would indicate that the underlying engagement constructs are robust.

Accordingly, this study addresses three research questions. First, can quantile-based engagement classification achieve acceptable accuracy across different educational contexts when thresholds adapt automatically to local data distributions? Second, do feature importance patterns remain stable across datasets with substantial differences in scale and institutional characteristics? Third, how do engagement category characteristics differ between large-scale distance education and smaller blended learning environments?

Answering these questions contributes to both theory and practice in learning analytics. First, the study provides empirical evidence regarding the cross-context validity of quantile-based engagement classification. Second, it identifies behavioural features that remain informative across educational settings with substantially different scales and delivery modes. Third, it establishes practical performance benchmarks that can guide institutions seeking to implement transferable engagement analytics frameworks. Demonstrating cross-context validity would support wider institutional adoption of quantile-based methods. Identifying stable behavioural predictors would guide feature selection in new implementations. Finally, clarifying how performance varies with dataset scale would help institutions form realistic expectations when deploying engagement analytics systems.

2. Literature Review

This literature review establishes the theoretical and methodological foundation for the proposed engagement classification framework. The review is organised into four interconnected areas. First, it examines the concept of student engagement in digital learning and its behavioural manifestations within learning management systems. Second, it reviews existing approaches to engagement classification, highlighting the limitations of arbitrary threshold-based, clustering-based, and supervised methods. Third, it discusses quantile-based classification as a statistically grounded alternative for defining engagement categories. Finally, it examines the role of cross-dataset validation in learning analytics research and the challenges associated with transferring analytical models across educational contexts. Together, these strands of literature provide the rationale for evaluating a quantile-based engagement classification framework across heterogeneous learning environments.

2.1 Student Engagement in Digital Learning

Student engagement encompasses behavioural, emotional, and cognitive dimensions (Fredricks et al., 2004). Behavioural engagement manifests through participation in learning activities, attendance, and time-on-task. Emotional engagement reflects students' affective responses to learning experiences. Cognitive engagement involves self-regulation and strategic learning approaches. While all three dimensions matter, digital learning environments most readily capture behavioural indicators through interaction logs.

Recent work has reinforced this framing. (Saqr, Fors and Nouri, 2018), in their longitudinal analysis of online engagement across a full programme, confirm that behavioural engagement signals from interaction logs reliably anticipate downstream academic outcomes when measured consistently over time. (Bergdahl *et al.*,

2024) further argue that the trustworthiness of any engagement instrument depends on the methodological transparency of how its categories are defined.

Learning management systems record every student action: accessing content, submitting assignments, participating in forums, watching videos, and taking assessments (Kuzilek, Hlosta and Zdrahal, 2017). These interaction logs provide granular data about when, how often, and which resources students engage with. Temporal patterns emerge from these logs. Some students access materials consistently throughout the course, while others concentrate activity around assessment deadlines.

The relationship between interaction volume and academic success remains complex. More interactions generally correlate with better outcomes, but the relationship is not linear (Conijn *et al.*, 2017). A student with 1,000 interactions distributed across diverse activities and consistent time periods demonstrates qualitatively different engagement than a student with 1,000 interactions concentrated in a single cramming session before the final exam. This distinction motivates the development of temporal-behavioural features that capture engagement quality rather than mere quantity. These behavioural differences motivate the need for classification approaches that can distinguish meaningful engagement patterns while remaining interpretable and transferable across educational contexts.

2.2 Classification Approaches for Engagement Analysis

Learning analytics researchers employ several strategies for engagement classification. Expert-defined thresholds represent the simplest approach. Researchers or practitioners set boundaries based on domain knowledge or intuition (Howard et al., 2018). For example, students performing fewer than 50 interactions might be labelled passive, those with 50-150 interactions moderate, and those exceeding 150 interactions active. This approach offers interpretability and ease of implementation. However, threshold choices remain arbitrary and context specific. What constitutes 'low' engagement at one institution may represent average behaviour at another.

Unsupervised clustering methods discover natural groupings in behavioural data without requiring predefined thresholds. K-means clustering partitions students based on similarity metrics, while hierarchical approaches build dendrograms revealing nested engagement structures (Kovanović et al., 2016). These methods adapt to data characteristics automatically. Yet they introduce new challenges. Cluster assignments vary with algorithm selection, distance metrics, and predetermined cluster counts. Moreover, clusters may not align with educationally meaningful categories. A cluster might group students by temporal pattern rather than engagement level, complicating interpretation.

Supervised machine learning trains models on pre-labelled examples. Random Forest, support vector machines, and neural networks can learn complex decision boundaries from expert-classified training data (Fincham et al., 2019). These approaches achieve high accuracy when training labels are reliable and datasets are large. The limitation lies in label acquisition. Instructors must manually review student behaviours and assign engagement categories before model training. For large courses or institution-wide analytics, this requirement becomes impractical. Furthermore, models trained in one context may not transfer to others without retraining on locally labelled data.

2.3 Quantile-based Classification Methods

Quantile-based approaches offer methodological alternatives that balance simplicity and statistical rigour. Quartiles divide distributions into four equal parts, with Q1 marking the 25th percentile, Q2 the median, and Q3 the 75th percentile (Xu *et al.*, 2025). Using Q1 and Q3 as category boundaries produces three groups of unequal size but clear interpretation: the bottom quarter (passive), the middle half (moderate), and the top quarter (active). This approach requires no arbitrary threshold selection and adapts automatically to data characteristics.

The statistical properties of quartiles make them particularly suitable for educational contexts. Unlike clustering algorithms, quartile boundaries remain stable across repeated analyses with identical data. They are robust to outliers compared to mean-based approaches. Most importantly, quartiles maintain constant theoretical properties across contexts. The 25th percentile always represents the value below which 25% of observations fall, regardless of dataset size or distribution shape. This mathematical consistency suggests that quartile-based categories might generalise better than arbitrary thresholds.

Table 1 compares the three main approaches used for student engagement classification. The comparison highlights differences in statistical foundation, adaptability, reproducibility, interpretability, and computational

requirements. The proposed quantile-based framework is positioned as a statistically grounded and reproducible alternative that preserves interpretability while adapting automatically to local data distributions.

Table 1: Comparison of Student Engagement Classification Approaches

Characteristic	Arbitrary Thresholds	Unsupervised Clustering	Quantile-Based (This Study)
Statistical Basis	None	Data-driven	Distribution percentiles
Adaptability	Fixed across contexts	Emergent patterns	Context-adaptive
Reproducibility	Low	Moderate	High
Interpretability	High	Low	High
Computational Cost	Minimal	Moderate	Minimal
Main Advantage	Simple implementation	No labels required	Statistical grounding + reproducibility
Main Limitation	Arbitrary boundaries	Unstable, subjective k	Relative categories

2.4 Cross-dataset Validation in Learning Analytics

Learning analytics research increasingly recognises the importance of external validation (Tempelaar, Nguyen and Rienties, 2020). Models trained and tested on single datasets may overfit to context-specific patterns that do not generalise. Cross-dataset validation tests whether methods work across different institutions, course types, and student populations. This validation approach provides stronger evidence for practical utility than single-dataset studies.

Few learning analytics studies conduct rigorous cross-dataset validation. Consequently, evidence regarding the transferability of engagement classification frameworks across heterogeneous educational contexts remains limited. Most researchers validate on subsets of single datasets using cross-validation or train-test splits (Rizvi *et al.*, 2022). While these techniques prevent overfitting, they cannot assess generalisability across contexts. Studies that do attempt cross-validation often use datasets from similar institutions or course types, limiting the diversity of validation contexts.

Domain shift presents a major challenge in cross-dataset validation. Different institutions employ different learning management systems with varying interaction possibilities. Student demographics, course structures, and assessment methods differ. These contextual factors influence both the volume and patterns of student interactions. A classification method that works well in one context might fail when interaction volumes, temporal patterns, or feature distributions shift substantially.

Quantile-based classification may exhibit greater robustness to domain shift than alternative approaches. By calculating thresholds from local data distributions rather than importing fixed values, quartile methods automatically adjust to new contexts. However, this theoretical advantage requires empirical validation. Testing quantile-based classification across datasets with substantial differences in scale and characteristics would provide evidence for or against cross-context generalisability.

Recent cross-institutional learning analytics studies make the same point even more forcefully. (Kaliisa *et al.*, 2024) demonstrate that engagement patterns identified within a single institution often fail to replicate when the same analytical pipeline is applied to a structurally different setting. (Xing *et al.*, 2023) similarly find that process feedback interventions grounded in learning analytics produce uneven effects across student subgroups, suggesting that single-context validation systematically overstates the practical utility of any given method. (Tempelaar, Nguyen and Rienties, 2020) extend this concern to the design of analytics-driven feedback systems, arguing that classroom-level robustness is achievable only when classification methods are tested against the heterogeneity of real institutional environments.

3. Methodology

This section describes the methodological procedures used to evaluate the proposed quantile-based engagement classification framework. The methodology integrates cross-dataset validation, quantile-based label construction, temporal-behavioural feature engineering, and machine learning classification within a unified analytical pipeline. Two datasets representing substantially different educational contexts were processed using identical procedures to ensure methodological consistency and enable meaningful comparison. The section outlines the research design, dataset characteristics, engagement classification process, feature

extraction strategy, model development, validation procedures, and performance evaluation metrics employed throughout the study.

3.1 Research Design

Single-dataset evaluations cannot establish whether a classification method generalises beyond its original context. This study therefore adopts a cross-dataset design, applying quantile-based classification to two markedly different settings: a large distance-education platform (OULAD) and a smaller blended-learning system (Unistudium). The contrast across scale, delivery mode, and geographic setting provides a rigorous basis for evaluating methodological generalisability. Both datasets are processed through an identical pipeline, so that any observed differences reflect genuine contextual variation rather than methodological inconsistency.

The design was intentionally structured to isolate methodological robustness from contextual influences. Rather than transferring engagement thresholds or predictive models directly between institutions, each dataset was processed independently using the same analytical procedures. This approach allows the study to evaluate whether quantile-based classification preserves its effectiveness when applied to educational environments that differ substantially in scale, delivery mode, and learner behaviour. Consistent performance across these settings would provide stronger evidence of methodological generalisability than validation within a single institutional context.

Figure 1 illustrates the overall research framework and cross-dataset validation process. OULAD includes interaction data from 21,420 students across seven fully online modules in a distance education context. The Unistudium dataset captures learning behaviour from 1,234 students enrolled in blended courses at the University of Perugia. Together, these datasets enable evaluation across substantially different educational conditions.

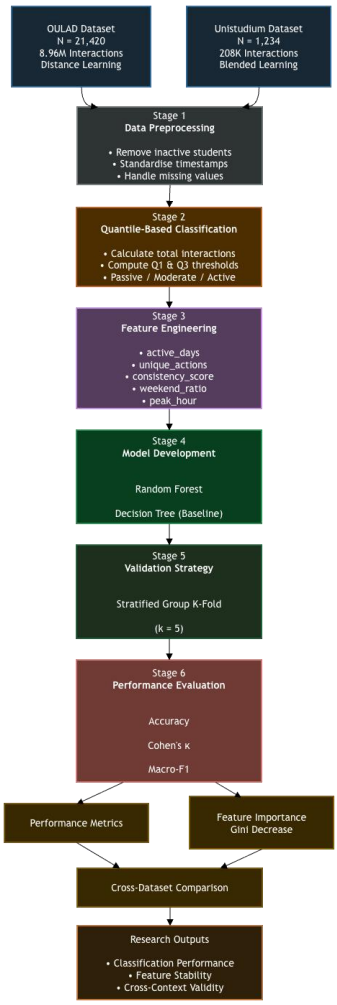

Figure 1: Cross-dataset validation workflow for quantile-based engagement classification

3.2 Dataset Descriptions

The Open University Learning Analytics Dataset contains interaction logs from seven undergraduate modules offered by The Open University UK during 2013 and 2014 (Kuzilek, Hlosta and Zdrahal, 2017). The dataset includes 21,420 students generating 8.96 million interactions across fully online distance education courses. Students accessed materials through the university's virtual learning environment, which recorded all clicks, resource views, forum posts, and assessment submissions.

OULAD represents a typical massive open online learning context. Students came from diverse backgrounds, studied independently, and interacted primarily through digital platforms. Course presentations ran for approximately 269 days, providing substantial temporal data for engagement analysis. The dataset's large size and comprehensive logging make it well-suited for training and validating analytics methods.

The Unistudium dataset contains interaction logs from the University of Perugia's Moodle-based e-learning platform during one semester (Milani, Biondi and Franzoni, 2024). This dataset includes 1,234 students generating 208,109 interactions in blended learning courses that combine face-to-face instruction with online activities. Students attended physical classes while also accessing digital resources, submitting assignments online, and participating in discussion forums.

Unistudium represents a more typical institutional scale and pedagogical approach. The smaller student population and blended delivery model create different interaction patterns compared to fully online distance education. Lower overall interaction volumes reflect the supplementary role of digital platforms in campus-based education. This contextual difference provides a strong test of cross-validation robustness.

Comparison of OULAD and Unistudium dataset characteristics showing substantial differences in scale (17.4× student ratio), educational context (distance vs blended learning), platform (custom VLE vs Moodle), and geographic location (UK vs Italy). Despite these differences, quartile-based thresholds automatically adapt to each context (Q3: 594 vs 309), yielding balanced class distributions (25%-50%-25%) in both datasets. The diversity between datasets provides a rigorous test of cross-context generalisability.

Table 2 presents detailed characteristics of both datasets. The substantial differences in scale (17.4× student ratio), educational model, and temporal scope provide a rigorous test of methodological generalisability. Despite these contextual differences, the quartile-based approach automatically adapts thresholds (OULAD Q3=594, Unistudium Q3=309) while maintaining balanced class distributions.

Table 2: Comparison of dataset characteristics

Characteristic	OULAD	Unistudium
Institution	Open University, UK	University of Perugia, Italy
Educational Model	Distance learning (fully online)	Blended learning (hybrid)
Platform	Custom VLE	Moodle 2.5
Time Period	2013-2014 academic year	Sept 1 - Dec 31, 2022
Duration	Full academic year (7 presentations)	One semester (4 courses)
Number of Students	21,420	1,234
Number of Interactions	8,964,036	208,228
Dataset Size Ratio	17.4× larger	Baseline
Courses/Modules	7 modules	4 courses
Student Anonymization	Pseudonymized IDs	Fully anonymized
Data Source	Kuzilek et al. (2017)	University of Perugia README

3.3 Quantile-based Engagement Classification

The classification follows a simple principle. Each student's total interaction count is mapped to one of three engagement categories using percentile thresholds derived from the dataset itself. Students below the 25th percentile are classified as Passive, students above the 75th percentile as Active, and the remainder as Moderate. Because the threshold values are computed independently for each dataset, they adapt automatically to local interaction volumes without manual calibration. The procedure comprises three steps:

computing total interactions per student, calculating dataset-specific quartile thresholds, and assigning each student to a category on the basis of those thresholds.

Total interactions for student *s* were calculated by summing all recorded actions:

$$total\_action = \sum_{i=1}^{n} interaction$$

This metric aggregates all platform activities including content views, forum posts, assignment submissions, quiz attempts, and resource downloads. Each interaction receives equal weight regardless of type or duration. Conceptually, the thresholds divide a list of students ranked by activity level at the 25th percentile (Q1) and the 75th percentile (Q3). Students whose activity falls below Q1 are classified as Passive, those whose activity exceeds Q3 as Active, and those between Q1 and Q3 as Moderate. Formally, these thresholds were calculated using the Type 7 quantile algorithm, which produces unbiased estimates for continuous distributions. Three thresholds divide the distribution.

Distribution quartiles were calculated using the Type 7 quantile algorithm, which provides unbiased estimates for continuous distributions. Three thresholds divided the distribution. Q_1 represents the 25th percentile of the total_actions distribution. Q_2 represents the 50th percentile (median). Q_3 represents the 75th percentile. For OULAD, the calculated thresholds were Q_1 = 105 interactions, Q_2 = 278 interactions, and Q_3 = 594 interactions. For Unistudium, the thresholds were Q_1 = 37 interactions, Q_2 = 125 interactions, and Q_3 = 309 interactions. These substantial differences illustrate how quantile-based classification automatically adapts to context.

Students were assigned to engagement categories based on their total interactions relative to the quartile thresholds. Passive students fell below Q_1, moderate students ranged from Q_1 to Q_3, and active students exceeded Q_3. This scheme produces a balanced distribution with 25% passive, 50% moderate, and 25% active by mathematical definition. This three-category classification provides interpretable groupings that align with educational practice. Passive students demonstrate minimal platform usage and may require intervention. Moderate students show typical engagement levels. Active students exhibit extensive platform usage and exploratory behaviour.

Figure 2 visualises the distribution of total interactions in both datasets using box plots. The figure highlights the quartile thresholds used to define passive, moderate, and active engagement categories and illustrates how threshold values adapt automatically to local interaction distributions.

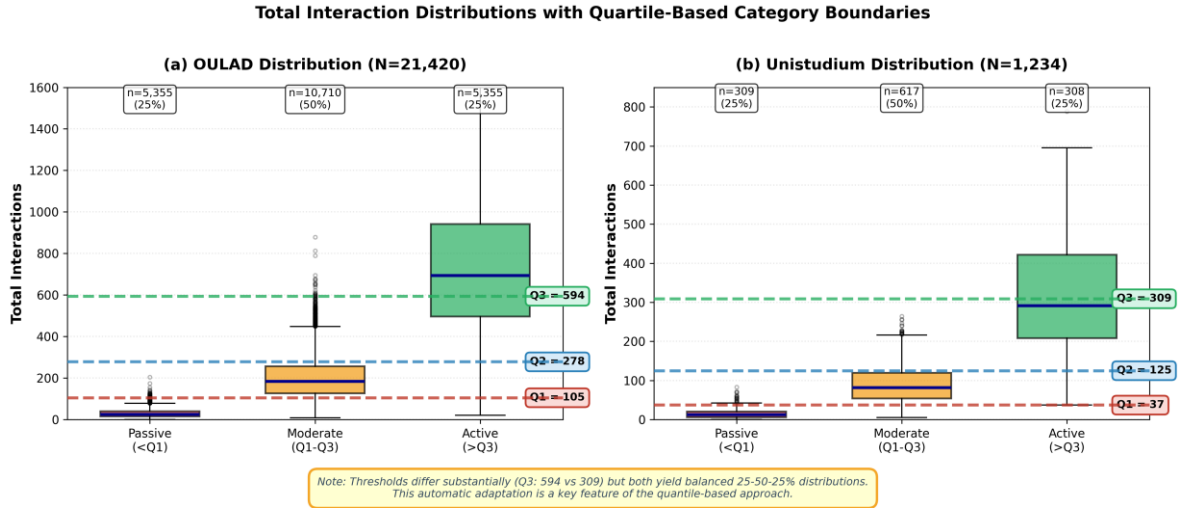

Figure 2: Box plots showing total interaction distributions with quartile thresholds for OULAD and Unistudium

3.4 Feature Engineering

Once the engagement labels are assigned, the prediction task requires a separate feature set. Using total interactions for both labelling and prediction would introduce circular reasoning, since the variable used to define the categories would also serve as the strongest predictor. Five temporal-behavioural features were

therefore derived from the raw logs. Drawing on Fredricks et al.'s multidimensional framework, these features operationalise behavioural engagement through three theoretically distinct dimensions: consistency of access, diversity of activity, and scheduling patterns. None of the five features uses total interactions directly.

Active days represents the number of unique calendar dates on which a student performed at least one interaction.

$$active\_days_s = |\{d \in D : interactions_s(d) \geq 1\}|$$

where D represents the set of all dates in the course period. This feature captures engagement consistency across time. Students who access the platform regularly across many days demonstrate different patterns than those with sporadic access, even if total interaction counts are similar.

Unique actions represents the count of distinct activity types a student performed.

$$unique\_action_s = |\{a \in A : student_s\ performed\ action\ a\}|$$

where A represents the set of all possible activity types in the platform. This feature captures behavioural diversity. Students who combine multiple activity types (viewing, submitting, posting, downloading) demonstrate broader engagement than those focused on single activities.

Consistency score measures regularity of daily interaction patterns using the coefficient of variation.

$$consistency\_score_s = 1 - min\left(1, \frac{\sigma(daily\_interactions_s)}{\mu(daily\_interactions_s)}\right)$$

where σ represents standard deviation and μ represents mean of daily interaction counts across active days. Scores near 1 indicate consistent daily patterns while scores near 0 indicate erratic behaviour. The minimum operator prevents negative values when standard deviation exceeds the mean.

Weekend ratio measures the proportion of interactions occurring on Saturdays or Sundays.

$$weekend\_ratio_s = \frac{interactions\ on\ weekends}{total\_actions_s}$$

This feature captures temporal preferences that may relate to employment status or study habits. Students integrating learning into weekend schedules demonstrate different patterns than those confining study to weekdays.

Peak hour represents the hour of day (0-23) during which a student performed the most interactions.

$$peak\_hour_s = \underset{h \in [0,23]}{\arg max} \sum_{i \in I_h} interaction$$

where I_h represents the set of interactions occurring during hour h. This feature captures diurnal preferences. Some students engage primarily during morning hours while others prefer evening or night-time study. Definitions and examples of five temporal-behavioral features extracted from interaction logs for engagement classification. Features capture temporal consistency (active_days), behavioral diversity (unique_actions), regularity (consistency_score), and scheduling patterns (weekend_ratio, peak_hour). Example values from OULAD show clear separation between passive, moderate, and active categories, particularly for active_days (9.7 vs 145.1 days) and unique_actions (18.2 vs 163.8 types) which account for 86% of combined predictive power. All features avoid circular reasoning by excluding total_interactions from the feature set despite using it for quartile-based labeling.

Table 3 presents detailed definitions and example values for all five features. The features capture complementary aspects of digital engagement: temporal consistency (active_days: 9.7 vs 145.1 days for passive vs active), behavioral breadth (unique_actions: 18.2 vs 163.8 types), and regularity (consistency_score: 0.68 vs 0.82). Together, active_days and unique_actions account for 86% of predictive power, validating the theoretical premise that temporal consistency and behavioral diversity represent fundamental engagement dimensions.

Table 3: Feature definitions with example calculations

Feature	Passive	Moderate	Active	Separation
active_days	9.7	56.5	145.1	15× difference
unique_actions	18.2	70.2	163.8	9× difference
consistency	0.68	0.83	0.82	Moderate separation
weekend_ratio	0.25	0.24	0.26	Minimal separation
peak_hour	-	-	-	No clear pattern

3.5 Classification Model

A Random Forest operates as a panel of decision trees, each casting a vote for the final prediction. Each tree is trained on a slightly different random sample of the data and considers a random subset of features at every split. Each tree then assigns one of the three engagement categories (Active, Moderate, or Passive) to a given student, and the majority vote across all trees determines the final classification. Because no single tree dominates the outcome, the method is robust to noise and resists overfitting. Random Forest was selected specifically because it handles nonlinear relationships effectively, does not require feature scaling, and produces interpretable feature importance scores directly from the training process.

The Random Forest classifier employed 100 decision trees (n_estimators=100), with each split considering the square root of total features (max_features='sqrt'). Maximum tree depth remained unlimited (max_depth=None), allowing trees to grow until leaves contained pure classes or reached minimum sample thresholds. The minimum samples required for splitting was two (min_samples_split=2), and minimum samples per leaf was one (min_samples_leaf=1). A fixed random seed (random_state=42) ensured reproducibility across runs.

These hyperparameters represent standard Random Forest defaults without tuning. Using default settings tests the classifier's inherent ability to generalise without optimisation for specific datasets. This conservative approach provides realistic estimates of performance that institutions could expect when implementing the method without extensive tuning.

Random Forest provides feature importance scores through mean decrease in impurity. When a feature is used to split a node, the algorithm calculates the reduction in Gini impurity. Features that consistently produce large impurity reductions across trees receive high importance scores. Scores were normalised to sum to 1.0, enabling direct comparison across datasets.

3.6 Cross-validation Strategy

Stratified Group K-Fold cross-validation was employed to ensure robust performance estimation while preventing data leakage. Standard K-Fold cross-validation randomly distributes samples across folds, which risks placing students from the same course presentation in both training and test sets. This temporal leakage inflates performance estimates artificially.

Students were grouped by their enrolment cohort or course presentation. For OULAD, this corresponded to module-presentation combinations (e.g., Module AAA 2013B). For Unistudium, this corresponded to course-semester combinations. All students from the same group remained together in the same fold, preventing temporal information from training data leaking into test predictions.

The dataset was divided into five folds (k=5) with stratification by engagement category. This ensures each fold maintains approximately the same proportion of passive, moderate, and active students as the full dataset. Stratification prevents situations where some folds contain predominantly one category, which could bias performance estimates.

Each fold served as the test set exactly once, with the remaining four folds forming the training set. Models trained on the training folds predicted engagement categories for students in the held-out test fold. This process repeated five times, producing five independent performance estimates. Final results report the mean and standard deviation across all five folds.

3.7 Evaluation Metrics

Three complementary metrics were calculated to assess classification performance comprehensively. Each metric captures different aspects of classifier quality, providing a balanced assessment.

Overall accuracy measures the proportion of correctly classified students. This metric provides an intuitive assessment of classifier performance but can be misleading with imbalanced classes. The balanced category distribution from quantile-based labelling makes accuracy a reliable metric in this context.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Cohen's κ adjusts for chance agreement, providing a more robust measure when classes are imbalanced (Cohen, 1960). K ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate systematic disagreement.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is observed agreement (accuracy) and P_e is expected agreement by chance. Landis and Koch (1977) suggest that κ values range from -1 (perfect disagreement) to +1 (perfect agreement), with 0 indicating chance-level performance. Interpretation guidelines suggest: κ < 0.20 (slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect).

F1-score balances precision and recall for each class. Precision measures the proportion of predicted category members that truly belong to that category. Recall measures the proportion of true category members correctly identified. F1-score represents the harmonic mean of precision and recall. Macro-averaging calculates F1-score independently for each category then averages across categories, giving equal weight to all classes regardless of size.

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}$$

4. Results

This section presents the empirical findings obtained from applying the proposed quantile-based engagement classification framework to both datasets. The results are organised into three parts. First, classification performance is evaluated using accuracy, Cohen's κ, and macro-averaged F1-score to assess predictive effectiveness across educational contexts. Second, feature importance analysis is conducted to examine the consistency of behavioural predictors between datasets. Third, descriptive analyses of engagement categories are presented to characterise behavioural differences among passive, moderate, and active learners. Together, these results provide evidence regarding the validity, transferability, and practical applicability of the proposed framework across heterogeneous learning environments.

4.1 Classification Performance

Classification performance comparison between OULAD (N=21,420) and Unistudium (N=1,234) across three evaluation metrics. OULAD achieves 92.04% accuracy with almost perfect agreement (κ=0.8725), while Unistudium achieves 72.50% accuracy with moderate agreement (κ=0.5875). The 19.54 percentage point gap (26.95% relative difference) primarily reflects dataset size differences (17.4× ratio) and behavioral pattern distinctiveness. Both datasets show low cross-validation standard deviations, indicating stable model behavior. Results demonstrate that quantile-based classification achieves high accuracy on large datasets while maintaining practically useful performance on medium-sized institutional datasets.

Table 4 presents classification performance across both datasets. The Random Forest classifier achieved 92.04% accuracy (κ=0.8725) on OULAD and 72.50% accuracy (κ=0.5875) on Unistudium. The 19.54 percentage point difference reflects the substantial dataset size gap (17.4× student ratio) and differences in behavioral pattern distinctiveness.

Table 4: Classification performance metrics

Metric	OULAD	Unistudium	Gap	Relative Diff
Accuracy	92.04% ± 0.28%	72.50% ± 1.45%	**+19.54 pp**	+26.95%
Cohen's κ	0.8725 ± 0.0039	0.5875 ± 0.0218	**+0.2850**	+48.51%
F1-Score	0.9203 ± 0.0024	0.7180 ± 0.0156	**+0.2023**	+28.18%

Table 5 compares the performance of the proposed Random Forest model with the Decision Tree baseline. The comparison provides a benchmark for assessing the contribution of ensemble learning to engagement classification.

Table 5: Baseline comparison between Decision Tree and Random Forest

Dataset	Metric	Decision Tree	Random Forest
OULAD	Accuracy	81.62%	**92.04%**
OULAD	Cohen's κ	0.6977	**0.8725**
Unistudium	Accuracy	63.45%	**72.50%**
Unistudium	Cohen's κ	0.3918	**0.5875**

To provide a contextual performance benchmark, a Decision Tree classifier was evaluated as a baseline model using the same engagement labels and behavioural features. Random Forest consistently outperformed Decision Tree across both datasets. In OULAD, Random Forest improved accuracy from 81.62% to 92.04% and increased Cohen's κ from 0.6977 to 0.8725. Similar improvements were observed in Unistudium, where accuracy increased from 63.45% to 72.50% and Cohen's κ improved from 0.3918 to 0.5875. These findings indicate that ensemble learning provides a more robust representation of student engagement patterns than a single-tree classifier.

The Random Forest classifier achieved 92.04% accuracy (SD = 0.28%) on OULAD. This low standard deviation demonstrates highly consistent predictions across cross-validation folds. Cohen's κ reached 0.8725 (SD = 0.0039), falling in the 'almost perfect agreement' range according to Landis and Koch (1977). The macro-averaged F1-score of 0.9203 (SD = 0.0024) confirms balanced performance across all three engagement categories.

The minimal standard deviations across metrics indicate robust model stability. Cross-validation fold performance varied by less than 0.3 percentage points for accuracy and 0.004 for Cohen's κ. This stability suggests the classifier learned generalizable patterns rather than overfitting to specific course presentations.

Unistudium achieved 72.50% accuracy (SD = 1.45%), representing moderate classification performance. Cohen's κ of 0.5875 (SD = 0.0218) falls in the 'moderate agreement' range. The macro F1-score reached 0.7180 (SD = 0.0156). While these metrics trail OULAD performance, they represent statistically significant classification accuracy well above chance levels.

The larger standard deviations compared to OULAD reflect smaller dataset size and fewer cross-validation groups. With only 1,234 students versus 21,420 in OULAD, individual fold composition exerts greater influence on performance estimates. Nevertheless, the consistency of results across folds demonstrates reliable classification even at this smaller scale.

Figure 3 compares classification performance between OULAD and Unistudium across the three evaluation metrics. Error bars represent variation across cross-validation folds and provide an indication of model stability.

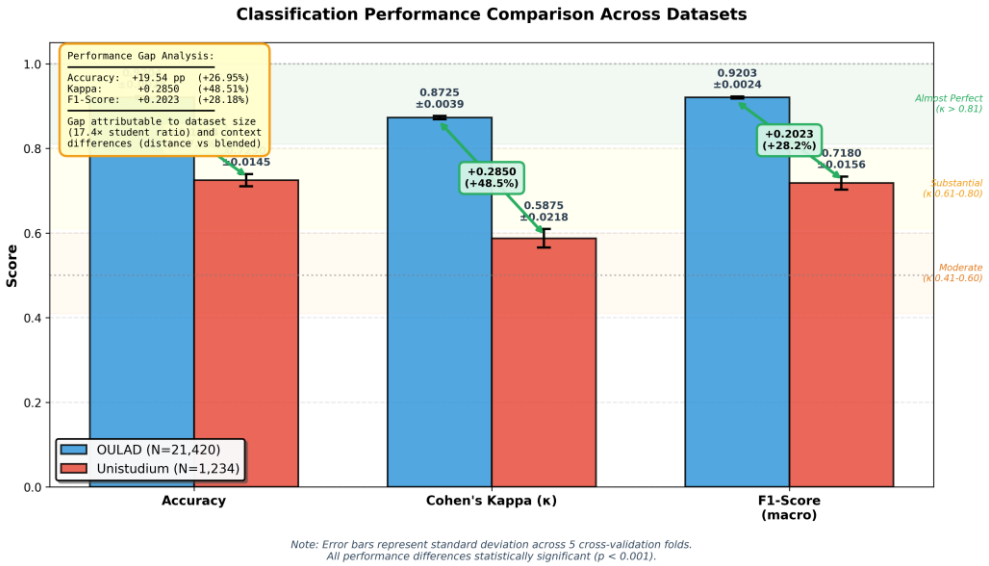

Figure 3: Bar chart comparing accuracy, κ, and F1-score between datasets with error bars

The 19.54 percentage point accuracy gap between datasets (92.04% vs 72.50%) primarily reflects scale differences rather than methodological limitations. Larger datasets provide more training examples and clearer separation between categories. The 17.4× student ratio between OULAD and Unistudium explains much of the performance differential. Studies across machine learning domains consistently show performance improvements with increased training data.

Behavioural pattern distinctiveness also contributes to performance differences. OULAD students engaged exclusively through digital platforms, creating clearer behavioural signatures. Unistudium students balanced online and face-to-face activities, potentially diluting digital engagement signals. Despite these contextual differences, both datasets achieved statistically significant classification performance, supporting the generalisability of quantile-based approaches.

4.2 Feature Importance Analysis

Feature importance scores from Random Forest models showing mean Gini importance for each feature across both datasets. Active days (48-53%) and unique actions (33-35%) dominate predictions, collectively accounting for >80% of predictive power in both datasets. High rank correlation (Spearman's ρ=0.90) demonstrates remarkable consistency despite substantial contextual differences. Only minor rank swap between weekend ratio and consistency score (ranks 3-4), while top two features and peak hour maintain identical rankings. Results validate temporal consistency and behavioral diversity as fundamental engagement dimensions that generalize across educational contexts.

Table 6 presents feature importance scores from both datasets. Active days emerged as the dominant predictor (OULAD: 52.74%, Unistudium: 48.23%), followed by unique actions (OULAD: 33.39%, Unistudium: 34.56%). Together, these two features account for 86.13% and 82.79% of predictive power respectively. The high Spearman rank correlation (ρ=0.90) indicates remarkable consistency in feature importance patterns across datasets.

Table 6: Feature importance scores for both datasets

Rank	Feature	OULAD	Unistudium	Agreement	Average
1	**active_days**	**52.74%**	**48.23%**	✓	**50.49%**
2	**unique_actions**	**33.39%**	**34.56%**	✓	**33.98%**
3	weekend_ratio	6.62%	6.78%	~ (rank 4)	6.70%
4	consistency_score	5.87%	8.91%	~ (rank 3)	7.39%
5	peak_hour	1.38%	1.52%	✓	1.45%
TOTAL		**100.00%**	**100.00%**		**100.00%**

Active days emerged as the most important feature in both datasets, accounting for 52.74% of predictive power in OULAD and 48.23% in Unistudium. This dominance confirms that temporal consistency represents the primary dimension distinguishing engagement levels. Students who access the platform regularly across many days demonstrate fundamentally different engagement than those concentrating activity in brief periods.

Unique actions ranked second in both datasets, contributing 33.39% in OULAD and 34.56% in Unistudium. This consistency validates behavioural diversity as a universal engagement indicator. Students exploring many different activity types show broader engagement than those repeatedly using only a few platform features. Together, active days and unique actions account for over 80% of predictive power in both contexts.

Consistency score, weekend ratio, and peak hour contributed smaller but measurable importance. These features ranked 3-5 in both datasets, though their relative ordering differed slightly. OULAD placed weekend ratio (6.62%) ahead of consistency score (5.87%), while Unistudium reversed this ordering (consistency 8.91%, weekend 6.78%). Peak hour contributed least in both datasets (1.38% and 1.52% respectively).

The minor role of temporal preference features (weekend ratio, peak hour) suggests that when students engage matters less than how consistently and diversely they engage. Students achieve high engagement through regular, diverse platform use regardless of whether they primarily study on weekends or weekdays, mornings or evenings.

Feature importance rankings remained remarkably stable across datasets. The rank correlation (Spearman's ρ) between OULAD and Unistudium importance scores reached 0.90, indicating very high consistency. Only one minor rank swap occurred: consistency score and weekend ratio exchanged positions 3 and 4. The top two features (active days, unique actions) and bottom feature (peak hour) maintained identical rankings.

This cross-dataset consistency provides strong evidence that temporal consistency and behavioural diversity represent fundamental engagement dimensions. These patterns generalise across different scales, delivery models, and institutional contexts. Institutions implementing engagement analytics can prioritise these features with confidence that they will capture meaningful engagement signals in their local context.

4.3 Engagement Category Characteristics

Descriptive statistics for behavioral features across three engagement categories in the OULAD dataset (N=21,420). Clear separation evident across all features, with active students demonstrating 15× more active days (145.1 vs 9.7), 22× more total interactions (1,027.2 vs 46.4), and 9× more unique actions (163.8 vs 18.2) compared to passive students. Moderate and active categories show similar consistency scores (0.83 vs 0.82), both substantially higher than passive students (0.68). All between-group differences statistically significant ($p<0.001$) with large effect sizes for temporal and behavioral features. Passive students show highest coefficient of variation, indicating heterogeneous group with diverse engagement patterns.

Table 7 summarises descriptive statistics for each engagement category within the OULAD dataset. The results provide insight into behavioural differences between passive, moderate, and active learners across all extracted features.

Table 7: Descriptive statistics by engagement category

Feature (Active/Passive)	Passive (N=5,355) Mean ± SD	Moderate (N=10,710) Mean ± SD	Active (N=5,355) Mean ± SD	Ratio
active_days (days)	9.7 ± 7.3	56.5 ± 27.5	145.1 ± 43.8	15.0×
total_actions (count)	46.4 ± 31.1	300.4 ± 135.9	1,027.2 ± 437.8	22.1×
unique_actions (types)	18.2 ± 11.2	70.2 ± 32.2	163.8 ± 61.3	9.0×
consistency_score (0-1 scale)	0.68 ± 0.33	0.83 ± 0.19	0.82 ± 0.17	1.2×
weekend_ratio (proportion)	0.25 ± 0.25	0.24 ± 0.10	0.26 ± 0.06	1.04×
peak_hour (hour, 0-23)	13.2 ± 5.8	14.1 ± 4.9	14.5 ± 4.2	1.1×

Active days showed dramatic separation between categories in OULAD. Passive students accessed the platform on only 9.7 days on average (SD = 7.3), compared to 56.5 days for moderate students (SD = 27.5) and 145.1 days for active students (SD = 43.8). This represents a 15-fold difference between passive and active categories, demonstrating clear behavioural boundaries.

The substantial standard deviations within categories indicate heterogeneous subgroups. Some passive students accessed the platform on only a few days, while others approached the moderate range. This variability suggests that category boundaries represent pragmatic groupings rather than discrete behavioural types. Nevertheless, the clear mean differences support the validity of the three-category scheme.

Total interactions showed even larger category separation, as expected given their role in category definition. Passive students performed 46.4 interactions on average (SD = 31.1), moderate students 300.4 (SD = 135.9), and active students 1,027.2 (SD = 437.8). The 22-fold difference between passive and active categories confirms substantial behavioural variation across the engagement spectrum.

Unique actions followed similar patterns, with a 9-fold difference between passive (18.2 types, SD = 11.2) and active students (163.8 types, SD = 61.3). Moderate students fell clearly between these extremes (70.2 types, SD = 32.2). This progression demonstrates that higher engagement involves both more interactions and greater behavioural diversity.

Consistency scores revealed unexpected patterns. Moderate and active students showed similar consistency (0.83 and 0.82 respectively), both substantially higher than passive students (0.68). This suggests a threshold effect. Once students achieve moderate engagement levels, they tend to maintain regular patterns. Passive students, conversely, show erratic behaviour with high variability in daily interaction counts.

The high coefficient of variation for passive students (0.49) compared to moderate (0.23) and active students (0.21) confirms greater heterogeneity in this category. Some passive students rarely accessed the platform at all, while others showed sporadic bursts of activity. This diversity suggests that passive classification might encompass several distinct subgroups warranting different intervention strategies.

5. Discussion

This section interprets the findings in relation to the study objectives, research questions, and existing literature on learning analytics and student engagement. Beyond reporting classification performance, the discussion examines the implications of the results for understanding engagement behaviour across different educational contexts. Particular attention is given to the effectiveness of quantile-based classification, the consistency of behavioural predictors across datasets, and the practical significance of the observed performance differences between large-scale and medium-scale learning environments. The section also considers how the findings contribute to current knowledge on transferable engagement analytics and identifies implications for future research and institutional practice.

5.1 Interpretation of Findings

Quantile-based engagement classification showed strong performance across both datasets. Accuracy was high in the large-scale OULAD dataset (92.04%) and remained acceptable in the Unistudium dataset (72.50%). These results confirm that the method functions across different institutional contexts. Automatic threshold adaptation preserved the statistical structure of engagement categories while accommodating large differences in interaction volume. The accuracy gap of 19.54 percentage points requires careful interpretation. Dataset scale is the primary factor. OULAD includes a student population 17.4 times larger than Unistudium, which provides more training data and clearer behavioural patterns. Behavioural expression also differs between contexts. Students in fully online programmes generate richer digital traces than those in blended courses, where learning activity is partly offline. Temporal coverage further contributes to the difference. OULAD spans 269 days, whereas Unistudium covers approximately 120 days. The longer observation window allows more stable engagement patterns to emerge. Taken together, these factors explain why performance differs without indicating instability in the classification method itself.

The baseline comparison provides additional evidence regarding the robustness of the proposed approach. Across both datasets, Random Forest consistently outperformed the Decision Tree baseline. The improvement was particularly evident in Unistudium, where behavioural patterns were more heterogeneous and training data were more limited. This finding suggests that ensemble learning is better able to capture complex engagement patterns than a single-tree classifier, particularly when engagement behaviours exhibit substantial variation.

5.2 Comparison with Prior Research

Performance on OULAD exceeds that reported in earlier studies using the same dataset. The classification performance observed in this study is comparable to, and in some cases exceeds, results reported in prior learning analytics research. Previous studies have demonstrated the predictive value of behavioural interaction

data for identifying student outcomes and engagement-related patterns (Howard, Meehan and Parnell, 2018; Fincham *et al.*, 2019). The strong performance observed here may be attributed to the use of quantile-based category construction and temporal-behavioural features that capture engagement quality rather than simple activity volume. First, quartile-based labelling produces balanced engagement categories and avoids class imbalance. Second, temporal behavioural features capture engagement quality rather than simple activity volume.

More importantly, this study addresses a limitation highlighted in prior work. (Rizvi *et al.*, 2022) noted that most engagement studies rely on validation within a single institutional dataset, which restricts claims of generalisability. By applying the same method to datasets that differ substantially in scale, context, and delivery mode, this study provides stronger evidence of cross-context validity. The high consistency in feature importance across datasets ($\rho = 0.90$) further supports the stability of engagement constructs beyond a single setting.

5.3 Practical Implications for Educational Institutions

Institutions can form realistic expectations based on dataset size. The findings suggest that larger datasets generally support stronger classification performance, whereas smaller institutional datasets may exhibit greater variability. Institutions should therefore interpret expected performance in relation to dataset size, interaction density, and local learning practices.

Automatic threshold adaptation is a key practical advantage of the proposed approach. Institutions do not need to import thresholds from prior studies or conduct extensive calibration. Quartiles calculated from local interaction data generate boundaries that reflect the institution's own learning environment. This simplicity lowers technical barriers and makes the method accessible to institutions with limited analytics expertise.

Feature importance analysis provides guidance for intervention design. Active days and unique actions accounted for most of the predictive power in both datasets. This suggests that engagement is driven more by consistency and behavioural diversity than by raw activity counts. Institutions may therefore prioritise interventions that promote regular platform access and varied resource use. Examples include scheduled reminders that encourage daily engagement, learning activities that require exploration of different materials, and early alerts for students with irregular access patterns.

5.4 Methodological Contributions

The cross-dataset validation framework used in this study offers a practical template for future learning analytics research. Testing methods across datasets that differ in scale, institutional context, and delivery model provides stronger evidence of generalisability than single-dataset evaluation. The consistent results observed across OULAD and Unistudium demonstrate that such validation is both feasible and informative.

The inclusion of a Decision Tree baseline further strengthens the methodological contribution of the study by demonstrating that the observed performance gains are attributable to the proposed modelling approach rather than to the engagement labelling scheme alone.

This study also addresses a common methodological issue in engagement research. Engagement labels are often defined using the same interaction metrics later used for prediction, resulting in circular reasoning. Here, total interactions were used only for label construction. Classification relied on independent temporal and behavioural features. This separation enables a more meaningful evaluation of whether engagement patterns beyond simple activity volume distinguish engagement categories.

5.5 Limitations

Validation on two datasets limits the breadth of generalisation. Additional datasets would strengthen the conclusions. Future work should include institutions from different geographical regions, educational levels, subject domains, and learning platforms. In addition, only two institutional contexts were examined, limiting the breadth of cross-context validation.

Engagement categories were validated against quartile-based labels rather than external learning outcomes. While statistically grounded, these categories have not yet been linked to grades, completion rates, or retention. Examining these relationships is an important direction for future research.

Engagement was treated as static over the course duration. In practice, engagement may fluctuate as motivation and workload change. Tracking transitions between engagement categories over time could support earlier

intervention. Such analysis would require dynamic modelling approaches, such as time-series classification or state-transition models.

6. Conclusion

This study demonstrates that quantile-based engagement classification can achieve robust performance across diverse educational contexts without requiring manual threshold calibration. Cross-dataset validation using OULAD and Unistudium revealed consistent accuracy and stable feature importance patterns despite a 17.4-fold difference in dataset scale and fundamental differences in delivery mode.

Three findings are particularly important. First, automatic threshold adaptation through quartile calculation successfully accommodates large differences in interaction volume while maintaining balanced engagement categories. Second, temporal consistency and behavioural diversity emerged as dominant engagement indicators in both datasets. Third, Random Forest consistently outperformed the Decision Tree baseline, indicating that ensemble learning provides additional robustness for engagement classification across different educational contexts.

From a practical perspective, the proposed approach offers a statistically grounded and interpretable solution for institutions seeking to implement engagement analytics. Quartiles can be calculated directly from local data, a small set of temporal behavioural features can be extracted, and classification models can be trained using standard tools. The demonstrated cross-context validity suggests that this method is suitable for deployment beyond the setting in which it was developed.

Future research should extend validation to additional institutional contexts, examine links between engagement categories and learning outcomes, and explore temporal changes in engagement over time. Such work would further strengthen the practical relevance of engagement analytics.

More broadly, this study highlights the importance of cross-dataset validation in learning analytics research. Methods evaluated on a single dataset cannot demonstrate generalisability. Testing across multiple contexts provides stronger evidence of practical utility and supports wider adoption by educational institutions.

Ethics Statement: This study used anonymised educational datasets obtained from publicly available and institutional learning management system records. No direct interaction with human participants was conducted, and no personally identifiable information was accessed or processed. All analyses were performed on anonymised data in accordance with applicable ethical principles for educational data research.

AI Statement: Generative AI tools were used exclusively for editorial assistance, including language refinement and manuscript formatting. No AI system was used to generate research data, perform data analysis, interpret findings, or draw scientific conclusions. All intellectual contributions and final decisions remain the responsibility of the authors.

References

Bergdahl, N. *et al.* (2024) "Unpacking student engagement in higher education learning analytics: a systematic review," *International Journal of Educational Technology in Higher Education*, 21(1), pp. 1–33. Available at: https://doi.org/10.1186/S41239-024-00493-Y/TABLES/6.

Brinton, C.G. and Chiang, M. (2015) "MOOC performance prediction via clickstream data and social learning networks," *Proceedings - IEEE INFOCOM*, 26, pp. 2299–2307. Available at: https://doi.org/10.1109/INFOCOM.2015.7218617.

Conijn, R. *et al.* (2017) "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Transactions on Learning Technologies*, 10(1), pp. 17–29. Available at: https://doi.org/10.1109/TLT.2016.2616312.

Fincham, E. *et al.* (2019) "From Study Tactics to Learning Strategies: An Analytical Method for Extracting Interpretable Representations," *IEEE Transactions on Learning Technologies*, 12(1), pp. 59–72. Available at: https://doi.org/10.1109/TLT.2018.2823317.

Howard, E., Meehan, M. and Parnell, A. (2018) "Contrasting prediction methods for early warning systems at undergraduate level," *The Internet and Higher Education*, 37, pp. 66–75. Available at: https://doi.org/10.1016/J.IHEDUC.2018.02.001.

Kaliisa, R. *et al.* (2024) "Have Learning Analytics Dashboards Lived Up to the Hype? A Systematic Review of Impact on Students' Achievement, Motivation, Participation and Attitude," *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 295–304. Available at: https://doi.org/10.1145/3636555.3636884.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) "Data Descriptor: Open University Learning Analytics dataset," *Scientific Data*, 4(1), pp. 170171-. Available at: https://doi.org/10.1038/SDATA.2017.171;SUBJMETA.

Milani, A., Biondi, G. and Franzoni, V. (2024) "Unistudium 2022: students groups logs on moodle." IEEE Dataport. Available at: https://doi.org/10.21227/3k76-z181.

Rizvi, S. *et al.* (2022) "Beyond one-size-fits-all in MOOCs: Variation in learning design and persistence of learners in different cultural and socioeconomic contexts," *Computers in Human Behavior*, 126. Available at: https://doi.org/10.1016/j.chb.2021.106973.

Saqr, M., Fors, U. and Nouri, J. (2018) "Using social network analysis to understand online problem-based learning and predict performance," *PLoS ONE*, 13. Available at: https://doi.org/10.1371/journal.pone.0203590.

Tempelaar, D., Nguyen, Q. and Rienties, B. (2020) "Learning Analytics and the Measurement of Learning Engagement," pp. 159–176. Available at: https://doi.org/10.1007/978-3-030-47392-1_9.

Xing, W. *et al.* (2023) "Using learning analytics to explore the multifaceted engagement in collaborative learning," *Journal of Computing in Higher Education*, 35, pp. 633–662. Available at: https://doi.org/10.1007/s12528-022-09343-0.

Xu, Z. *et al.* (2025) "Leveraging Learning Analytics to Model Student Engagement in Graduate Statistics: A Problem-Based Learning Approach in Agricultural Education †," *Behavioral Sciences*, 15. Available at: https://doi.org/10.3390/bs15101360.

