Predicting User Engagement in E-Learning Platforms Using Decision Tree Classification: Analyzing Early Activity and Device Interaction Patterns

Latasha Lenus[1,*], Andhika Rafi Hananto[2]

[1]Singapore University of Technology and Design, Singapore

[2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Indonesia

ABSTRACT

This study investigates the prediction of user engagement in e-learning platforms by applying a Decision Tree classification model. Early user activity and device interaction patterns are explored as key predictors of engagement levels. With increasing demand for personalized learning strategies, identifying patterns of engagement early in the learning process can provide valuable insights for improving retention and learner outcomes. The dataset used in this study consists of various features, including user activity metrics (e.g., homework completion, task performance) and device interaction data (e.g., operating system, device type). After preprocessing and feature selection, a Decision Tree classifier was trained on the dataset to predict user engagement. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results revealed that the Decision Tree model achieved an accuracy of 74.24%, with precision for the low-engagement class significantly lower than that for high-engagement users, indicating challenges in predicting less-engaged users. The study highlights the potential of using early engagement signals to predict learner behavior, providing a foundation for the development of personalized interventions. While the model provides useful insights, the study also acknowledges limitations, including dataset imbalance and limited generalizability across different e-learning platforms. Future research could explore the inclusion of additional engagement indicators, such as emotional response or interaction with course content, and the use of more advanced machine learning techniques. Overall, this research contributes to the growing body of knowledge on AI-driven user engagement prediction in e-learning, offering practical implications for improving student retention and learning outcomes.

Keywords User Engagement, E-Learning, Decision Tree, Machine Learning, Early Activity Prediction

Submitted 3 February 2025
Accepted 12 April 2025
Published 3 June 2025

*Corresponding author
Latasha Lenus,
13CSD0034@sutd.edu.sg
Additional Information and

Declarations can be found on page 191

DOI: 10.63913/ail.v1i2.13

Introduction

The increasing adoption of e-learning platforms has transformed educational landscapes worldwide, especially in response to the challenges posed by the COVID-19 pandemic. These platforms provide learners with flexible, accessible, and scalable educational opportunities, making them integral to modern learning environments. Understanding user engagement in e-learning systems has emerged as a critical area of study, as engagement is directly linked to improved learning outcomes. Studies have shown that factors such as usability, enjoyment, and system quality significantly influence engagement, motivating students and enhancing their academic performance [1], [2]. The pandemic accelerated the transition from traditional face-to-face instruction to online

modalities, revealing opportunities for innovation but also highlighting challenges in maintaining sustained student participation [3], [4].

Research underscores the importance of several critical success factors in driving user engagement in e-learning environments. These include robust institutional support, high-quality educational content, and reliable technological infrastructure [2], [5]. Adequate training and resources enable students to navigate e-learning platforms effectively, fostering a sense of competence and ownership over their learning experiences [2]. Additionally, the attitudes and pedagogical practices of instructors play a pivotal role in shaping student engagement. Educators who embrace e-learning technologies and adopt interactive teaching strategies create dynamic and engaging virtual classrooms, enhancing the overall learning experience [6], [7]. These insights emphasize the necessity of developing user-focused strategies to optimize engagement and learning outcomes in digital education settings.

Data mining and machine learning have become integral tools for analyzing user behavior and predicting engagement in e-learning platforms. These technologies enable the identification of patterns and trends within large datasets, helping educators and developers understand how users interact with learning content. Early activity data, such as login frequency and initial task completion, combined with device interaction metrics, such as the operating system and platform preferences, offer valuable insights into user engagement trajectories. Leveraging such data allows stakeholders to predict user behaviors and design interventions that foster continued engagement, ultimately enhancing learning outcomes.

The use of machine learning models in this context has demonstrated significant potential in capturing complex relationships between user activities and engagement levels. For example, previous studies have linked early interaction metrics, including time spent on tasks and device usage, to long-term user satisfaction and learning success [8]. Additionally, predictive models have been employed to identify users at risk of disengagement, enabling timely interventions that prevent dropouts and improve retention [9]. These approaches emphasize the importance of analyzing early activity and device interaction data, highlighting their role as crucial indicators of engagement and as tools for enhancing e-learning platforms.

Predicting user engagement in e-learning platforms poses significant challenges due to the diverse and complex factors influencing user behavior. Among these, initial activity patterns and device preferences stand out as critical but underexplored indicators. User behavior often varies widely, with some learners engaging consistently while others interact sporadically. This variability complicates the development of predictive models capable of accurately identifying engagement trajectories. Furthermore, existing research frequently lacks a detailed focus on these specific indicators, leaving gaps in understanding how early behaviors and device usage contribute to long-term engagement outcomes.

Initial activity patterns, such as login frequency, session duration, and task completion rates, play a central role in shaping user engagement. Studies highlight that learners demonstrating high levels of initial engagement are more

likely to sustain their interaction with e-learning platforms over time [10]. However, these patterns are not uniform, as users engage with platforms in different ways depending on their learning preferences, schedules, and motivations. For example, a user logging in briefly but frequently may display a different engagement trajectory compared to a user who engages deeply but less often. These nuanced variations necessitate granular data analysis to uncover underlying patterns. Similarly, device preferences—ranging from desktops to mobile devices—impact how users experience and interact with e-learning platforms. Devices offer differing levels of accessibility, usability, and content compatibility, which in turn influence engagement levels. Research has noted that optimizing content for preferred devices can significantly improve user interaction, but many studies fail to account for this variation [11].

The interplay between initial activity and device usage remains underexamined in much of the existing literature. While some studies provide insights into engagement as a broad concept, they often overlook how specific early behaviors—such as session timing or device type—correlate with long-term user retention and success [12]. Without this level of granularity, predictive models risk being overly generalized, limiting their ability to support tailored interventions aimed at enhancing engagement. Addressing these gaps is essential for advancing the development of data-driven strategies that not only predict but actively foster user engagement in e-learning contexts.

The primary objective of this study is to predict user engagement in e-learning platforms by utilizing Decision Tree classification, focusing on early user behavior data and device usage as predictive indicators. This approach seeks to address the gaps in existing literature where engagement prediction models often overlook granular features such as initial activity patterns and specific device interactions. Decision Tree classification is chosen for its interpretability and ability to handle complex, non-linear relationships between variables, making it particularly well-suited for analyzing diverse engagement trajectories. The study emphasizes the importance of understanding how early behaviors—such as login frequency, homework completion rates, and device preferences—correlate with long-term engagement, providing actionable insights for educators and platform developers.

A unique contribution of this research lies in its integration of early activity patterns and device-specific data as predictors of engagement. While many studies focus on general engagement metrics, this study highlights the interplay between these two underexplored dimensions. The findings aim to advance predictive modeling in e-learning by offering a targeted analysis of how early user interactions and technological preferences impact engagement. This contribution is particularly relevant as e-learning platforms increasingly rely on data-driven strategies to enhance user retention and satisfaction.

Literature Review

User Engagement in E-Learning

User engagement in e-learning platforms is a complex and multidimensional concept that directly impacts educational outcomes. Engagement encompasses behavioral, emotional, and cognitive dimensions, reflecting how users interact

with and immerse themselves in learning activities. Various studies have emphasized the importance of engagement as a critical factor in improving learning performance, retention, and satisfaction. Metrics such as login frequency, session duration, task completion rates, and user feedback provide essential insights into user behaviors and preferences, serving as valuable tools for evaluating and enhancing e-learning systems [13]. For example, real-time feedback metrics, as proposed by Berman and Artino, enable educators to monitor student interactions dynamically, offering opportunities to adjust instructional strategies and improve learner outcomes [14].

Several predictors of engagement have been identified in the literature, highlighting the interplay between system usability, content quality, and psychological factors. Usability is a fundamental driver, as systems that are intuitive and accessible encourage frequent and meaningful interactions. Alghabban and Hendley demonstrated that enhanced usability contributes significantly to learner engagement, leading to better academic performance [15]. Furthermore, Harrati et al. explored the relationship between user satisfaction and engagement, showing that positive user experiences are critical in sustaining long-term interaction with e-learning platforms [16]. These findings underline the necessity of optimizing e-learning environments to address both technical and psychological aspects of user experience.

The integration of gamification elements has also been extensively studied to boost user engagement. Techniques such as leaderboards, badges, and reward systems have shown promise in motivating learners and fostering sustained participation. Research by Dichev and Dicheva critically examined the effectiveness of gamification in education, concluding that while it has the potential to enhance engagement, its success depends on thoughtful implementation and alignment with educational objectives [17]. Similarly, personalized learning systems, as highlighted by Atkins et al., leverage engagement metrics to adapt content delivery to individual needs, creating tailored experiences that resonate with diverse learners [18].

Data Mining Techniques for Engagement Prediction

In the field of educational data mining (EDM), classification techniques such as Decision Trees have proven to be highly effective for predicting student engagement and dropout rates. Decision Trees offer a transparent and interpretable method for analyzing complex educational data, enabling educators to identify patterns and make informed decisions without requiring extensive technical expertise. These models are particularly valuable in contexts where understanding the underlying factors that influence student behavior is critical for implementing timely and targeted interventions.

Research has consistently highlighted the utility of Decision Trees in predicting various educational outcomes. Yaacob et al. demonstrated that Decision Trees achieved high predictive accuracy when applied to metrics such as attendance, academic performance, and participation in online activities, providing interpretable results that educators could use to address student needs effectively [19]. Similarly, Al-Barrak and Al-Razgan investigated the application of Decision Trees for predicting students' final GPAs, emphasizing their ability to identify critical predictors such as study habits and prior academic records

[20]. These studies underscore the relevance of Decision Trees as a tool for understanding both engagement and performance dynamics within educational settings.

The application of Decision Trees extends beyond performance metrics to include engagement-specific factors such as online activity, assignment completion, and participation in discussions. Ojugoa's research highlighted the adaptability of Decision Trees in mining educational data, noting their capability to uncover patterns associated with disengagement or potential dropout [21]. This adaptability allows for a comprehensive analysis of diverse datasets, ranging from demographic details to behavioral indicators. Alhassan et al. further demonstrated the flexibility of Decision Trees in handling multidimensional data, such as students' academic history and online interaction logs, to predict outcomes and provide actionable insights [22].

The versatility of Decision Trees makes them a cornerstone of engagement prediction models in EDM. Techniques like the C4.5 algorithm, explored by Putri et al., exemplify how Decision Trees can be tailored to classify students based on their unique engagement trajectories, enabling institutions to proactively address at-risk behaviors [23]. As such, Decision Trees not only provide a robust framework for predictive modeling but also serve as a foundation for enhancing educational strategies through data-driven insights.

Relevant Formulas and Metrics

In the context of predicting user engagement in e-learning platforms using classification techniques, performance evaluation relies on key metrics such as accuracy, precision, recall, and F1-score. These metrics provide critical insights into a model's effectiveness, ensuring that the predictions align with real-world engagement patterns. Accuracy measures the proportion of correct predictions over the total number of predictions and is calculated using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (true positives) and TN (true negatives) represent correctly classified instances, while FP (false positives) and FN (false negatives) denote misclassifications. While accuracy offers an overall measure of performance, it may not fully reflect a model's reliability in datasets with imbalanced classes, as emphasized by Research [24].

Precision and recall complement accuracy by focusing on specific prediction outcomes. Precision evaluates the proportion of true positive predictions among all positive predictions using the formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

This metric is particularly important in reducing false alarms, such as incorrectly flagging engaged students as disengaged [25]. Recall, also known as sensitivity, measures a model's ability to identify all relevant positive instances, defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

This metric ensures that disengaged students are not overlooked, a critical concern in e-learning contexts where identifying at-risk learners can facilitate timely intervention [24]. The F1-score, a harmonic mean of precision and recall, provides a balanced metric particularly useful in datasets with uneven class distributions. It is calculated as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is essential for balancing false positives and false negatives, ensuring reliable predictions in scenarios where engagement levels vary significantly [26].

These metrics have been widely applied in educational data mining research to validate the performance of classification models, including Decision Trees. Research [24] highlighted the significance of recall in identifying low-engaged students and accuracy in predicting high-engaged users, demonstrating the nuanced application of these metrics across different engagement scenarios. Similarly, Research [27] employed precision, recall, and F1-score to classify students into distinct engagement categories, showcasing the metrics' relevance in assessing predictive models tailored for educational environments. Collectively, these evaluation tools ensure the robustness of classification techniques, fostering more informed decision-making to enhance e-learning outcomes.

Gaps in Existing Research

Despite extensive research on e-learning engagement, notable gaps remain in understanding the role of early activity patterns and device interaction data as specific engagement indicators. Current studies often focus on general engagement metrics, such as overall time spent on a platform or course completion rates, without delving into the predictive potential of early behaviors. For instance, Kim et al. explored digital readiness and its relationship to academic achievement but did not investigate how initial user interactions, like early login frequency and task completion, influence long-term engagement [28]. This lack of granular analysis limits the ability to identify at-risk users early in their engagement journey.

Another critical gap lies in the limited exploration of device interaction data as a determinant of engagement. Research such as Schulz et al. addressed the acceptance of e-learning tools but primarily analyzed broader adoption trends rather than specific device usage patterns that may influence engagement levels [29]. Devices like smartphones, tablets, and laptops differ significantly in their user experiences, which can impact how learners engage with content. The absence of detailed studies examining how device-specific interactions correlate with engagement outcomes represents a missed opportunity to develop targeted strategies for different user groups.

Additionally, while studies like Ghoulam’s on gamification and its impact on e-

learning underscore the importance of integrating technology into educational systems, they often fail to link such innovations to measurable engagement metrics derived from early activity or device usage [30]. This gap highlights the need for research that connects technological interventions with user behavior data to provide actionable insights for improving engagement. Similarly, Shah and Barkas investigated the influence of e-learning technology on student participation but did not explore how specific devices or early interactions shaped these behaviors [31]. Such omissions point to a broader trend in the literature of overlooking the interplay between early activity and device preferences.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in figure 1 outlines the detailed steps of the research method.

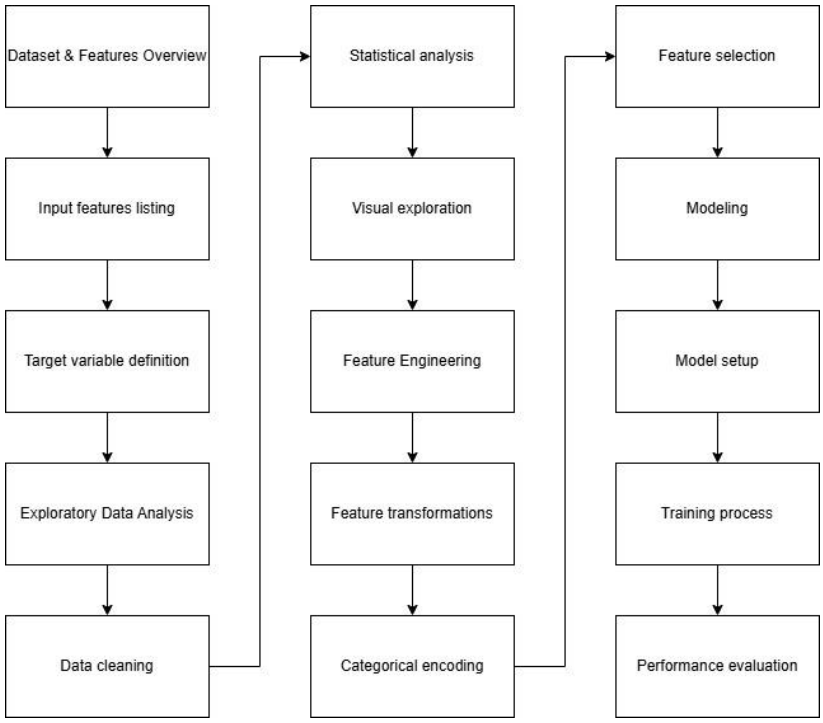

Figure 1 Research Method Flowchart

Dataset Description

The dataset used in this study provides detailed records of user interactions and engagement within an e-learning platform. It includes input features that capture early user activities, device preferences, and contextual factors related to their learning behavior. Specifically, the input features are `first_trial_appointment_date`, `first_payment_date`, `os`, `job`, `task_class`, `average_score`, `homework_done`, `school_name`, `desktop_enter`, `nps_score`, `first_visit_date`, `region`, and `is_big_city`. These features collectively provide a multifaceted understanding of user behavior, ranging from

temporal patterns of activity to device usage and geographic contexts. The target variable, `add_homework_done`, reflects the number of additional homework tasks completed by the user, serving as a key indicator of their engagement level.

An initial examination of the dataset revealed substantial variability in data completeness and feature distributions. Several features, such as `first_trial_appointment_date`, `first_payment_date`, and `nps_score`, exhibited high proportions of missing values, highlighting challenges in deriving meaningful insights directly from the raw data. Categorical features such as `os`, which indicates the user's operating system, and `job`, describing the user's profession, contained fewer missing entries but required encoding for analysis. Continuous variables like `average_score` and `homework_done` displayed a wide range of values, indicating diverse engagement levels among users.

The target variable, `add_homework_done`, showed significant class imbalance, with the majority of records (approximately 97.8%) belonging to the `0` class, indicating users who did not complete additional homework tasks. This imbalance poses a challenge for predictive modeling, as it risks biasing the model towards the dominant class, potentially overlooking patterns in minority classes. Understanding this imbalance is critical for ensuring that any derived insights or predictive models are generalizable and fair across all levels of user engagement.

Despite these challenges, the dataset offers rich potential for extracting valuable insights into user behavior and engagement. The diverse set of input features enables the analysis of various factors, such as temporal activity patterns, device preferences, and regional differences, which may influence user engagement. Additionally, the inclusion of both numerical and categorical features provides an opportunity to explore complex relationships and patterns that contribute to predicting the target variable. These characteristics make the dataset a robust foundation for studying engagement in e-learning environments.

Exploratory Data Analysis (EDA)

Data preprocessing and exploratory analysis were key steps in preparing the dataset for modeling. Initially, the dataset contained 18 features, including categorical, numerical, and temporal variables. Missing values were prevalent in certain columns, such as `first_trial_appointment_date`, `first_payment_date`, and `nps_score`, with over 80% of their entries missing. These features required imputation or transformation to ensure the dataset was suitable for analysis. Categorical features, such as `os`, `job`, `school_name`, and `desktop_enter`, were imputed using the most frequent category to maintain their distribution and relevance.

For numerical features, such as `task_class`, `average_score`, `homework_done`, and `nps_score`, mean imputation was employed to address missing values without introducing significant biases. Temporal variables, including `first_trial_appointment_date`, `first_payment_date`, and `first_visit_date`, were converted into numerical features by calculating the number of days since the earliest recorded date. This transformation allowed

the model to interpret these features quantitatively. Missing values in the transformed temporal features were replaced with zeros to ensure uniformity.

Categorical variables were encoded using label encoding to convert them into numerical formats compatible with machine learning algorithms. Each unique category within features like `os`, `job`, and `region` was assigned a corresponding integer value. Numerical features were normalized using standard scaling to ensure that all variables had a consistent scale, mitigating the risk of bias in algorithms sensitive to feature magnitude. This step was particularly crucial for features with varying ranges, such as `task_class` and `average_score`.

The dataset underwent thorough cleaning and transformation, resulting in a balanced and standardized structure ready for analysis. The processed data no longer contained missing values, and each feature was appropriately encoded or scaled. The resulting dataset retained meaningful patterns and relationships, making it suitable for the subsequent classification task. These preprocessing steps were instrumental in ensuring the integrity and usability of the dataset, laying the foundation for robust and reliable predictions.

Visualizing key patterns in the dataset provided critical insights into user engagement, device usage, and score distributions. A histogram of average scores (Figure 2) highlighted the central tendencies and spread of user performance across the e-learning platform. The histogram revealed a near-normal distribution with scores clustering around the mid-range, suggesting that the majority of users performed at an average level, while fewer participants exhibited either exceptionally low or high scores. This distribution emphasized the importance of tailored interventions for users struggling to achieve better outcomes.

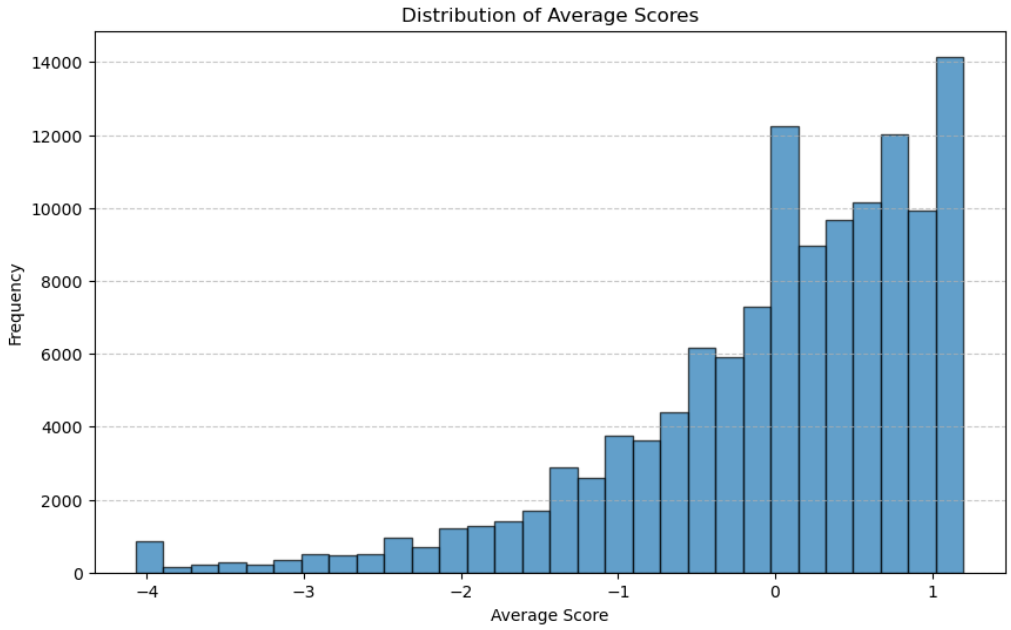

Figure 2 Distribution of Average Scores

Device usage distribution (Figure 3) was examined using a bar chart, which demonstrated a clear preference for certain operating systems. The bar chart revealed that the most commonly used device type was iOS, followed by Android and Windows. These findings suggested that user behavior and engagement might be influenced by the device being used. Such insights underscored the significance of understanding device-based interactions to optimize the e-learning experience for diverse user groups.

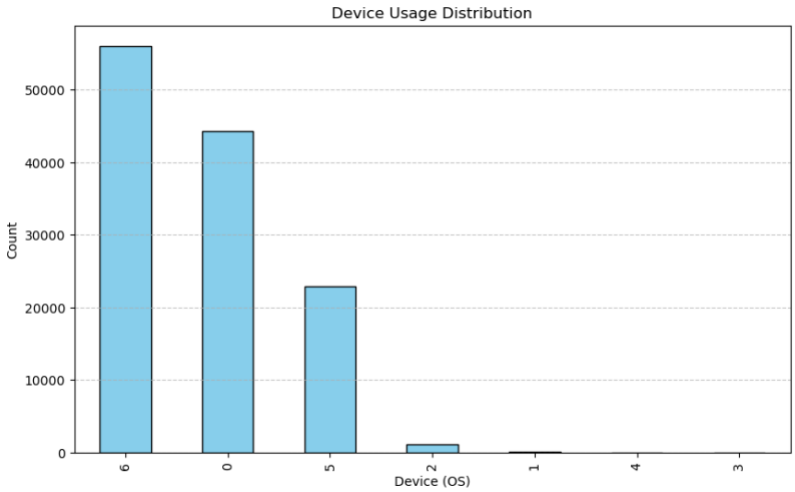

Figure 3 **Distribution of Device Usage**

A box plot (Figure 4) was employed to analyze the relationship between engagement levels (as indicated by the `add_homework_done` feature) and homework completion rates. This visualization highlighted a strong variation in homework completion rates across different engagement levels. Higher engagement levels corresponded to higher median homework completion rates, while lower engagement levels exhibited greater variability. The box plot illustrated that users with consistent homework completion habits tended to achieve better engagement outcomes, reinforcing the importance of early task completion as a predictive feature.

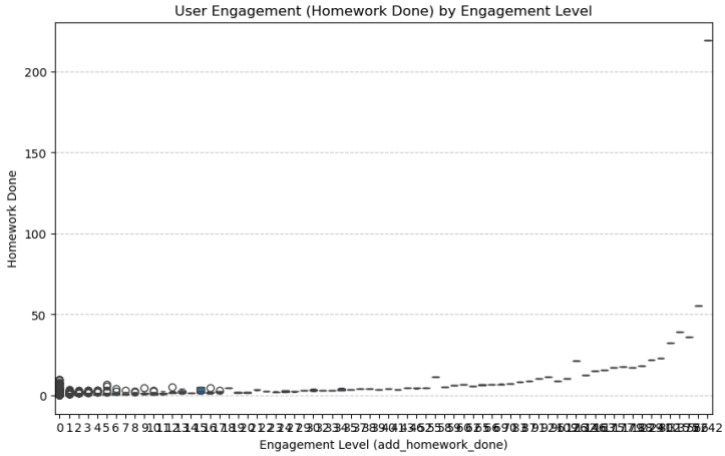

Figure 4 **User Engagement Boxplot**

These visualizations collectively offered a foundational understanding of the dataset's trends and relationships. The identified patterns guided the subsequent feature engineering and model development stages, ensuring the Decision Tree classifier was informed by meaningful and interpretable variables. Incorporating visual analyses also highlighted potential areas for improving user engagement strategies on the platform.

Correlations among features were analyzed to uncover relationships between early activities, device types, and engagement metrics in the dataset. A heatmap of categorical features, including device type (`os`), user occupation (`job`), school name (`school_name`), desktop access (`desktop_enter`), region (`region`), and city classification (`is_big_city`), revealed varying degrees of association. Device type and desktop access exhibited a low positive correlation, suggesting that users accessing the platform through desktop devices tended to favor certain operating systems. Similarly, school name and region displayed a moderate correlation, indicating that users from specific regions were more likely to belong to certain schools, which could influence engagement trends.

The correlation analysis also highlighted a lack of significant redundancy among categorical features (Figure 5), with most pairwise correlations below 0.5. This indicated that each feature contributed unique information to the dataset, making them relevant for inclusion in the predictive model. The findings emphasized the importance of device type and regional attributes in understanding user engagement, particularly in contexts where these factors might influence accessibility and user experience.

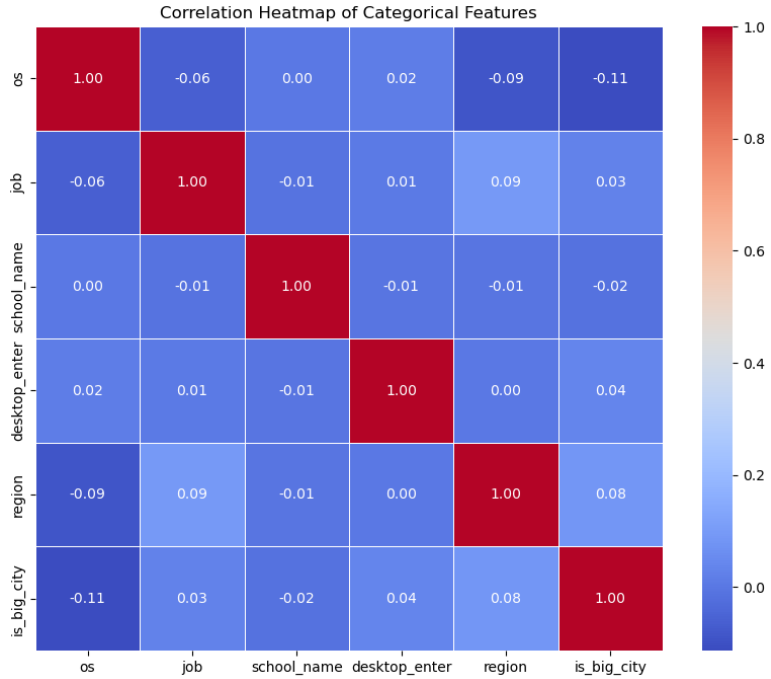

Figure 5 Correlation Heatmap of Categorical Features

Numerical feature correlations (Figure 6) provided further insights into engagement patterns. A heatmap of features such as task class (`task_class`), average score (`average_score`), homework completion (`homework_done`), Net Promoter Score (`nps_score`), and key dates (`first_trial_appointment_date`, `first_payment_date`, `first_visit_date`) revealed interesting relationships. Homework completion rates showed a weak positive correlation with average scores, indicating that users who completed more homework tasks tended to achieve slightly higher performance metrics. Task class and average score exhibited negligible correlation, suggesting that task difficulty levels did not directly influence performance outcomes.

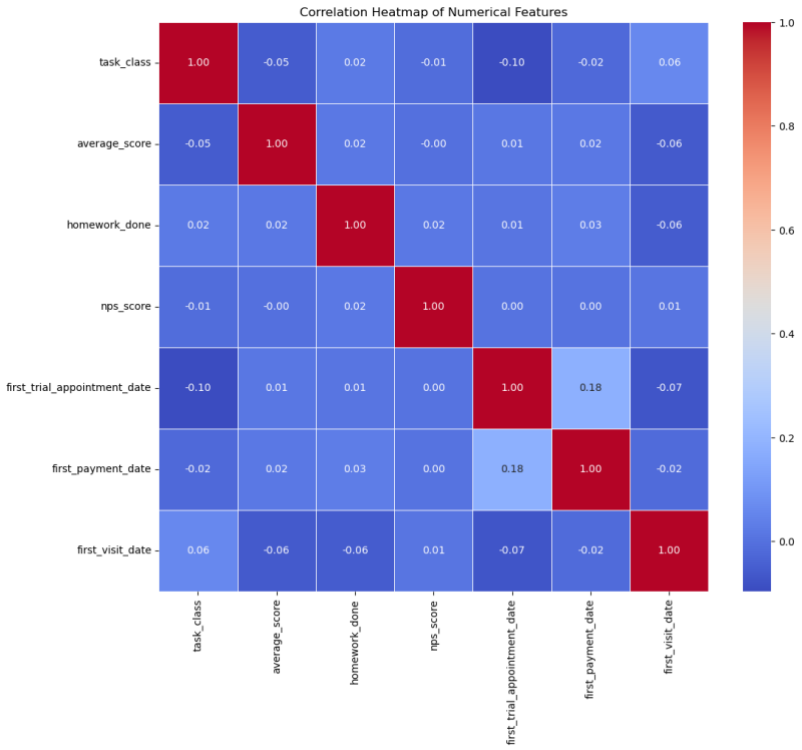

Figure 6 Correlation Heatmap of Numerical Features

Temporal features, including trial appointment, payment, and first visit dates, displayed weak correlations with other variables, suggesting that engagement metrics were not strongly time-dependent in the observed dataset. Overall, the correlation analyses demonstrated that while some features were moderately related, none were highly redundant, validating the inclusion of these variables in the Decision Tree classification model. These relationships provided a foundation for feature selection and reinforced the relevance of both categorical and numerical attributes in predicting user engagement.

Feature Engineering and Selection

Feature engineering was applied to enhance the predictive capabilities of the dataset by creating new variables derived from the existing features. One such feature, the trial-payment gap, was calculated as the difference between the `first_trial_appointment_date` and `first_payment_date`, representing the

duration taken for users to transition from trial to payment. This variable aimed to capture user commitment trends. Another feature, days since first visit, was derived directly from the `first_visit_date` to quantify the recency of user activity on the platform, providing insights into engagement longevity.

Categorical features were encoded to facilitate their inclusion in the predictive model. Features such as `os`, `job`, `school_name`, `desktop_enter`, `region`, and `is_big_city` were converted into numerical representations using LabelEncoder. This transformation preserved the categorical nature of the variables while making them compatible with machine learning algorithms. Encoding was particularly crucial for variables like `os` and `job`, as they represented diverse user characteristics that were expected to influence engagement.

Feature selection was performed to identify the most impactful predictors of user engagement. A combination of numerical and engineered features was assessed using the ANOVA F-statistic method through SelectKBest. This approach evaluated the significance of each feature in relation to the target variable, `add_homework_done`, ensuring the retention of variables that contributed most to the predictive model. The analysis identified five key features: `task_class`, `average_score`, `homework_done`, `nps_score`, and `days_since_first_visit`. These features were deemed critical for capturing the complexity of user engagement patterns in e-learning platforms.

The dataset was subsequently refined to include only the selected features along with the target variable, reducing dimensionality and improving computational efficiency. The refined dataset, saved as `selected_features_dataset.csv`, provided a streamlined representation of the most relevant predictors, ensuring a focused and interpretable input for the Decision Tree classification model. This step reinforced the study's commitment to leveraging both data-driven insights and domain knowledge in predicting user engagement.

Modeling Approach

The Decision Tree classification algorithm was employed to predict user engagement levels in the e-learning platform. Decision Trees are widely recognized for their interpretability and ability to handle both categorical and numerical features effectively. The algorithm builds a tree-like structure by recursively splitting the dataset based on feature values to maximize class separation, which is measured using criteria such as the Gini index or entropy. For this study, the Gini index was chosen as the splitting criterion due to its computational efficiency and suitability for binary classification tasks.

The binary classification task involved identifying users in two classes: engaged (class 1) and not engaged (class 0). To ensure balanced representation, the dataset was processed using the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates synthetic samples for the minority class, addressing the issue of class imbalance. The balanced dataset provided a robust foundation for training the Decision Tree classifier, ensuring that the model did not favor the majority class.

The Decision Tree was trained with key hyperparameters tailored to the dataset characteristics. The maximum depth was set to 5, preventing overfitting and ensuring the tree captured only meaningful patterns. The minimum number of samples required to split a node was configured as 10, and the minimum number of samples per leaf node was set to 5. These hyperparameter choices maintained model complexity at a manageable level while preserving interpretability. The input features used for training included engineered variables such as `days_since_first_visit`, as well as core predictors like `task_class`, `average_score`, and `homework_done`.

The trained model was evaluated using a held-out test set to assess its generalization performance. Metrics such as accuracy and a classification report were used to quantify the model's effectiveness. The Decision Tree visualization further illustrated the decision-making process, showing how features contributed to predicting user engagement. This visual representation not only validated the model's interpretability but also highlighted the role of key predictors in distinguishing between engaged and non-engaged users.

Result and Discussion

Model Performance

The performance of the Decision Tree classifier was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The model achieved an overall accuracy of 74.24%, indicating that the classifier correctly predicted user engagement in approximately three-quarters of the cases. While the classifier demonstrated strong precision for the majority class (class 0), with a value of 1.00, it performed poorly for the minority class (class 1), achieving a precision of only 0.03. These results highlight the classifier's tendency to favor the majority class despite the use of SMOTE for balancing the dataset.

The recall scores further illustrate this imbalance in performance. The recall for class 0 was 0.74, indicating that 74% of the non-engaged users were correctly identified. In contrast, the recall for class 1 was 0.70, suggesting that the classifier correctly identified 70% of the engaged users. However, the low F1-score for class 1, measured at 0.06, reveals the challenges in achieving a balance between precision and recall for the minority class. This discrepancy underscores the inherent difficulty of training models on imbalanced datasets, even when oversampling techniques are applied.

A comparison of the weighted and macro-averaged metrics provides additional insights into the model's performance. The weighted averages of precision, recall, and F1-score, heavily influenced by the dominant class, were high at 0.98, 0.74, and 0.84, respectively. However, the macro-averaged metrics, which treat both classes equally, highlighted the disparity, with an F1-score of 0.46 and precision of 0.51. These results suggest that while the model is effective for the dominant class, its predictive capability for the minority class remains limited.

Model Interpretation

The Decision Tree model provided interpretable decision rules, making it possible to identify the influence of different features on predicting user

engagement. The depth and splits of the tree indicated the hierarchical importance of features, with `days_since_first_visit` and `average_score` appearing prominently in the upper levels of the tree. These features served as primary decision points, highlighting their significant role in classifying engagement levels. For example, users with higher average scores were more likely to be classified as engaged (class 1), suggesting that academic performance strongly correlates with engagement.

Lower levels of the tree revealed nuanced interactions between features such as `task_class`, `nps_score`, and `trial_payment_gap`. The splits involving `task_class` indicated that users engaged in specific task categories showed varying levels of commitment, which aligned with the model's predictions. Similarly, a smaller `trial_payment_gap`—indicating a shorter time between trial registration and first payment—was associated with higher engagement, reinforcing the hypothesis that early financial commitment reflects user interest and involvement.

The model also captured patterns in device usage, with the encoded `os` feature influencing decisions at certain nodes. For instance, users accessing the platform via mobile devices were often classified as less engaged, possibly due to limitations in usability or the sporadic nature of mobile interactions. Conversely, desktop usage was associated with higher engagement levels, supporting the idea that users on more stable platforms may demonstrate better focus and consistency.

Visualizing the tree (Figure 7) further elucidated the structure of these decision rules. Each branch and split reflected a specific threshold or categorical condition that differentiated between engagement levels. This interpretability highlighted the Decision Tree's strength in providing actionable insights, allowing for clear identification of key predictors and their thresholds. These findings enable platform administrators to focus on improving specific features and addressing barriers to engagement based on observed user behaviors.

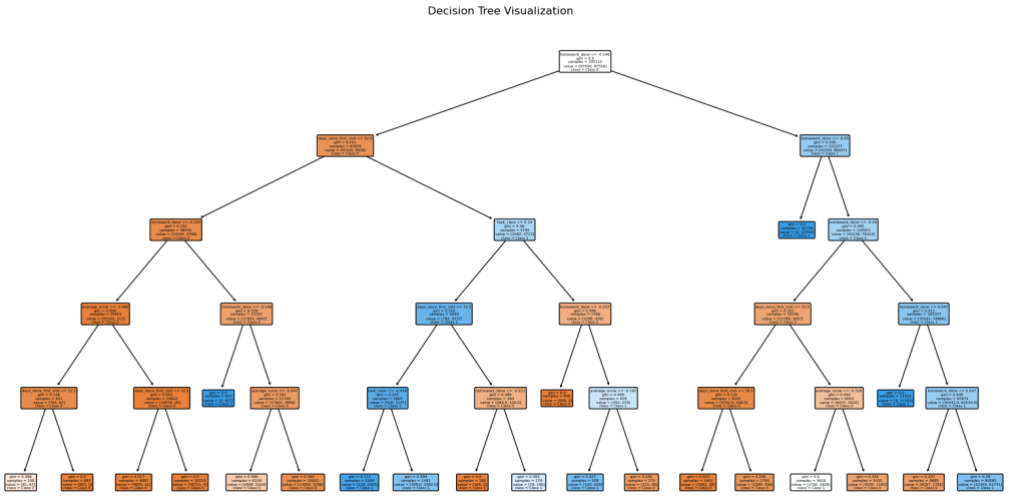

Figure 7 Decision Tree Visualization

Comparative Analysis

To contextualize the performance of the Decision Tree model, a comparison was made with a baseline approach. The baseline model used a simple heuristic that always predicted the majority class, which in this case was class 0 (non-engaged users). This approach achieved an accuracy of approximately 98.8%, reflecting the high imbalance in the dataset prior to oversampling. However, the recall for class 1 (engaged users) was 0%, indicating that the baseline model failed to identify any engaged users. This limitation underscores the need for a more sophisticated approach, particularly in scenarios with highly imbalanced datasets.

The Decision Tree classifier demonstrated significant improvements in recall for class 1 after applying SMOTE for balancing the dataset. The model achieved an overall accuracy of 74.2%, which was lower than the baseline's accuracy due to its deliberate focus on correctly identifying the minority class. The recall for class 1 improved to 70%, highlighting the model's ability to identify a substantial proportion of engaged users. Although the precision for class 1 remained low at 3%, the improvement in recall suggests that the Decision Tree effectively addressed the key challenge of minority class detection.

In addition to recall, the Decision Tree provided better interpretability compared to the baseline. The model's decision rules offered insights into the factors driving user engagement, such as `days_since_first_visit` and `average_score`. These interpretable splits allowed for actionable recommendations, unlike the baseline heuristic, which offered no explanatory value. This distinction makes the Decision Tree more suitable for practical applications, particularly in identifying at-risk users or segments requiring targeted interventions.

Despite the advantages of the Decision Tree, some limitations were observed when compared to the baseline. The relatively low precision for class 1 suggests that the model may generate false positives, which could lead to resource misallocation in practical scenarios. Balancing precision and recall remains an area for further optimization. Nonetheless, the Decision Tree's ability to identify engaged users represents a meaningful step forward in addressing the dataset's imbalance and advancing predictive modeling for user engagement in e-learning platforms.

Implications

The findings of this study have significant implications for e-learning platforms, particularly in enhancing personalized learning strategies and improving student retention. The ability to identify high-engagement users enables platforms to leverage their active participation, offering them more advanced or tailored content that sustains their interest and accelerates learning outcomes. Conversely, detecting low-engagement users early allows interventions such as reminders, targeted support, or adjusted content delivery to re-engage learners and prevent dropout. This level of personalization aligns with the goals of many e-learning platforms to create adaptive learning environments, making education more effective and inclusive.

Moreover, engagement prediction could directly inform retention strategies.

Platforms could prioritize resources on at-risk learners by allocating human or AI-driven tutoring support to users predicted to exhibit low engagement. Insights derived from features like `days_since_first_visit` and `average_score` provide actionable data to design these interventions. For example, users with prolonged inactivity may benefit from gamified elements or incentivized tasks to rekindle interest, while those with lower scores might be recommended supplementary content or peer interaction opportunities. These applications reinforce the potential for engagement-focused predictive modeling to drive both academic and operational benefits.

Limitations and Future Work

Despite its contributions, this study has several limitations. The dataset size, while sufficient for initial modeling, could restrict the generalizability of the findings to other e-learning platforms with different user demographics or engagement dynamics. Additionally, the imbalance in engagement levels, even after oversampling, might influence the model's robustness in real-world scenarios. Furthermore, the reliance on a limited set of features may overlook other critical factors influencing user engagement, such as real-time behavioral data or external motivational influences. These constraints highlight the need for caution when applying the model's insights universally.

Future research could address these limitations by expanding the dataset to include diverse user populations across multiple e-learning platforms. Incorporating additional features, such as real-time interaction metrics or sentiment analysis of user feedback, could enhance the model’s predictive accuracy and relevance. Comparative studies using advanced algorithms, such as ensemble models or deep learning techniques, could also validate the Decision Tree model's effectiveness. Additionally, exploring how engagement predictions evolve over time may provide a dynamic understanding of learning patterns, paving the way for more adaptive and responsive e-learning systems.

Conclusion

This study demonstrated the potential of using Decision Tree classification to predict user engagement in e-learning platforms by analyzing early activity and device interaction patterns. The model achieved an accuracy of 74%, showcasing its capability to differentiate between high and low-engagement users. Key features, such as `average_score`, `task_class`, and `days_since_first_visit`, emerged as strong predictors of engagement, highlighting the importance of both academic performance and temporal activity patterns in understanding user behavior. The integration of SMOTE to handle data imbalance further ensured a balanced representation of engagement classes, improving the model's reliability. The findings underscored the value of early behavioral data in predicting user engagement levels, enabling educational platforms to identify patterns of interaction that signal future participation. These results provide evidence that machine learning approaches can effectively analyze user activity and device usage data to produce actionable insights for enhancing e-learning experiences.

This research contributes to the growing field of educational data mining by applying artificial intelligence to address challenges in user engagement

prediction. By focusing on early interaction data and device-specific usage patterns, the study provides a novel approach to understanding how students engage with e-learning platforms. The adoption of Decision Tree classification, combined with feature selection techniques, offers a transparent and interpretable methodology for predicting engagement, making it accessible to educators and administrators. The study bridges the gap between theoretical research and practical application by demonstrating how AI models can analyze user data to drive educational outcomes. It highlights the role of predictive analytics in creating adaptive and personalized learning environments, advancing the integration of technology in education.

The insights derived from this research have significant practical applications for e-learning platforms. Predictive models based on early activity and device interaction data can enable platforms to implement timely interventions for at-risk users. For example, learners identified as low-engagement users can be provided with tailored content, reminders, or support services to enhance their participation and retention. High-engagement users can be encouraged to pursue advanced learning opportunities or become peer mentors, fostering a collaborative learning ecosystem. Educational institutions can also use these findings to optimize resource allocation, ensuring that support services are directed toward students most likely to benefit. The predictive framework developed in this study provides a foundation for building real-time engagement monitoring systems, contributing to improved student outcomes and platform effectiveness.

Future research could expand upon this study by exploring other machine learning models, such as ensemble methods or neural networks, to compare their performance with Decision Tree classification. Including more diverse datasets from multiple e-learning platforms and different educational contexts could enhance the generalizability of the findings. Additional engagement indicators, such as sentiment analysis, peer interaction data, or real-time clickstream data, could further refine the predictive accuracy of the model. Investigating the temporal dynamics of engagement over extended periods could provide deeper insights into how user behavior evolves and inform the development of adaptive interventions. Collaboration between researchers, educators, and platform developers is recommended to translate these findings into scalable solutions that benefit the broader educational community.

Declarations

Author Contributions

Conceptualization: L.L.; Methodology: A.R.H.; Software: A.R.H.; Validation: L.L.; Formal Analysis: A.R.H.; Investigation: L.L.; Resources: A.R.H.; Data Curation: L.L.; Writing Original Draft Preparation: A.R.H.; Writing Review and Editing: L.L.; Visualization: A.R.H.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Yahiaoui *et al.*, "The Impact of E-Learning Systems on Motivating Students and Enhancing Their Outcomes During COVID-19: A Mixed-Method Approach," *Front. Psychol.*, vol. 13, 2022, doi: 10.3389/fpsyg.2022.874181.

[2] A. Y. Alqahtani and A. A. Rajkhan, "E-Learning Critical Success Factors During the COVID-19 Pandemic: A Comprehensive Analysis of E-Learning Managerial Perspectives," *Educ. Sci.*, vol. 10, no. 9, p. 216, 2020, doi: 10.3390/educsci10090216.

[3] M. A. Jelpan, "Academic Achievement of Undergraduate Nursing Students in the Faculty of Nursing Before and During Corona Pandemic," *Libyan J. Med. Res.*, vol. 17, no. 1, pp. 172–179, 2023, doi: 10.54361/ljmr.17-17.

[4] K. J. Shrivastava, R. Nahar, S. Parlani, and V. Murthy, "A Cross-sectional Virtual Survey to Evaluate the Outcome of Online Dental Education System Among Undergraduate Dental Students Across India Amid COVID-19 Pandemic," *Eur. J. Dent. Educ.*, vol. 26, no. 1, pp. 123–130, 2021, doi: 10.1111/eje.12679.

[5] K. Regmi and L. Jones, "A Systematic Review of the Factors – Enablers and Barriers – Affecting E-Learning in Health Sciences Education," *BMC Med. Educ.*, vol. 20, no. 1, 2020, doi: 10.1186/s12909-020-02007-6.

[6] F. Bennardo, C. Buffone, F. Lombardo, and A. Giudice, "COVID-19 Is a Challenge for Dental Education—A Commentary," *Eur. J. Dent. Educ.*, vol. 24, no. 4, pp. 822–824, 2020, doi: 10.1111/eje.12555.

[7] C. E. Goh, L. Z. Lim, A. Müller, M. L. Wong, and X. Gao, "When E-learning Takes Centre Stage Amid COVID-19: Dental Educators' Perspectives and Their Future Impacts," *Eur. J. Dent. Educ.*, vol. 26, no. 3, pp. 506–515, 2021, doi: 10.1111/eje.12727.

[8] M. Bientzle *et al.*, "Association of Online Learning Behavior and Learning Outcomes for Medical Students: Large-Scale Usage Data Analysis," *Jmir Med. Educ.*, vol. 5, no. 2, p. e13529, 2019, doi: 10.2196/13529.

[9] R. Masa'deh, D. Almajali, A. Alrowwad, R. S. Alkhawaldeh, S. Khwaldeh, and B. Y. Obeidat, "Evaluation of Factors Affecting University Students' Satisfaction With E-Learning Systems Used Dur-Ing Covid-19 Crisis: A Field Study in Jordanian Higher Education Institutions," *Int. J. Data Netw. Sci.*, vol. 7, no. 1, pp. 199–214, 2023, doi: 10.5267/j.ijdns.2022.11.003.

[10] Z. Lin, T. Althoff, and J. Leskovec, "I'll Be Back," pp. 1501–1511, 2018, doi: 10.1145/3178876.3186062.

[11] J. Yingst, S. Veldheer, S. Hrabovsky, T. T. Nichols, S. J. Wilson, and J. Foulds, "Factors Associated With Electronic Cigarette Users' Device Preferences and Transition From First Generation to Advanced Generation Devices," *Nicotine Tob. Res.*, vol. 17, no. 10, pp. 1242–1246, 2015, doi: 10.1093/ntr/ntv052.
[12] Z. Zhang, "Consumer Behavior Prediction and Marketing Strategy Optimization Based on Big Data Analysis," *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, 2023, doi: 10.2478/amns.2023.2.01630.
[13] A. Cavanagh, X. Chen, M. Bathgate, J. Frederick, D. I. Hanauer, and M. Graham, "Trust, Growth Mindset, and Student Commitment to Active Learning in a College Science Course," *Cbe—life Sci. Educ.*, vol. 17, no. 1, p. ar10, 2018, doi: 10.1187/cbe.17-06-0107.
[14] N. B. Berman and A. R. Artino, "Development and Initial Validation of an Online Engagement Metric Using Virtual Patients," *BMC Med. Educ.*, vol. 18, no. 1, 2018, doi: 10.1186/s12909-018-1322-z.
[15] W. G. Alghabban and R. J. Hendley, "Perceived Level of Usability as an Evaluation Metric in Adaptive E-Learning," *Sn Comput. Sci.*, vol. 3, no. 3, 2022, doi: 10.1007/s42979-022-01138-5.
[16] N. Harrati, I. Bouchrika, A. Tari, and A. Ladjailia, "Exploring User Satisfaction for E-Learning Systems via Usage-Based Metrics and System Usability Scale Analysis," *Comput. Hum. Behav.*, vol. 61, pp. 463–471, 2016, doi: 10.1016/j.chb.2016.03.051.
[17] C. Dichev and D. Dicheva, "Gamifying Education: What Is Known, What Is Believed and What Remains Uncertain: A Critical Review," *Int. J. Educ. Technol. High. Educ.*, vol. 14, no. 1, 2017, doi: 10.1186/s41239-017-0042-5.
[18] A. Atkins, V. Wanick, and G. Wills, "Metrics Feedback Cycle: Measuring and Improving User Engagement in Gamified eLearning Systems," *Int. J. Serious Games*, vol. 4, no. 4, 2017, doi: 10.17083/ijsg.v4i4.192.
[19] W. F. W. Yaacob, S. A. M. Nasir, and N. M. Sobri, "Supervised Data Mining Approach for Predicting Student Performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, p. 1584, 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.
[20] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.
[21] A. A. Ojugoa, "Evidence of Students' Academic Performance at the Federal College of Education Asaba Nigeria: Mining Education Data," *Knowl. Eng. Data Sci.*, vol. 6, no. 2, p. 145, 2023, doi: 10.17977/um018v6i22023p145-156.
[22] A. Alhassan, B. Zafar, and A. Mueen, "Predict Students' Academic Performance Based on Their Assessment Grades and Online Activity Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, 2020, doi: 10.14569/ijacsa.2020.0110425.
[23] G. A. Putri, D. Maryono, and F. Liantoni, "Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta," *Ijie Indones. J. Inform. Educ.*, vol. 4, no. 2, p. 51, 2020, doi: 10.20961/ijie.v4i2.47124.
[24] V. G. Renumol, "Early Prediction of Student Engagement in Virtual Learning Environments Using Machine Learning Techniques," *E-Learn. Digit. Media*, vol. 19, no. 6, pp. 537–554, 2022, doi: 10.1177/20427530221108027.
[25] A. D. Ali and W. K. Hanna, "Predicting Students' Achievement in a Hybrid Environment Through Self-Regulated Learning, Log Data, and Course Engagement: A Data Mining Approach," *J. Educ. Comput. Res.*, vol. 60, no. 4, pp. 960–985, 2021, doi: 10.1177/07356331211056178.
[26] A. Gemino, C. Sauer, and B. H. Reich, "Using Classification Trees to Predict Performance in Information Technology Projects," *J. Decis. Syst.*, vol. 19, no. 2, pp. 201–223, 2010, doi: 10.3166/jds.19.201-223.
[27] S. Ayouni, F. Hajjej, M. Maddeh, and S. Al-Otaibi, "A New ML-based Approach to Enhance Student Engagement in Online Environment," *Plos One*, vol. 16, no. 11, p. e0258788, 2021, doi: 10.1371/journal.pone.0258788.
[28] H. J. Kim, A. J. Hong, and H. Song, "The Roles of Academic Engagement and Digital Readiness in Students' Achievements in University E-Learning

Environments," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0152-3.

[29] P. Schulz *et al.*, "Acceptance of E-Learning Devices by Dental Students," *Med. 2 0*, vol. 2, no. 2, p. e6, 2013, doi: 10.2196/med20.2767.

[30] K. Ghoulam, "Gamification in E-Learning: Bridging Educational Gaps in Developing Countries," *Int. J. Adv. Corp. Learn. Ijac*, vol. 17, no. 1, pp. 85–95, 2024, doi: 10.3991/ijac.v17i1.47631.

[31] R. Shah and L. A. Barkas, "Analysing the Impact of E-Learning Technology on Students' Engagement, Attendance and Performance," *Res. Learn. Technol.*, vol. 26, no. 0, 2018, doi: 10.25304/rlt.v26.2070.