Leveraging machine learning and clickstream data to improve student performance prediction in virtual learning environments

Zakaria Khoudi, Nasreddine Hafidi and Mourad Nachaoui
Equipe Mathématiques et Interactions, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni Mellal, Morocco, and
Soufiane Lyaqini
LIPIM Laboratory, USMS ENSA Khouribga, Khouribga, Morocco

Abstract

Purpose – The purpose of this research is to evaluate the utility of clickstream data and machine learning algorithms in predicting student performance and enhancing online learning experiences. By leveraging clickstream data and machine learning algorithms, the study aims to predict student performance accurately, enabling timely and personalized interventions. This approach seeks to reduce high failure and dropout rates in online courses, ultimately enhancing educational outcomes and preserving the reputation of educational institutions.

Design/methodology/approach – This study utilizes clickstream data from the Open University Learning Analytics Data set (OULAD) to predict student performance in virtual learning environments. The approach involves extracting and organizing data into weekly and monthly interactions. Various machine learning models, including traditional methods (Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, XGBoost) and advanced time-series models (LSTM-XGBoost, GRU), are employed to analyze the data. The GRU model demonstrated the highest accuracy, offering insights into student engagement and learning patterns.

Findings – The study reveals that integrating clickstream data with machine learning models provides a robust framework for predicting student performance in virtual learning environments. Among the methods tested, the GRU algorithm outperformed six baseline models, achieving an accuracy of 90.13%. These findings underscore the effectiveness of using advanced time-series models to monitor and improve student engagement and success rates in online education.

Originality/value – This research introduces a novel approach to student performance prediction by combining traditional and advanced time-series machine learning models with clickstream data. The study's originality lies in its comprehensive analysis of both weekly and monthly student interactions, providing educators with a powerful tool for early intervention. The findings contribute to the growing body of literature on learning analytics, offering practical solutions to enhance online education's effectiveness and reduce dropout rates.

Keywords Clickstream data, Students at risk, VIF, Student performance prediction, Machine learning, Virtual learning environment, GRU algorithm

Paper type Research paper

1. Introduction

Online learning has emerged as a transformative mode of instruction in higher education, driven by advancements in computer and network technologies. The COVID-19 pandemic further accelerated its adoption, particularly in 2020, as most teaching and learning activities shifted online. In developing regions, online learning offers significant advantages, such as overcoming time and location barriers, reducing educational costs and expanding access for individuals unable to attend traditional in-person classes (Ouyang *et al.*, 2022; Rizvi *et al.*, 2019). However, despite its potential, online learning faces significant challenges, including high failure and dropout rates. Studies indicate alarmingly low completion rates for online courses, with some reporting rates below 15%, and in extreme cases, as low as 5% (Hlosta *et al.*, 2017). For instance, research by He *et al.* (2015) revealed that only 795 out of 51,306 students enrolled in online courses successfully completed their programs, resulting in a completion rate of just 1.5%. These high failure and dropout rates not only undermine the effectiveness of online learning but also tarnish the reputation of educational institutions. To address these challenges, early identification of at-risk students through accurate performance prediction is critical. By leveraging predictive techniques, institutions can allocate resources more effectively, providing timely interventions and support to reduce the likelihood of failure (Huang *et al.*, 2020). Unlike traditional classroom settings, online learning lacks direct interaction

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2398-6247.htm**

Information Discovery and Delivery

[DOI 10.1108/IDD-08-2024-0120]

Conflict of interest: The authors declare that they have no conflict of interest.

Availability of data and materials: The data used this study are available from https://analyse.kmi.open.ac.uk/open-dataset

Received 17 August 2024
Revised 17 November 2024
28 January 2025
Accepted 30 January 2025

between instructors and students, making it difficult to monitor progress. However, technological advancements have mitigated this limitation by enabling the collection and storage of vast amounts of data on student activities within digital platforms (Kuzilek *et al.*, 2017). Clickstream data, which captures users' sequential interactions and navigation behaviors, offers a rich resource for analyzing learner engagement. This data-driven approach allows researchers to identify the underlying causes of student failure, enabling early interventions and enhancing the overall learning experience (Baig *et al.*, 2023). Predictive analytics not only helps identify areas of weakness but also supports the development of personalized learning plans tailored to improve academic performance (Batool *et al.*, 2023). By adopting advanced predictive analytics techniques, educational institutions can achieve significant improvements in student outcomes and the overall quality of instruction (Pallathadka *et al.*, 2023). In summary, while online learning presents unique challenges, the integration of data-driven strategies holds promise for addressing these issues and fostering a more effective and inclusive educational environment.

Numerous outcomes are predicted using a range of data analytics and machine learning approaches, with deep artificial neural networks (ANN) being particularly notable for their sophisticated learning capabilities (Coelho and Silveira, 2017). Traditional feature engineering techniques are revolutionized by deep learning, distinguished by its hierarchical learning approach comprising numerous computing layers and the ability for systems to learn from previous examples (Sarwat *et al.*, 2022). Nevertheless, there still needs to be more research in academic publications about the effectiveness of deep learning in learning analytics, especially when anticipatorily resolving student performance problems. Several ANN models, including long short-term memory (LSTM) and recurrent neural networks (RNN), are used to improve student outcomes proactively (Liu *et al.*, 2022). These models analyze learners' performance across daily, weekly, or monthly intervals, using the course schedule as a data series. To detect and assist students who may be in danger of underperforming and to ensure prompt interventions for enhanced academic achievement, the research community is putting more and more emphasis on these sequential analytic methodologies.

This paper presents an approach to predicting student performance that combines machine learning techniques with clickstream data. The two main stages of our method are data application and data extraction. Before analyzing, we first extract and organize clickstream data from OULAD (Open University Learning Analytics Data set) (Kuzilek *et al.*, 2017). We next divide the data into two subsets: weekly and monthly interactions for each activity that students complete in the virtual learning environment. Next, we use a blend of conventional and advanced machine learning models. This pertains to two advanced time-series models (LSTM-XGBoost and GRU), where we apply the Variance Inflation Factor (VIF) to evaluate and select relevant features, and five traditional methods (K-Nearest Neighbors, Random Forest, XGBoost and Naive Bayes). The key contributions of this paper are outlined as follows:

Q1. How can clickstream data and machine learning algorithms effectively predict student performance in virtual learning environments?

Q2. What are the comparative advantages of integrating traditional and advanced time-series machine learning models in analyzing and predicting students' engagement and success rates in a virtual learning environment?

The remainder of this paper is structured as follows: Section 2 provides a comprehensive literature review, exploring the use of machine learning models to predict student performance, with a focus on state-of-the-art approaches that leverage clickstream data to identify at-risk students in virtual learning environments and enhance educational outcomes. Section 3 outlines the methodology employed in our study, providing a detailed explanation of our approach to data collection, preparation and analysis. Section 4 presents the settings and findings from our experimental results, comparing the performance of different machine learning algorithms used in our analysis. Section 5 discusses two significant contributions and explores their implications for teaching and learning. Additionally, it highlights the limitations of the research and proposes directions for future studies. Finally, Section 6 concludes the paper by summarizing our contributions.

2. Literature review

The increasing reliance on online learning environments has brought about new challenges in understanding and improving student performance. Virtual learning environments differ significantly from traditional classrooms due to the absence of direct interaction between instructors and students. This makes it crucial to identify key factors affecting student performance and to leverage available data, such as clickstream data, for predictive and analytical purposes. Among these factors, student engagement emerges as a critical determinant, often reflected through actions such as participation in forums, timely completion of assignments and consistent interaction with course materials. Understanding these behaviors, coupled with advanced machine learning techniques, can help educators develop targeted interventions and improve student outcomes.

This literature review explores the use of machine learning models for predicting student performance. It begins by outlining the classification models employed in this domain and then delves into state-of-the-art approaches leveraging clickstream data to enhance educational outcomes.

2.1 Machine learning models

Machine learning offers a variety of models for classification tasks, broadly categorized into traditional models and advanced time-series models. These models play a pivotal role in processing the data collected from online learning platforms, including clickstream logs, to predict student engagement and performance effectively.

1. Traditional Models: Traditional models include well-established techniques such as Logistic Regression (LR), Naive Bayes, K-Nearest Neighbors (KNN), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). These models are often used for their simplicity, interpretability and efficiency in handling structured data. For example, LR and Naive Bayes are frequently employed for binary classification tasks, while RF and XGBoost excel in handling high-dimensional data sets and providing feature importance insights. Despite their

simplicity, traditional models often serve as a baseline for evaluating the effectiveness of more complex algorithms:

- *LR (Logistic Regression)*: By using a logistic function to estimate probabilities, logistic regression is a statistical technique for binary classification. By using a logistic (sigmoid) function to estimate probabilities, it simulates the connection between a dependent binary variable and one or more independent variables. Values between 0 and 1 are the outputs of this function, and they represent the likelihood that the dependent variable falls into a specific category. Many domains, including the social sciences, medicine, and machine learning, employ logistic regression extensively for problems involving the prediction of binary outcomes (Acito, 2023; Zou *et al.*, 2019).
- *Naive Bayes*: A popular probabilistic machine learning technique used in many domains, such as image recognition and natural language processing, is the Naive Bayes classifier. It is predicated on the Bayes theorem, which determines the likelihood of a hypothesis in light of observed data. The "naive" part comes from the presumption that features are conditionally independent, which makes calculations easier but may not accurately represent interdependence in the actual world (Blanquero *et al.*, 2021).
  In order to accommodate many data types, recent research (Smith *et al.*, 2023) has expanded the classical Naive Bayes model with innovations like Gaussian Naive Bayes, Multinomial Naive Bayes and Complement Naive Bayes. Because of its ease of use, effectiveness and competitive performance, naive Bayes classifiers continue to be widely used in a wide range of research fields where probabilistic categorization is crucial. Further improvements enhance their skills and adaptability to modern applications.
- *K-Nearest Neighbors (KNN)*: The k-Nearest Neighbors (KNN) classifier is an example of instance-based learning, also known as lazy learning, in which all computation is postponed until the function is evaluated and the function is only locally approximated. It is a nonparametric technique for problems involving classification. To discover the k closest training instances in the feature space to a given test point, the KNN classifier searches through the feature space. The test point is then assigned the label that is most prevalent among the k closest neighbors by a majority vote in the classification process. The success of the classifier is highly dependent on the selection of k and the distance metric, which is usually known as Euclidean distance. In many real-world settings, KNN may achieve excellent accuracy despite its simplicity, mainly when the decision border is irregular. However, because of the curse of dimensionality in high-dimensional environments, the algorithm's performance is reliant on the quality of the data (Fan *et al.*, 2023; Taunk *et al.*, 2019).
- *Random Forest*: By establishing a" forest" of trees to increase prediction accuracy and manage overfitting, the Random Forest approach for classification expands on the ease of use of decision trees via ensemble learning. To do classification tasks, it builds many decision trees during training and outputs the class that represents the majority vote of the classes predicted by each tree. This approach introduces randomization in two ways to improve generalization: randomly selected subsets of the data are used to train each tree, and randomly selected subsets of features are taken into account at each split in the tree. By lowering the correlation between trees, this method improves the accuracy and resilience of the model as a whole. The Random Forest algorithm is well recognized for its proficiency in managing huge data sets with high dimensionality, preserving accuracy even in situations when a significant amount of the data is missing and handling missing values. It is a well-liked solution for a variety of categorization issues due to its adaptability and simplicity of use (Haddouchi and Berrado, 2019; Paul *et al.*, 2018).
- *XGboost*: Extreme Gradient Boosting, or XGBoost, is a potent machine learning method that is often used for classification problems because of its remarkable efficiency and predicted accuracy. It is an extension of gradient-boosting techniques and a member of the ensemble learning family. XGBoost constructs an ensemble of decision trees sequentially, with each tree fixing the mistakes of the one before it. For classification, XGBoost converts each decision tree's output into a class probability using a logistic regression model. These probabilities are combined to get the final forecast, usually with the use of a weighted total. Thanks to control settings and regularization strategies, XGBoost stands out for its resilience against overfitting. XGBoost has several benefits, such as feature significance ranking, support for binary and multiclass classification and management of missing data. Additionally, users may fine-tune hyperparameters with customization, which makes it a flexible option for a range of classification jobs (Asselman *et al.*, 2023; Aydin and Ozturk, 2021).

2. Advanced Time-Series Models: Advanced time-series models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are particularly adept at capturing sequential patterns in data. These models are well-suited for analyzing time-dependent variables in student behaviors, such as weekly or monthly activity patterns. Hybrid approaches, like LSTM-XGBoost, combine the strengths of sequential data modeling and ensemble learning, enabling the extraction of temporal and structural features for robust predictions. These advanced models are particularly effective in addressing complex educational challenges, such as identifying at-risk students and forecasting performance trends:

- *LSTM-XGboost*: Combining the best features of XGBoost and Long Short-Term Memory (LSTM) neural networks, LSTM-XGBoost is a potent hybrid machine learning model that offers a strong foundation for classification problems. A kind of recurrent neural network (RNN) called an LSTM is intended to identify long-range relationships in sequential data. It is well-suited for jobs where temporal patterns are essential, such as voice recognition and natural language processing (Karim *et al.*, 2017), because of its proficiency with time series and sequential data. One well-known ensemble learning technique for handling structured data well is called XGBoost. By building an ensemble of decision trees, it

can manage complicated feature interactions and provide good prediction performance (Asselman *et al.*, 2023).

To capture complex temporal connections in data, LSTM-XGBoost makes use of LSTM's sequence modeling capabilities. It takes pertinent characteristics out of sequential input and combines them with XGBoost's potent ensemble learning. The combination of structured and sequential data processing allows the model to perform well in a variety of classification tasks (Zang *et al.*, 2023).

In order to determine the final classification, the LSTM-XGBoost model feeds the LSTM output into an XGBoost classifier after first using LSTM to identify sequential patterns in the input data. The capacity to handle both temporal and structural information is an advantage of this hybrid technique.

- *GRU*: One kind of recurrent neural network (RNN) architecture that has become popular in machine learning and natural language processing is the Gated Recurrent Unit (GRU). This sort of RNN architecture works well with sequential data, especially when it comes to classification tasks. An adaptation of conventional RNNs, the GRU incorporates gating techniques to mitigate the vanishing gradient issue and allow for the capturing of long-range dependencies in sequences (Chung *et al.*, 2014; Dey and Salem, 2017).

A GRU cell is essentially made up of two gates: the Update Gate (z_t) and the Reset Gate (r_t). At each time step, these gates control the flow of information across the network and how it is modified. Whether new information from the current input and the previous hidden state should be added to the current hidden state is determined by the update gate, while the reset gate determines whether information from the previous hidden state should be kept or ignored. The following mathematical equations determine the GRU's behavior:

$$\begin{aligned}
&\text{Reset Gate } (r_t): \quad r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \\
&\text{Update Gate } (z_t): \quad z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \\
&\text{Current Memory Content } (\widetilde{h}_t): \quad \widetilde{h}_t = \tanh(W \cdot [r_t \cdot h_{t-1}, x_t]) \\
&\text{Hidden State } (h_t): \quad h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \widetilde{h}_t
\end{aligned}$$

Given sequential input data, a GRU-based neural network is trained to learn representations and provide class predictions in the context of classification tasks. To produce classification predictions, the input sequence is usually represented by its final hidden state (h_t). This feature representation is then passed through one or more fully connected layers. To reduce the classification loss, optimization approaches based on gradient descent and backpropagation are used to train the network. Regarding natural language processing tasks, such as text sentiment analysis, named entity identification and audio recognition, the GRU has shown encouraging performance in a number of classification tests. A potent recurrent neural network architecture, the Gated Recurrent Unit (GRU) algorithm is well-known for its efficacy in sequential data processing and classification tasks because of its gating mechanisms and capacity to capture dependencies in input sequences. This makes it a valuable tool in the toolbox of deep learning models for classification (Chung *et al.*, 2014; Dey and Salem, 2017).

2.2 State of the art: using clickstream data for predicting and managing student performance

Clickstream data, which records the sequential interactions and navigational behavior of students within online platforms, has emerged as a powerful resource for understanding and managing student performance. Several studies have explored the use of this data in conjunction with machine learning techniques to develop predictive models and actionable insights.

Wang *et al.* (2022) proposed a Convolutional Residual Recurrent Neural Network (CRRNN) that combines short-term activity features with long-term trends in online student behavior. This approach achieved superior accuracy compared to traditional models, highlighting the importance of sequential data analysis in identifying at-risk students. Similarly, (Liu *et al.*, 2022) employed LSTM models, achieving an accuracy of 89.25%, emphasizing how specific interactions, such as reviewing course content and participating in quizzes, influence student outcomes.

In another study, Mubarak *et al.* (2021a) demonstrated the use of LSTM networks to analyze video clickstream data in Massive Open Online Courses (MOOCs), achieving prediction accuracies between 82% and 93%. This outperformed traditional baseline models such as Logistic Regression and Support Vector Machines (SVM). Similarly, (Waheed *et al.*, 2020) explored the application of deep artificial neural networks on clickstream data from virtual learning environments, achieving classification accuracies ranging from 84% to 93%. These findings highlight the effectiveness of deep learning in early identification of underperforming students.

Graph-based techniques have also been applied to analyze student interaction patterns. For example, (Khoushehgir and Sulaimany, 2023) introduced a novel method for dropout prediction in MOOCs by utilizing network topology. This graph-based approach converted enrollment data into network structures, demonstrating significant improvements in dropout prediction compared to traditional methods.

Hao *et al.* (2022) developed a predictive framework using genetic algorithms and stochastic Bayesian networks to analyze MOOC data. This model explored the interplay of environmental variables, course content and student characteristics, achieving high accuracy in predicting academic success and providing personalized support strategies.

Gaftandzhieva *et al.* (2022) examined the relationship between student participation in online courses and academic outcomes using Moodle Learning Management System data. Their study employed Random Forest (RF), XGBoost, KNN and SVM for performance prediction, with RF achieving the highest accuracy at 78%. These results emphasize the importance of data-driven insights in supporting at-risk students.

Zerkouk *et al.* (2024) utilized logistic regression and XGBoost to predict dropout scenarios on a Canadian distance learning platform. Their analysis incorporated a data set with 49 parameters, including sociodemographic and behavioral features, achieving an accuracy of 82%. This approach underscores the potential of predictive analytics in identifying students at risk of attrition.

In addition to these, (Waheed *et al.*, 2023) employed LSTM networks to detect students likely to fail a course, achieving an

accuracy of 84.57% on data from 22,437 students. This study demonstrated the effectiveness of LSTM for early interventions, highlighting the critical role of interpretability in deep learning for education.

For instance, (Adnan *et al.*, 2021) found that Random Forest (RF) fared better than SVM and KNN when examining the usefulness of demographics, clickstream data, evaluation scores and their combinations in creating predictive models. Beneficial is clickstream data, which records in-depth exchanges between students and the learning environment. It provides perceptions of how engaged and behaved pupils are, which may provide information about their study habits, emotional health and degree of knowledge. For instance, very involved students may often participate in discussion forums by posing queries or answering those of their classmates.

With an emphasis on XGBoost in particular, (Asselman *et al.*, 2023) explored how to improve Performance Factors Analysis (PFA) for the prediction of student performance via the application of ensemble learning methods. The results show that, on a number of data sets, XGBoost significantly outperforms other models and traditional PFA in terms of prediction accuracy. This work introduces a more efficient method for predicting student outcomes, which significantly advances the field of educational data mining. These developments should make it easier to provide more individualized learning opportunities.

(Al-Azazi and Ghurab, 2023) investigated patterns in student performance in MOOCs using a virtual learning environment, delving into the field of learning analytics. The research presented ANN-LSTM, a unique multiclass day-wise prediction model that combines artificial neural networks with long short-term memory (LSTM). By the third month of the course, our model outperformed the conventional baseline models in terms of accuracy, achieving a rate of about 70%. Compared to the accuracy rates of 53% for Recurrent Neural Networks (RNN) and 57% for Gated Recurrent Units (GRU), this was much higher. Moreover, the ANN-LSTM model demonstrated accuracy increases ranging from 6% to 14%, outperforming current state of the art models. The results highlight how well the LSTM model predicts student success in MOOCs at an early stage.

Table 1 summarizes key studies that utilized clickstream data for predicting and managing student performance, showcasing the diverse methodologies and results achieved.

These findings highlight the critical role of clickstream data and machine learning in transforming virtual learning environments. By focusing on student engagement and leveraging advanced models, researchers and educators can develop effective strategies to improve learning outcomes and reduce dropout rates.

3. Methodology

An outline of the methodology of our suggested approach is given in this section. The comprehensive data set description subsection 3.1, which highlights the data set's selection and salient characteristics, opens the discussion. We next explain the preparation procedures, which include normalization and data cleaning subsection 3.2. The next step is feature extraction subsection 3.3, where the goal is to find and extract important features to build a refined sub-data set that can be used for model training and testing. A variety of classifiers, including LR, NB, KNN, Random Forest and XGBoost, as well as cutting-edge neural networks like LSTM-XGBoost and GRU subsection 2.1, are included in our model portfolio. To thoroughly evaluate model performance, we lastly examine the assessment criteria subsection 4.1 using measures like accuracy, specificity, sensitivity and area under curve (AUC) (subsection 4.2, subsection 4.3). The methodology is expounded upon in Figure 1.

3.1 Data description

The Open University Learning Analytics Data set (OULAD) (Kuzilek *et al.*, 2017) comprises data from 22 courses involving a total of 32,593 students. This data set encompasses a wide range of information, including students' assessment results and detailed logs of their interactions with the Virtual Learning Environment (VLE), which are represented through daily summaries of student clicks amounting to 1,06,55,280 entries. The data set is categorized into three primary domains:

1 Demographic data;
2 Performance data (reflecting students' results and achievements); and
3 Learning behavior data (comprising logs of student activities within the VLE).

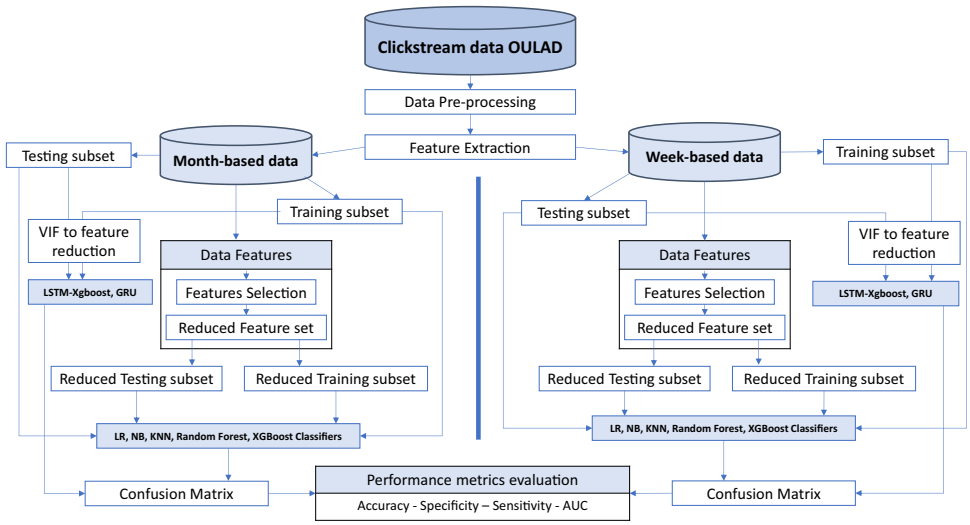
To adhere to ethical and privacy standards, the data set has undergone a meticulous anonymization process, ensuring that

Table 1 Key studies using clickstream data to predict student performance

Study	Data set used	Method used	Accuracy (%)
Aljohani *et al.* (2019)	Open university learning analytics	LSTM	90
Yang *et al.* (2017)	MOOC video-watching clickstreams	Time-series neural networks	84
Waheed *et al.* (2020)	Virtual learning environments (VLE) clickstreams	Deep artificial neural networks	84
Wu *et al.* (2019)	Weblog data	Support vector machines (SVM)	81.22
Casey and Azcona (2017)	Student activity patterns	Random Forest classification	85
Aouifi *et al.* (2020)	Video viewing behavior data	Text graph convolutional networks	67.23
Chu *et al.* (2021)	In-video activity data	Clustering guided Meta-Learning	87
Zhou *et al.* (2015)	Online behavior and website access records	Specificity and sensitivity analysis	65
Mubarak *et al.* (2021b)	Video-clickstream data from MOOCs	LSTM (long Short-Term memory)	89
Park *et al.* (2017)	Clickstream data with statistical change methods	Bayesian change detection	78
Kórösi *et al.* (2018)	Short MOOC clickstream data	Logistic regression	82

Source(s): Created by authors

Figure 1 Proposed approach methodology



Source(s): Created by authors

individual students cannot be identified. Figure 2 illustrates the structured nature of the data set, which consists of several interconnected tables. These tables include "studentInfo" (containing student demographics and module results), "courses" (providing a list of all modules and their respective presentations), "studentRegistration" (housing registration information for module presentation), "assessments" (providing details about assessments in module presentation), "studentAssessment" (containing results of students' assessments), "studentVle" (offering information regarding student interactions with the VLE) and "vle" (providing data about materials available within the VLE).

3.2 Data preprocessing

The crucial first stage in getting raw data ready for analysis or modeling is data preprocessing, which involves organizing, converting and cleaning the data to improve its applicability and quality. To guarantee accurate and significant outcomes in machine learning processes, recent research by Géron (2022) highlights its significance. We concentrated on three particular tables in this study since they support the goals of the investigation. First, the table "studentInfo" provides demographic data and is the source of the categorization label, which is the course's outcome. The second table (called "vle") is referred to as the module presentation and contains details on various kinds of activities. Important information on students' clickstream interactions inside the VLE is presented in the third table, "studentVLE." Table 2 presents the specifics of the three tables.

After a thorough evaluation of the OULAD, it was determined that concentrating on a single course would provide significant insights into the dynamics of instruction and learning in that particular setting. We analyzed the number of students who did not withdraw from seven different courses within the raw data set to determine which course had the most extensive data set. After this analysis, it was clear that the course designated as "BBB" had the most significant number of students who had not withdrawn, which made it the best option for our study. It should be noted that this specific course fell under the social sciences category (Kuzilek *et al.*, 2017).

Before merging the "studentInfo," "vle" and "studentVle" data sets, the "final_result" feature within the studentInfo data set included four distinct classes: "Pass," "Fail," "Withdrawn" and "Distinction." Among the total sample of 7,692 students, 39.20% (3,015 students) were classified as "Pass," 8.68% (668 students) received a "Distinction," 30.14% (2,318 students) were marked as" Withdrawn" and the remaining 21.98% (1,691 students) were categorized as "Fail" (Figure 3). The samples marked as "Withdrawn" were subsequently ignored.

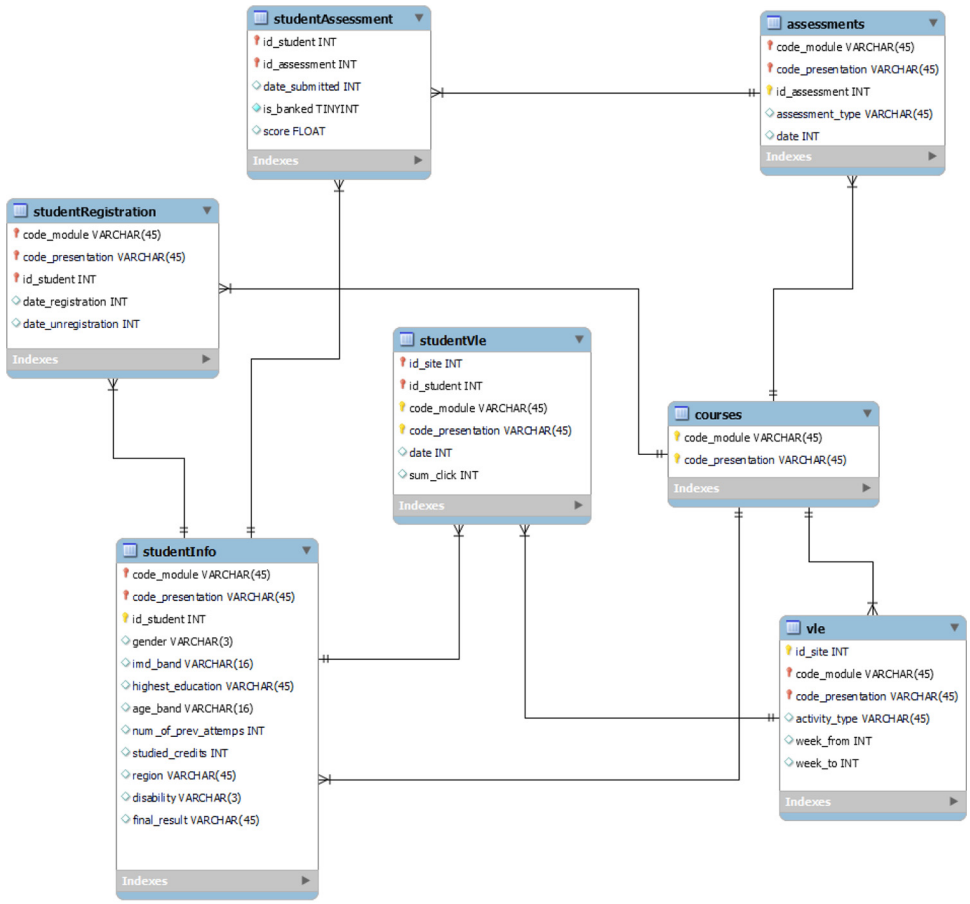
After merging, in the pursuit of simplicity, the "final_result" feature within the merged data set, which initially comprised three distinct classes: "Pass," "Fail" and "Distinction," was refined to a binary classification system consisting of "Pass" and "Fail" more precisely, the "Distinction" label was consolidated with the "Pass" label. Consequently, among the total sample of 5,372 students, 69.92% (3,756 students) were designated as "Pass" while the remaining 30.08% (1,616 students) were categorized as "Fail" This modification simplifies the data set, as depicted in (Figure 4), making it more straightforward for analytical purposes by categorizing students simply as either passing or failing.

A reform that translated "Fail" to the numeric value 0 and "Pass" to the numeric value 1 was put in place to standardize data for classification models. The representation of student performance was standardized by this change, making it compatible with a range of classification techniques. Therefore, by classifying students who fail as 0 and those who pass as 1, the binary encoding facilitates accurate predictions and insights and streamlines analysis and model training.

In the course BBB, there were 12 activity categories (Kuzilek *et al.*, 2017), which comprised:

1 *homepage:* the homepage is typically the initial landing page for the course, providing essential information, announcements and navigation links for students.

Figure 2 Data set schema



Source(s): Created by authors

2 *forumng:* this refers to discussion forums or message boards where students can interact, discuss course-related topics, ask questions and share ideas.
3 *subpage:* are supplementary content pages that offer additional information, resources or materials related to specific course topics or modules.
4 *oucontent:* represents Open University content modules, which likely contain course materials such as readings, presentations or multimedia resources.
5 *resource:* provide various learning materials, such as documents, PDFs, videos or external links, to aid students in their studies.
6 *quiz:* are used for assessment and evaluation, allowing students to take quizzes or tests related to course content.
7 *url:* contain web links to external resources, websites or references that support course learning objectives.
8 *glossary:* contain definitions and explanations of key terms and concepts used in the course, assisting students with understanding subject-specific terminology.
9 *sharedsubpage:* collaborative or shared content pages that students work on together, contributing to a group project or assignment.
10 *ouelluminate:* refer to modules related to web conferencing or virtual classroom sessions, enabling real-time interaction and lectures.
11 *oucollaborate:* involve collaborative tools or activities that encourage students to work together on course-related projects or tasks.
12 *questionnaire:* are used to collect feedback, surveys or responses from students to assess their understanding, satisfaction or opinions about the course.

3.3 Feature extraction

To facilitate practical analysis or modeling, feature extraction is the process of identifying and reducing pertinent information from raw data into a smaller and more meaningful collection of features. According to recent research by Tang *et al.* (2024), feature extraction methods simplify and emphasize essential patterns in data, which improves the performance of machine learning algorithms. We conducted a thorough feature extraction procedure in our study using the OULAD, with the primary goal being to record and visualize click counts related to various activity categories. Two unique feature sets were produced as a result of this work, which was made possible by the following extraction techniques:

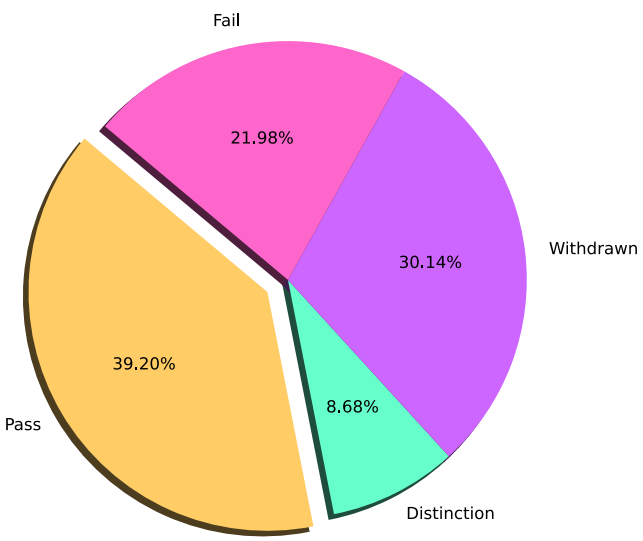
- *Combining Click Counts Depending on Time:* We used a multigranular method in the first stage of feature extraction; aggregating click counts at the weekly and monthly temporal levels. This aggregate's main goal was

Table 2 The three tables' specifics

Feature	Description	Type
studentInfo		
code_module	Module identification code where the student is registered	Categorical
code_presentation	Presentation identification code during which the student is registered on the module	Categorical
id_student	The unique student identification number	Numeric
gender	Student's gender	Categorical
region	The geographic region where the student lived while taking the module-presentation	Categorical
highest_education	The highest student education level on entry to the module presentation	Categorical
imd_band	The IMD band of the place where the student lived during the module-presentation	Categorical
age_band	A band of student's age	Categorical
num_of_prev_attempts	The number of times the student has attempted this module	Numeric
studied_credits	The total number of credits for the modules the student is currently studying	Numeric
disability	Indicates whether the student has declared a disability	Categorical
final_result	Student's final result in the module-presentation	Categorical
studentVle		
code_module	The module identification code	Categorical
code_presentation	The presentation identification code	Categorical
id_student	The unique student identification number	Numeric
id_site	The VLE material identification number	Numeric
date	The day of student's interaction with the material	Numeric
sum_click	The number of times the student interacted with the material	Numeric
vle		
id_site	The identification number of the material	Numeric
code_module	The identification code for the module	Categorical
code_presentation	The identification code of the presentation	Categorical
activity_type	The role associated with the module material	Categorical
week_from	The week from which the material is planned to be used	Numeric
week_to	The week until which the material is planned to be used	Numeric

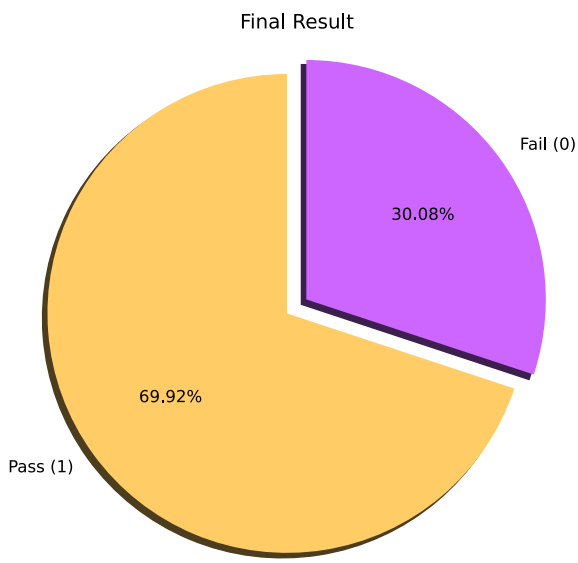
Source(s): Created by authors

Figure 3 Final_result feature distribution before merging



Source(s): Created by authors

Figure 4 Final_result feature distribution after merging



Source(s): Created by authors

to summarize the dynamics of click counts by identifying weekly and monthly trends throughout the length of the course (T). We divided the course duration into two separate parts to make this easier:

- T0 Segment (week0, month0): This section reflected the time frame before the course's formal start. To create an aggregate measure during this period, we combined click counts using a summation approach.
- Week 1 to week 38 or month 1 to month 8 in the T1 to Tn Segment: This section covered the time frame from the official start of the course until its conclusion. As in the T0 phase, we kept adding up the number of clicks over this period.

A reformulation was done to standardize the data for the classification models. Labels "week0" through "week38" were translated to numeric values between 0 and 38, and labels "month0" through "month8" were translated to numeric values between 0 and 8.

- *Changes to the Data Structure:* We started a transformation procedure to reorganize the data into a new format in the feature extraction step that followed. This included keeping an eye on every kid at certain times based on how many cliques they were connected to.

After undergoing this transformation, two unique feature sets were produced, which were designated "week" and "month" respectively. These feature sets followed the structured pattern shown in Figure 5. These sets used rows to correlate to time-based observations of students' click counts and columns to reflect the different activity categories.

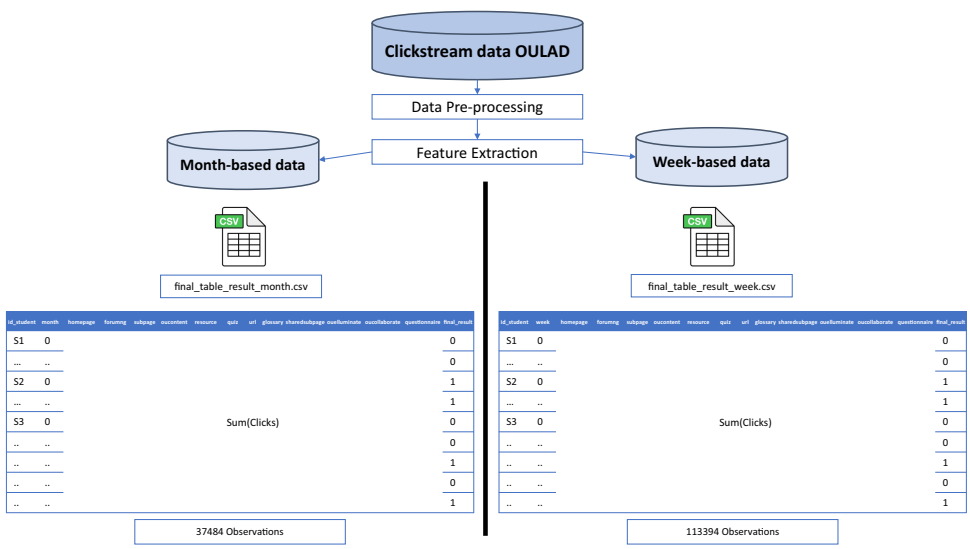
Notably, the data set's size was significantly increased by this change. In particular, a comprehensive data set consisting of 1,13,394 observations in the week-based data set and 37,484 observations in the month-based data set was generated from the original cohort of 5,372 sample students.

This method was justified by the need to avoid producing high-dimensional feature sets while still preserving the time and activity dimensions.

A detailed summary of all clicks logged for each activity per month is shown in Figure 6, which is part of the study of student involvement with the virtual learning platform. The information encompasses a variety of interactive components found on the platform, such as forum participation (forumng), homepage navigation, content consumption via pages (subpage, oucontent), resource access (resource), quiz attempts (quiz), external link navigation (url) and other features like the use of glossaries, shared subpages, live online sessions (ouelluminate, oucollaborate) and questionnaire engagement. Important findings from the data point to substantial participation in forum discussions (forumng) and homepage visits, indicating these as students' main means of contact. The fact that participants consistently interact with these activities indicates how important they are to the online learning process. The data also shows a significant level of resource access and content consumption (oucontent), highlighting the value of instructional resources in digital learning contexts. The data set's temporal dimension shows variations in activity levels, with a general pattern of declining engagement toward the end of the month as seen by a decrease in clicks for all activities. This pattern is explained by a number of things, including the conclusion of the course, changes in the academic calendar and dwindling platform novelty.

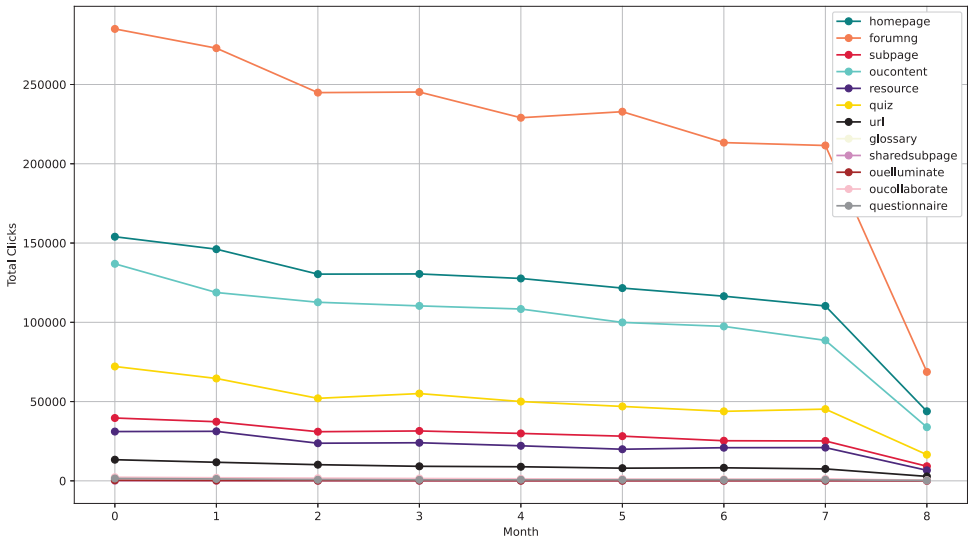
The total weekly clicks for every activity are shown in Figure 7, which compiles the data into a single line chart to facilitate cross-activity comparison. The data set spans weeks 0 through 38, with week 0 having the most excellent click-through rate at almost 4,39,000. After this high, there is a general decreasing trend in clicks. However, there are notable variations throughout the time. Interestingly, there is a noticeable spike in engagement from week 13 to week 22, with click counts typically ranging from 1,15,000 to 1,72,000. There is a noticeable downward trend in clicks as week

Figure 5 Procedure for extracting features



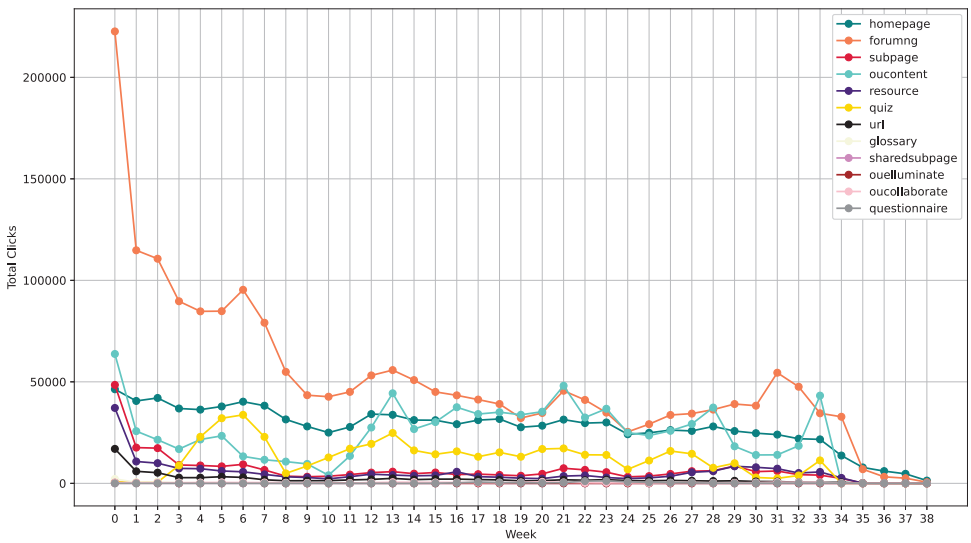
Source(s): Created by authors

Figure 6 Total clicks for each activity by month



Source(s): Created by authors

Figure 7 Total clicks for each activity by week



Source(s): Created by authors

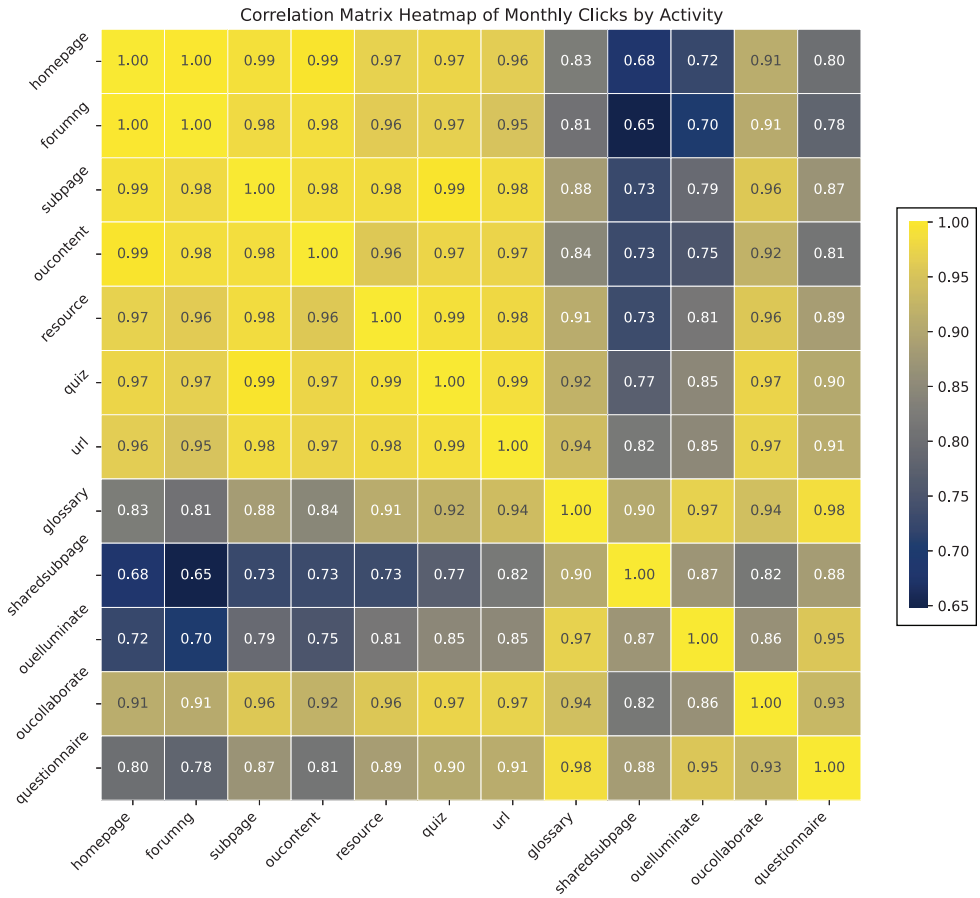
38 approaches; by week 38, there are around 1900 clicks, which is the lowest level of engagement to date. As a result, Figure 7 summarizes the weekly activity patterns in the data set rather well.

A quantitative evaluation of the linear links between different sorts of activities based on their involvement (clicks) throughout the months is provided by the correlation matrix shown in Figure 8. The range of correlation values is −1 to 1, meaning:

- A value close to 1 implies a strong positive correlation, indicating that as clicks for one activity increase, clicks for the other activity also tend to increase.
- A value close to −1 indicates a strong negative correlation, suggesting that as clicks for one activity increase, clicks for the other activity tend to decrease.
- A value around 0 suggests little to no linear correlation between the clicks for two activities.

From the matrix displayed in Figure 8, we observe high positive correlations between most activities, with values notably close to 1 for pairs such as homepage and forumng, subpage and quiz, indicating that engagements in these activities tend to increase or decrease together over the months. This could suggest that when users are active, they engage with multiple aspects of the online learning environment, not just focusing on a single type of activity. The glossary, sharedsubpage and ouelluminate activities show lower correlation values with others, though still positive, indicating a less strong but still positive relationship

Figure 8 Correlation heatmap matrix for monthly activity clicks



Source(s): Created by authors

in their usage patterns compared to the more central activities like homepage, forumng and quiz.

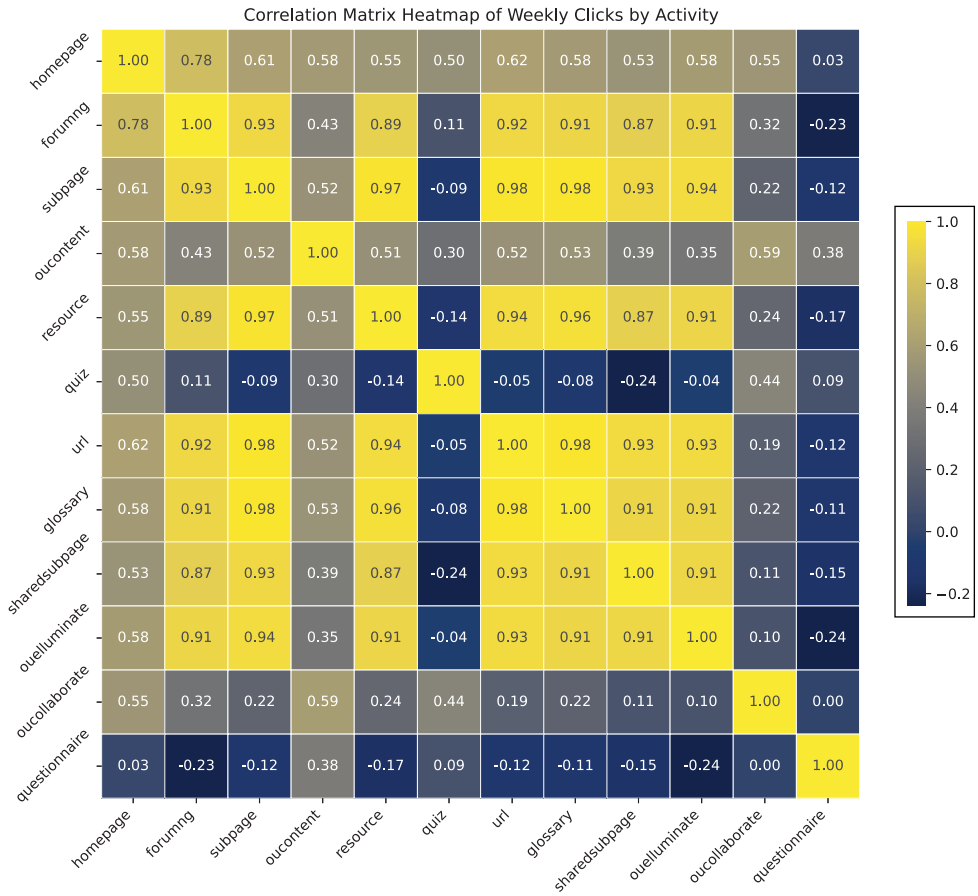
The correlation matrix in Figure 9, which analyzes weekly clicks across various activities, illustrates the relationships among these activities on a weekly basis. There is a strong positive correlation of 0.783 between homepage and forumng, indicating a close link between activities on the homepage and forum engagements. Furthermore, forumng and subpage exhibit a very strong positive correlation of 0.927, showing that forum interactions and subpage clicks are highly intertwined. Another very strong positive correlation of 0.973 between subpage and resource suggests a close relationship between subpage navigations and resource utilizations. URL and glossary display a strong positive correlation of 0.984, revealing a tight connection between URL accesses and glossary consultations. Conversely, lower correlations, such as between quiz and other activities like forumng (0.107) or resource (−0.136), indicate a weaker direct relationship in weekly student engagement patterns in these areas.

4. Experimental results

Using the Open University Learning Analytics Data set (OULAD) Clickstream data, the seven machine learning algorithms described in subsection 2.1 are being examined for the experimental findings to determine which method or methods most effectively improve students' performance. It is essential to emphasize that the text shows that no single machine learning approach consistently produces the best accurate predictions. Thus, to compare their efficacy, several models are used. For advanced models like XGBoost-LSTM and GRU, the Variance Inflation Factor (VIF) is applied to evaluate and select relevant features by addressing multicollinearity, thereby ensuring optimal performance. A high correlation between multiple factors indicates the presence of multicollinearity, which can be problematic as it complicates isolating the individual contributions of independent variables to the dependent variable. To address this issue, VIF is calculated for each feature to detect multicollinearity. If the VIF is high, variables with the highest VIF values are removed to reduce multicollinearity (Katrutsa and Strijov, 2017).

A 10-fold cross-validation approach was used in this work to improve training data utilization and guarantee the reliability of our findings (Nti *et al.*, 2021; Wong and Yeh, 2019). The data set is divided into ten equal halves using this procedure. The training set receives 80% of the data for each fold, while the test set receives the remaining 20%, following the standard training-to-testing ratio of 8:2. This kind of approach makes it easier to evaluate the model thoroughly using the training set and then determine how generalizable it is using the test set.

Figure 9 Correlation heatmap matrix for weekly activity clicks



Source(s): Created by authors

To preserve the integrity of the experimental design, the ratio of "Fail" to "Pass" outcomes in both the training and test data sets was carefully controlled. This was achieved by employing the "final_result" attribute as a stratifying variable, ensuring that the distribution of outcomes remained consistent across all folds of the cross-validation. Specifically, this stratification was applied to two distinct data sets: one based on monthly data and the other on weekly data, as detailed in Tables 3 and 4, respectively. These tables illustrate the counts of students for each outcome ("Fail" or "Pass") in the training and test data sets, reflecting the meticulous effort to maintain balance and representativeness in both temporal data sets.

For setting the parameters of our models for both the monthly and weekly data sets, we ensured that all traditional models utilized normalized input data, within the range of 0 to 1. Specifically, the

Table 3 Test and training for month-based data

	1 (Pass: 80.44%)	0 (Fail: 19.56%)
Data (n = 37,484)	30,154	7,330
Training data (n = 29,987)	24,123	5,864
Test data (n = 7,497)	6,031	1,466

Source(s): Created by authors

Table 4 Test and training for week-based data

	1 (Pass: 83.40%)	0 (Fail: 16.60%)
Data (n = 1,13,394)	94,570	18,824
Training data (n = 90,715)	75,656	15,059
Test data (n = 22,679)	18,914	3,765

Source(s): Created by authors

Logistic Regression model was trained using hyperparameters set to max_iter = 1000, solver = 'lbfgs' and random_state = 42. The Naive Bayes model was utilized with its default hyperparameters to maintain simplicity and standardization. For the K-Nearest Neighbors (KNN) model, we selected an optimal k value of 5. The Random Forest (RF) model was adjusted by tuning two key hyperparameters, with n_estimators set to 100. Finally, the XGBoost model was applied with its default parameters, and consistency was ensured by setting the random_state to 42.

For our two advanced time-series models, we utilized the original input data, comprising 5,372 sequences, each representing an individual student. The models were structured to accommodate 9 time steps per sequence for monthly data and 39 time steps per sequence for weekly data, reflecting the temporal dynamics of student interactions. Each

time step incorporated 12 features based on activity counts, providing a detailed snapshot of student engagement over time. Additionally, we aligned these sequences with 5,372 corresponding targets, which represented the final result for each student.

The LSTM-XGBoost model integrates both LSTM and XGBoost technologies, featuring an input layer, dual LSTM hidden layers for weekly data, and triple LSTM hidden layers for monthly data, culminating in a fully connected and an output layer. It employs the Adam stochastic gradient descent as its optimization function and categorical binary cross-entropy for the loss function. With a default learning rate and a dropout rate of 0.2, it also includes a dense layer with a sigmoid activation function. Specifically, for weekly data, represented as (39,12) input shape, the model comprises 64 hidden units in the first LSTM layer and 32 in the second, with a batch size of 128 and 4 epochs. In contrast, the monthly data model, with inputs shaped (9,12), features 64 hidden units across the first two LSTM layers and an additional 64 in the third, with a smaller batch size of 8 and 10 epochs. The XGBoost component, which processes data generated by the LSTM model, is configured with parameters including a maximum depth of 8, a learning rate of 0.1 and the objective set to "binary:logistic," across 10 training rounds.

The GRU model, designed for both weekly and monthly data analysis, is composed of an input layer, two GRU hidden layers, a fully connected layer and an output layer, with an Adam stochastic gradient descent optimization function, and categorical binary cross-entropy as the loss function. Different learning rates are applied, 0.01 for monthly and 0.0001 for weekly data, complemented by a Dense layer with sigmoid activation. For the weekly data set, characterized by a (39,12) input shape, the model features 128 hidden units in the first GRU layer and 64 in the second, with a batch size of 8 and 4 epochs. Similarly, for the monthly data, with a (9,12) input shape, it maintains the same hidden unit configuration but adjusts the batch size to 8 and increases the epochs to 8.

4.1 Evaluation criteria

The results of our models are presented and analyzed in the part that follows, with an emphasis on how to evaluate them using confusion matrices, as described by Tharwat (2020), Vujović *et al.* (2021). By monitoring both accurate and inaccurate predictions for every attribute value, the confusion matrix is essential in elucidating the link between actual and anticipated class characteristics and facilitating an evaluation of the model's classification accuracy. Using this matrix in our research is primarily intended to evaluate how well our models predict student achievement. Four outcomes are distinguished by the matrix, which is shown in Table 5 (Tharwat, 2020; Vujović *et al.*, 2021). True Positives (TP) represent learner responses that the model accurately predicts; True Negatives (TN) represent responses that are incorrectly classified as correct; False Positives (FP) represent incorrect responses that are mistakenly predicted as correct; and False Negatives (FN) represent correct responses that are incorrectly labeled as incorrect.

Table 5 Confusion matrix for a binary classifier

		Predicted
	Correct	**Incorrect**
Actual		
Correct	True positive (TP)	False positive (FP)
Incorrect	False negative (FN)	True negative (TN)

Source(s): Created by authors

Typically, to assess the performance outcomes forecasted by binary classification models, machine learning offers a suite of metrics for examining predictive precision. We have utilized the following among these metrics:

- *Accuracy*: This is one of the most frequently utilized metrics for evaluating performance. It calculates the proportion of students' answers that are correctly classified, essentially measuring the ratio of correctly classified answers to the total number of answers. The formula for accuracy is given by:

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

- *Precision*: This metric indicates the percentage of positive students' answers that are correctly classified out of the total number of expected positive responses. The precision formula is:

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

- *Recall*: Also known as sensitivity, recall quantifies the number of positive students' answers correctly classified relative to the total number of answers that should have been classified as positive. The recall formula is:

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

- *F1-Score*: This metric serves as the harmonic mean of precision and recall, offering a statistical measure to rate performance. High F1-Score values denote superior classification performance. The formula for the F1-Score is:

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

- *Area Under Curve (AUC)*: This is a commonly used metric for assessing predictions in binary classification tasks, where 0.5 is often used as the default probability threshold. However, it is advised to use different criteria in situations when this threshold becomes insufficient. The most often used technique for visually representing the effectiveness of a binary classifier at different thresholds is the Receiver Operating Characteristic (ROC) curve. Plotting the True Positive Rate versus the False Positive Rate is required in this situation, where the False Positive Rate is determined by taking the Specificity into account. As a complete measure of classifier performance, the AUC (ascertained from the ROC curve) estimates the likelihood that the classifier will score a randomly picked positive example higher than a randomly selected negative example (Tharwat, 2020).

4.2 Results

Using an Intel Core i5 CPU running at 1.8 GHz and 8 Go of RAM, we have developed our models to test our experiment. We used the assessment measures outlined in the paragraph subsection 4.1 to evaluate these models' performance once they were created.

First, we have started the analysis by comparing the misclassification rate of each algorithm. The two tables provided compare the performance of various classification models on month-based and week-based data sets. The first Table 6 presents the results for traditional machine learning models and a variant of XGBoost, while the second Table 7 showcases the performance of advanced neural network-based models, specifically LSTM-XGBoost and GRU (Gated Recurrent Unit). To analyze and compare the models, we consider the misclassification rate, which is a crucial evaluation metric for classification tasks. The misclassification rate is calculated as the number of incorrect predictions divided by the total number of predictions. Lower misclassification rates indicate better model performance.

For each model, the misclassification rate is calculated as:

$$\text{Misclassification Rate} = \frac{\text{Correct} + \text{Incorrect}}{\text{Incorrect}}$$

Figure 10 shows the misclassification rates for each model across the month-based and week-based data sets LR (Logistic Regression) 0% for both month-based and week-based data sets, indicating perfect classification, which demonstrates an indication of overfitting due to the absence of incorrect predictions. NB (Naive Bayes) 1.13% for the month-based data set and significantly higher at approximately 296.90% for the week-based data set, indicating a major issue in the calculation for the week-based data set. KNN (K-Nearest Neighbors): 5.95% for the month-based data set and 4.35% for the week-based data set, showing relatively good performance with a slight improvement in the week-based data set. Random Forest 2.87% for the month-based data set and 3.01% for the week-based data set, demonstrating consistent performance across both data sets. XGBoost 0.66% for the month-based data set and 0.19% for the

Table 6 Confusion matrix of traditional models on month-based and week-based data sets

	Month-based data set		Week-base data set	
	Correct	**Incorrect**	**Correct**	**Incorrect**
LR				
Correct	6,031	1466	18,914	3,765
Incorrect	0	0	0	0
NB				
Correct	5,966	1447	5,025	689
Incorrect	65	19	13,889	3,076
KNN				
Correct	5,694	1382	18,172	3,562
Incorrect	337	84	742	203
Random Forest				
Correct	5,866	1422	18,393	3,624
Incorrect	165	44	521	141
XGBoost				
Correct	5,991	1457	18,881	37,56
Incorrect	40	9	33	9

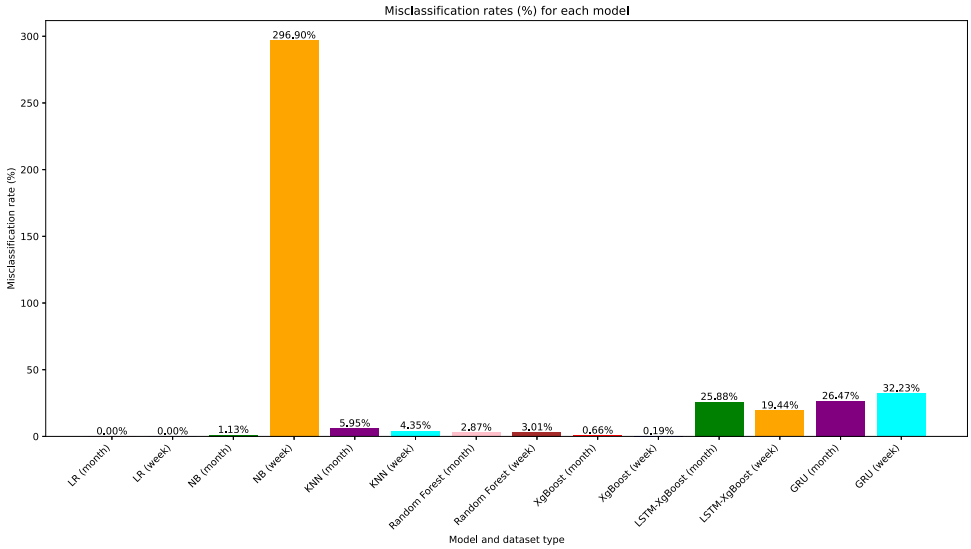
Source(s): Created by authors

Table 7 Confusion matrix of LSTM-XGBoost and GRU on month-based and week-based data sets

	Month-based data set		Week-based data set	
	Correct	**Incorrect**	**Correct**	**Incorrect**
LSTM-XGBoost				
Correct	746	108	760	140
Incorrect	23	198	9	166
GRU				
Correct	756	94	738	75
Incorrect	13	212	31	231

Source(s): Created by authors

Figure 10 Misclassification rates of all models on month-based and week-based data sets



Source(s): Created by authors

week-based data set, showing excellent performance and improvement in the week-based data set. LSTM-XGBoost 25.88% for the month-based data set and 19.44% for the week-based data set, showing a high misclassification rate but improvement in the week-based data set. GRU 26.47% for the month-based data set and 32.23% for the week-based data set, indicating relatively high misclassification rates, with performance deteriorating in the week-based data set.

The traditional models generally show lower misclassification rates compared to the advanced neural network-based models, with XGBoost and Random Forest exhibiting particularly strong performance across both data sets. The misclassification rate for Naive Bayes in the week-based data set seems anomalously high, suggesting a potential error in calculation; such a rate exceeds logical percentage bounds and likely indicates data handling. The advanced models (LSTM-XGBoost and GRU) have higher misclassification rates, which might be due to the complexity of these models and their sensitivity to the data set size.

According to Table 8, GRU stands out with the highest scores across almost all metrics, especially in accuracy *(90.04%)* and F1-score *(93.39%)*, indicating exceptional performance in balancing precision and recall, making it highly effective for the month-based data set. LSTM-XGBoost also shows strong performance, particularly in F1-score (91.92%), suggesting its effectiveness in handling complex patterns within the data set. Traditional models like LR, NB, KNN, Random Forest and XGBoost show lower performance compared to deep learning models. However, they still achieve reasonable scores, with LR and XGBoost performing notably well in terms of accuracy and F1-score. Interpretation of models on week-based data set GRU again demonstrates the highest performance, particularly in accuracy *(90.13%)* and F1-score *(93.29%)*, indicating its

Table 8 Performance metrics of models for month-based and week-based data sets

Data set	Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Month-based data set	LR	80.44	80.44	100	89.16
	NB	79.83	80.48	98.92	88.75
	KNN	77.07	80.46	94.41	86.88
	Random Forest	78.83	80.48	97.26	88.08
	XGBoost	80.03	80.43	99.33	88.89
	LSTM-XGBoost	87.81	87.35	97.00	91.92
	GRU	*90.04*	*88.94*	*98.30*	*93.39*
Week-based data set	LR	83.39	83.39	100	90.94
	NB	35.72	87.94	26.56	40.80
	KNN	81.02	83.61	96.07	89.41
	Random Forest	81.72	83.53	97.24	89.87
	XGBoost	83.29	83.40	99.82	90.88
	LSTM-XGBoost	86.13	84.44	98.82	91.07
	GRU	*90.13*	*90.77*	*95.96*	*93.29*

Source(s): Created by authors

superior ability to accurately classify positive cases without significantly increasing false positives. LSTM-XGBoost and XGBoost show commendable performance, particularly in F1-score, suggesting these models are well-suited for the week-based data set, balancing precision and recall effectively. NB displays a significant drop in performance, especially in recall (26.56%) and F1 score (40.80%), indicating it struggles with correctly identifying positive cases in the week-based data set. LR, KNN and Random Forest exhibit good performance, with LR surprisingly reaching a perfect recall score (100%), though this indicates a bias toward predicting the positive class. Overall comparison and conclusion deep learning models (GRU, LSTM-XGBoost) consistently outperform traditional machine learning models in both data sets, underscoring their capability to handle complex, temporal data more effectively. The week-based data set appears to be more challenging for certain models (e.g. NB, LR), this is due to the characteristics of the data and differences in how models handle temporal granularity. The disparity in performance between data sets for models like NB highlights the importance of model selection based on data set characteristics. High recall scores for models like LR in the week-based data set indicates a tendency toward over-predicting the positive class. Precision, recall and F1-score provide a more nuanced understanding of model performance than accuracy alone, especially in imbalanced data sets. Choosing the right model depends on the specific needs of the application, including the importance of false positives vs false negatives (precision vs recall), the balance between these metrics (F1-score) and the overall correctness of predictions (accuracy). Deep learning models offer substantial benefits for complex pattern recognition and temporal data but require more computational resources and data to train effectively.

The Table 9 lists the ROC curve area for seven models across two different data sets: month-based and week-based. For the month-based data set, the GRU model shows the highest ROC curve area *(0.84)*, indicating the best performance among the listed models. The LSTM-XGBoost also performs significantly well compared to other models like LR, NB, KNN, Random Forest and XGBoost. This suggests that GRU, a type of recurrent neural network, is highly effective for the month-based data set, potentially due to its ability to capture temporal dependencies. LSTM-XgBoost also demonstrates strong performance (AUC of 0.81), suggesting that the combination of LSTM (a type of recurrent neural network) with XGBoost (a gradient boosting framework) is effective for this data set. Traditional machine learning models like LR (Logistic Regression), NB (Naive Bayes), KNN (K-Nearest Neighbors), Random Forest and XGBoost show relatively lower AUC values (ranging from 0.51 to 0.55), indicating less effectiveness in classification for this data set. For the week-based data set, The GRU model outperforms others with an AUC of *(0.86)*, reinforcing its efficacy in handling sequential data and its strong ability to classify the week-based data set accurately. LSTM-XGBoost follows with an AUC of 0.76, suggesting that while effective, it is slightly less so than in the month-based data set or when compared directly to GRU. XGBoost shows a notable improvement in performance for the week-based data set (AUC of 0.61) compared to the month-based data set, indicating its relatively better suitability for this data set. Similar to the month-based data set, traditional models like LR, NB, KNN, and Random Forest present lower AUC values (ranging from 0.53 to 0.59) than the deep learning models, though they show a slight improvement over their performance on the month-based data set. Deep learning models, particularly GRU and LSTM-XGBoost, consistently outperform traditional machine learning models across both data sets, highlighting their strength in capturing complex patterns and dependencies within the data. The week-based data set appears to allow for slightly better model discrimination than the month-based data set, as seen in the generally higher AUC values across all models. This is due to the nature of the data, where shorter time frames provide more relevant or discriminative features for classification.

Table 9 ROC curve area of all models for month-based and week-based data sets

Data set	Models	ROC curve area
Month-based data set	LR	0.55
	NB	0.54
	KNN	0.51
	Random Forest	0.52
	XGBoost	0.54
	LSTM-XGBoost	0.81
	GRU	*0.84*
Week-based data set	LR	0.59
	NB	0.56
	KNN	0.53
	Random Forest	0.56
	XGBoost	0.61
	LSTM-XGBoost	0.76
	GRU	*0.86*

Source(s): Created by authors

4.3 Feature importance

The process of determining and prioritizing the variables (or features) in a predictive model according to how well they contribute to the model's prediction accuracy is known as feature importance. It facilitates comprehension of the behavior of the model, the underlying structure of the data, and the link between features and the prediction objective (Saarela and Jauhiainen, 2021). The model's performance, interpretability and generalization may all be enhanced by choosing, modifying, or even removing features using feature significance scores. The relevance of each feature in predicting an outcome is determined by ranking them based on their importance scores, which are obtained using the Random Forest classic machine learning model and shown in Table 10:

- Homepage: this feature has the highest importance score, suggesting that interactions or activities related to the homepage are most predictive or influential in determining the model's outcome. This implies that engagements on the homepage are critical indicators.
- Forumng: the second most important feature is related to forum engagements. This high importance score indicates that participation in forums is a significant predictor.
- Subpage: this suggests that navigation to subpages within the platform is also an important behavior in predicting

Table 10 Feature importance

Feature	Importance
homepage	*0.235803*
forumng	*0.192346*
subpage	*0.128884*
quiz	*0.109694*
oucontent	*0.106909*
resource	*0.102140*
url	0.073573
glossary	0.020130
oucollaborate	0.016494
questionnaire	0.007927
ouelluminate	0.004556
sharedsubpage	0.001544

Source(s): Created by authors

the outcome, though less so than homepage interaction and forum participation.

- Quiz: quiz interactions are the fourth most impactful, highlighting the relevance of assessments in the predictive model. This relates to academic performance.
- Oucontent and Resource: these features are closely ranked and suggest that content and resource utilization are nearly as influential as quiz interactions in the model.
- URL: general URL interactions have a moderate impact, indicating the importance of how students navigate and access various links within the platform.
- Glossary, Oucollaborate, Questionnaire, Ouelluminate and Sharedsubpage: these features have progressively lower importance scores, suggesting they have less influence on the model's predictions. Activities like using a glossary, collaborating, filling questionnaires, participating in ouelluminate sessions and accessing shared subpages play smaller roles in the outcome being predicted.

For experimentation, we select the first six features with the highest importance values (homepage, forumng, subpage, quiz, oucontent and resource) for traditional models. For GRU and LSTM-XGBoost, we do not use the feature importance method. This is because both models have the capability to weigh features within the neural network mechanism. To compare and interpret the results of traditional machine learning models after feature importance with those before feature importance, we examined the changes in performance metrics (accuracy, precision, recall, F1 score) for both month-based and week-based data sets. According to Table 11, improvements are observed in the performance of LR, NB (notably improved from a much lower baseline), KNN, Random Forest and XGBoost, with each model showing increased accuracy, precision, recall and F1-scores. A significant discrepancy in NB's performance is noted (a drastic drop in accuracy and recall, leading to a much lower F1-score), which has been rectified after feature importance. The process of feature importance has led to either improvement of model performances for traditional machine learning models. This indicates that removing irrelevant features helps in focusing the model on the most relevant information, potentially reducing overfitting and improving generalization.

According to Table 12, the AUC values for traditional models are generally low but show a slight improvement or consistency for LR, NB, KNN, Random Forest and XGBoost when compared to before feature importance. The highest AUC value in week-based data set after feature importance is 0.60 for XGBoost, suggesting a modest improvement in the model's ability to distinguish between classes. The slight decrease in AUC values for some models after feature selection in the month-based data set suggests that the removal of certain features has reduced the models' ability to distinguish between classes. However, the changes are marginal, indicating that the impact of feature importance on the discriminative ability of these models is limited. The generally low AUC values for traditional models in both data sets after feature importance highlight a challenge in achieving high discriminative ability. The superior performance of the GRU and LSTM-XGBoost models before feature importance suggests that more complex models are necessary to effectively capture the underlying patterns in the data for these tasks.

5. Discussion and implications

This section focuses on analyzing the study's findings in relation to the research questions, outlining their broader significance for educational technology and pedagogy. The key goal is to evaluate the utility of clickstream data and machine learning algorithms in predicting student performance and enhancing online learning experiences. Additionally, the

Table 11 Performance metrics of traditional models for month-based and week-based data sets after feature importance

Data set	Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Month-based data set	LR	80.44	80.44	100	89.16
	NB	80.44	80.44	100	89.16
	KNN	76.73	80.40	93.98	86.66
	Random Forest	78.61	80.44	96.99	87.95
	XGBoost	80.05	80.45	99.33	88.90
Week-based data set	LR	83.40	83.40	100	90.95
	NB	81.71	83.49	97.31	89.87
	KNN	80.25	83.72	94.73	88.89
	Random Forest	81.68	83.56	97.15	89.84
	XGBoost	83.33	83.40	99.89	90.90

Source(s): Created by authors

Table 12 ROC curve area of traditional models for month-based and week-based data sets after feature importance

Data set	Models	ROC curve area
Month-based data set	LR	0.54
	NB	0.53
	KNN	0.49
	Random Forest	0.51
	XGBoost	0.52
Week-based data set	LR	0.59
	NB	0.56
	KNN	0.54
	Random Forest	0.55
	XGBoost	0.60

Source(s): Created by authors

discussion highlights the comparative effectiveness of different machine learning models and contextualizes these findings within existing literature. The section also addresses the implications for educators and institutions, as well as the limitations of the current research, before proposing directions for future work.

5.1 Discussion

This section elaborates on the findings to address the key investigation points and discusses their implications for educational technology and pedagogy. The study emphasizes the transformative potential of leveraging clickstream data and machine learning algorithms to predict student performance and enhance online learning experiences. By comparing the findings with previous research, we also explore their practical implications for educators and propose directions for future research.

To address the first research question, the study demonstrates that clickstream data, when combined with machine learning, provides an effective approach to predicting student outcomes in virtual learning environments. The results reveal that patterns of engagement and activity can be used to identify students who may be at risk of underperforming or dropping out. This finding aligns with prior research, such as (Aljohani *et al.*, 2019) and (Mubarak *et al.*, 2021b), which also achieved high predictive accuracy using advanced machine learning methods. By enabling early detection, these models empower educators to implement timely, targeted interventions, enhancing student retention and performance. The importance of data-driven strategies in improving the effectiveness of online education is further underscored by these findings.

The second research question investigates the comparative advantages of advanced time-series machine learning models over traditional methods. The study reveals that models such as GRU and LSTM-XGBoost outperform simpler algorithms by effectively capturing complex temporal patterns in student behavior, achieving an accuracy of 90.13%. These findings highlight the potential of modern machine learning models to deliver nuanced insights into student engagement trends, facilitating more precise and impactful educational interventions. Compared to traditional approaches, advanced time-series models provide a deeper understanding of the factors influencing student success in virtual learning environments.

The study's findings are contextualized by comparing them with prior research on clickstream data and machine learning in education. As outlined in the literature review, Table 1 presents a summary of several studies that utilized clickstream data to predict student performance. The table highlights a diverse range of data sets, methods and accuracies, showcasing the evolution of predictive approaches in this field. Our model's accuracy of 90.13% is comparable to, or slightly exceeds, that of similar studies, such as (Aljohani *et al.*, 2019) and (Mubarak *et al.*, 2021b), which reported accuracies of 90% and 89%, respectively. These results underscore the reliability and effectiveness of the proposed method, marking significant advancements in machine learning applications for educational data mining. The findings affirm that combining clickstream data with advanced machine learning models offers a scalable and impactful solution for improving online education outcomes.

5.2 Implications for teaching and learning

The findings of this study have significant implications for educational practice, particularly in the realm of online learning. By integrating clickstream data and machine learning algorithms, institutions can create more responsive, personalized, and data-driven environments. Below are some key implications and recommendations based on the study's findings:

- *Real-Time Monitoring and Proactive Support:* Educators can utilize predictive analytics to monitor student performance in real time, enabling proactive support for at-risk students. Real-time monitoring allows for the early identification of students facing academic challenges or disengagement. This early detection is crucial in preventing students from falling too far behind. Institutions should invest in user-friendly dashboards that enable instructors to view and analyze student behavior in real time. These dashboards should highlight students showing early signs of distress, such as inactivity or poor assessments, prompting immediate intervention. Educators can use data to send personalized reminders or check-ins, offering support or directing students to resources like tutoring, counseling, or study groups. Colleges and universities should also develop early warning systems based on predictive models to alert instructors or advisors when intervention is necessary, facilitating a proactive approach to student success. By leveraging these predictive capabilities, interventions become more timely, specific and tailored to individual needs, enhancing the overall learning experience as students feel supported and engaged.
- *Adaptive and Personalized Learning Strategies:* The insights generated by predictive models enable the design of adaptive and personalized learning strategies. By analyzing student interactions with course materials, predictive models identify patterns related to learning preferences, strengths, and areas for improvement. This information allows educators to craft customized learning paths that enhance student engagement and retention. Institutions should adopt or develop adaptive learning systems that adjust the difficulty, content, and format of materials based on student performance and engagement. For example, if a student struggles with a concept, the system could provide additional resources or alternative explanations tailored to their learning style. Educators can use learning analytics to individualize content

delivery, ensuring that struggling students receive additional practice while advanced learners can progress more quickly. This approach could involve modular content, where students focus on areas needing improvement, creating a flexible and efficient learning environment. Additionally, by combining predictive analytics with gamification, educators can enhance engagement by presenting students with challenges and rewards that motivate them to stay on track and improve performance. This adaptive environment improves motivation, sustains interest, and directly impacts retention and completion rates.

- *Integration of Learning Analytics into Institutional Practices:* The effectiveness of predictive analytics provides a foundation for policymakers to advocate for the broader integration of learning analytics into institutional practices. These predictive models can guide individual teaching strategies and broader institutional decisions, including curriculum design and student support services. Institutions should develop a comprehensive data strategy that collects and analyzes diverse student data, including academic performance, engagement metrics and noncognitive factors like motivation and well-being. Integrating data from multiple sources offers a holistic view of student success, enabling more personalized interventions. To ensure successful integration, faculty development programs are crucial. These programs should focus on data literacy, helping instructors interpret and act on data insights. Empowering faculty to use analytics tools will enhance teaching, improve course design, and support student success. Policymakers should use data-driven insights to advocate for changes in educational policy, such as modifying course structures, improving support systems, or refining assessments. Data can help justify investments in learning technologies, making it easier to allocate resources effectively. By fostering a culture of data-driven decision-making, institutions can improve not only individual courses but also the overall student experience, leading to better educational outcomes.
- *Ethical Considerations and Data Privacy:* The use of predictive analytics in education presents important ethical considerations. The collection and use of student data must be done ethically, transparently, and responsibly to avoid privacy concerns and biases in predictive models. Institutions should ensure that students are fully informed about the data being collected, its intended use and the potential benefits for their learning. Transparent communication and explicit consent will help build trust and alleviate privacy concerns. It's essential to monitor and reduce biases in machine learning algorithms. Institutions should implement bias audits to ensure that predictive models do not disproportionately flag certain student groups as "at-risk" based on factors like ethnicity, gender, or socioeconomic background. Data privacy and security should remain top priorities, and institutions must invest in robust cybersecurity measures to protect student data and comply with regulations.
- *Future Directions and Continuous Improvement:* To maximize the impact of predictive analytics on teaching and learning, continuous improvement of tools and strategies is crucial. Educational institutions should foster an environment of ongoing feedback and iteration, regularly evaluating and refining models and teaching methods. Institutions should support research initiatives that explore the effectiveness of predictive analytics in different educational contexts. Regular evaluation of learning models and intervention strategies ensures they remain relevant and effective in addressing evolving student needs. Collaboration between institutions can facilitate the sharing of best practices, tools, and models, enabling the creation of standardized metrics and benchmarks to compare the effectiveness of various learning analytics approaches. Ultimately, all efforts should aim at improving the student experience. Involving students in the development and feedback process will ensure that predictive analytics tools are user-friendly, helpful, and aligned with student expectations.

The integration of predictive analytics and clickstream data into online education has the potential to revolutionize teaching and learning. By leveraging these tools, educators can offer more personalized, adaptive, and timely interventions, improving engagement and retention. Institutions, guided by data-driven decision-making, can provide a more supportive and responsive learning environment, leading to better outcomes for students. However, careful attention must be given to ethical considerations, data privacy, and continuous improvement to ensure the effective and responsible use of these tools.

6. Conclusion

The essential advances in comprehending and improving student performance in virtual learning environments made possible by the combination of clickstream data analysis and machine learning algorithms must be emphasized to wrap up the study. The research shows how sophisticated time-series models may provide a more in-depth understanding of student engagement trends, which can lead to more successful interventions. This study adds to the body of knowledge on educational technology while also providing educators and policymakers with valuable recommendations for enhancing the effectiveness of online learning.

While the study presents significant contributions, there are notable limitations. The analysis is conducted on specific data sets, which may limit the generalizability of the findings to other educational contexts or systems. Additionally, the study focuses on short- to medium-term predictions, leaving long-term forecasting unexplored. Another limitation is the technical expertise required to implement and maintain advanced machine learning models, which might pose challenges for institutions lacking adequate resources. These limitations suggest avenues for future research, which could include testing the models on diverse data sets to enhance generalizability, exploring the integration of multimodal data sources (e.g., forum posts, quiz scores, and video interactions) to improve prediction accuracy, and examining long-term patterns in student performance. Furthermore, future research could address the ethical implications of using student data for predictive analytics, ensuring that privacy and equity are maintained in all applications.

References

Acito, F. (2023), "Logistic regression", *Predictive Analytics with Knime: Analytics for Citizen Data Scientists*, Springer, pp. 125-167.

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A.A., Abid, M. and Khan, S.U. (2021), "Predicting at-risk students at different percentages of course length for early intervention using machine learning models", *IEEE Access*, Vol. 9, pp. 7519-7539.

Al-Azazi, F.A. and Ghurab, M. (2023), "Ann-lstm: a deep learning model for early student performance prediction in mooc", *Heliyon*, Vol. 9 No. 4, p. e15382.

Aljohani, N.R., Fayoumi, A. and Hassan, S.-U. (2019), "Predicting at-risk students using clickstream data in the virtual learning environment", *Sustainability*, Vol. 11 No. 24, p. 7238.

Aouifi, H.E., Hajji, M.E., Es-Saady, Y. and Douzi, H. (2020), "Predicting learner's performance through video viewing behavior analysis using graph convolutional networks", *2020 fourth international conference on intelligent computing in data sciences (icds)*, pp. 1-6, doi: 10.1109/ICDS50568.2020.9268730.

Asselman, A., Khaldi, M. and Aammou, S. (2023), "Enhancing the prediction of student performance based on the machine learning xgboost algorithm", *Interactive Learning Environments*, Vol. 31 No. 6, pp. 3360-3379.

Aydin, Z.E. and Ozturk, Z.K. (2021), "Performance analysis of xgboost classifier with missing data", *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)*, Vol. 2 No. 2, p. 2021.

Baig, M.A., Shaikh, S.A., Khatri, K.K., Shaikh, M.A., Khan, M.Z. and Rauf, M.A. (2023), "Prediction of students performance level using integrated approach of ml algorithms", *International Journal of Emerging Technologies in Learning*, Vol. 18 No. 1.

Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.-Y. and Hussain, A. (2023), "Educational data mining to predict students' academic performance: a survey study", *Education and Information Technologies*, Vol. 28 No. 1, pp. 905-971.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M.R. (2021), "Variable selection for naïve bayes classification", *Computers & Operations Research*, Vol. 135, p. 105456.

Casey, K. and Azcona, D. (2017), "Utilizing student activity patterns to predict performance", *International Journal of Educational Technology in Higher Education*, Vol. 14 No. 1, pp. 1-15.

Chu, Y.-W., Tenorio, E., Cruz, L., Douglas, K., Lan, A.S. and Brinton, C.G. (2021), "Click-based student performance prediction: a clustering guided meta-learning approach", *2021 ieee international conference on big data (big data)*, pp. 1389-1398, doi: 10.1109/BigData52589.2021.9671729.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014), "Empirical evaluation of gated recurrent neural networks on sequence modeling", *arXiv preprint arXiv:1412.3555*.

Coelho, O.B. and Silveira, I. (2017), "Deep learning applied to learning analytics and educational data mining: a systematic literature review", Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie), Vol. 28, p. 143.

Dey, R. and Salem, F.M. (2017), "Gate-variants of gated recurrent unit (gru) neural networks", *2017 ieee 60th international midwest symposium on circuits and systems (mwscas)*, pp. 1597-1600, doi: 10.1109/MWSCAS.2017.8053243.

Fan, Z., Huang, Y., Xi, C. and Liu, Q. (2023), "Multi-view adaptive k-nearest neighbor classification", *IEEE Transactions on Artificial Intelligence*, Vol. 5 No. 3.

Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y.K. and Doneva, R. (2022), "Exploring online activities to predict the final grade of student", *Mathematics*, Vol. 10 No. 20, p. 3758.

Géron, A. (2022), *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow*, O'Reilly Media.

Haddouchi, M. and Berrado, A. (2019), "A survey of methods and tools used for interpreting random Forest", *2019 1st international conference on smart systems and data science (icssd)*, pp. 1-6.

Hao, J., Gan, J. and Zhu, L. (2022), "Mooc performance prediction and personal performance improvement via bayesian network", *Education and Information Technologies*, Vol. 27 No. 5, pp. 7303-7326.

He, J., Bailey, J., Rubinstein, B. and Zhang, R. (2015), "Identifying at-risk students in massive open online courses", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29 No. 1.

Hlosta, M., Zdrahal, Z. and Zendulka, J. (2017), "Ouroboros: early identification of at-risk students without models based on legacy data", *Proceedings of the seventh international learning analytics & knowledge conference*, pp. 6-15.

Huang, A.Y., Lu, O.H., Huang, J.C., Yin, C. and Yang, S.J. (2020), "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs", *Interactive Learning Environments*, Vol. 28 No. 2, pp. 206-230.

Karim, F., Majumdar, S., Darabi, H. and Chen, S. (2017), "Lstm fully convolutional networks for time series classification", *IEEE Access*, Vol. 6, pp. 1662-1669.

Katrutsa, A. and Strijov, V. (2017), "Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria", *Expert Systems with Applications*, Vol. 76, pp. 1-11.

Khoushehgir, F. and Sulaimany, S. (2023), "Negative link prediction to reduce dropout in massive open online courses", *Education and Information Technologies*, Vol. 28 No. 8, pp. 1-20.

Kóröси, G., Esztelecki, P., Farkas, R. and Tóth, K. (2018), "Clickstream-based outcome prediction in short video moocs", *2018 international conference on computer, information and telecommunication systems (cits)*, pp. 1-5, doi: 10.1109/CITS.2018.8440182.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017), "Open university learning analytics dataset", *Scientific Data*, Vol. 4 No. 1, pp. 1-8.

Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S. and Wei, Y. (2022), "Predicting student performance using clickstream data and machine learning", *Education Sciences*, Vol. 13 No. 1, p. 17.

Mubarak, A.A., Cao, H. and Ahmed, S.A. (2021a), "Predictive learning analytics using deep learning model in moocs' courses videos", *Education and Information Technologies*, Vol. 26 No. 1, pp. 371-392.

Mubarak, A.A., Cao, H., Zhang, W. and Zhang, W. (2021b), "Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in moocs'

courses", *Computer Applications in Engineering Education*, Vol. 29 No. 4, pp. 710-732.

Nti, I.K., Nyarko-Boateng, O. and Aning, J. (2021), "Performance of machine learning algorithms with different k values in k-fold cross-validation", *International Journal of Information Technology and Computer Science*, Vol. 13 No. 6, pp. 61-71.

Ouyang, F., Zheng, L. and Jiao, P. (2022), "Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020", *Education and Information Technologies*, Vol. 27 No. 6, pp. 7893-7925.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K. (2023), "Classification and prediction of student performance data using various machine learning algorithms", *Materials Today: Proceedings*, Vol. 80, pp. 3782-3785.

Park, J., Denaro, K., Rodriguez, F., Smyth, P. and Warschauer, M. (2017), "Detecting changes in student behavior from clickstream data", *Proceedings of the seventh international learning analytics & knowledge conference*, pp. 21-30.

Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R. and Kundu, S. (2018), "Improved random Forest for classification", *IEEE Transactions on Image Processing*, Vol. 27 No. 8, pp. 4012-4024, doi: 10.1109/TIP.2018.2834830.

Rizvi, S., Rienties, B. and Khoja, S.A. (2019), "The role of demographics in online learning; a decision tree based approach", *Computers & Education*, Vol. 137, pp. 32-47.

Saarela, M. and Jauhiainen, S. (2021), "Comparison of feature importance measures as explanations for classification models", *SN Applied Sciences*, Vol. 3 No. 2, pp. 1-12.

Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A. and Ashraf, I. (2022), "Predicting students' academic performance with conditional generative adversarial network and deep svm", *Sensors*, Vol. 22 No. 13, p. 4834.

Smith, D., Khorsandroo, S. and Roy, K. (2023), "Supervised and unsupervised learning techniques utilizing malware datasets", *2023 ieee 2nd international conference on ai in cybersecurity (icaic)*, pp. 1-7.

Tang, H., Tang, Y., Su, Y., Feng, W., Wang, B., Chen, P. and Zuo, D. (2024), "Feature extraction of multi-sensors for early bearing fault diagnosis using deep learning based on minimum unscented kalman filter", *Engineering Applications of Artificial Intelligence*, Vol. 127, p. 107138.

Taunk, K., De, S., Verma, S. and Swetapadma, A. (2019), "A brief review of nearest neighbor algorithm for learning and classification", *2019 international conference on intelligent computing and control systems (iccs)*, pp. 1255-1260.

Tharwat, A. (2020), "Classification assessment methods", *Applied Computing and Informatics*, Vol. 17 No. 1, pp. 168-192.

Vujović, Ž. *et al.* (2021), "Classification model evaluation metrics", *International Journal of Advanced Computer Science and Applications*, Vol. 12 No. 6, pp. 599-606.

Waheed, H., Hassan, S.-U., Aljohani, N.R., Hardman, J., Alelyani, S. and Nawaz, R. (2020), "Predicting academic performance of students from vle big data using deep learning models", *Computers in Human Behavior*, Vol. 104, p. 106189.

Waheed, H., Hassan, S.-U., Nawaz, R., Aljohani, N.R., Chen, G. and Gasevic, D. (2023), "Early prediction of learners at risk in self-paced education: a neural network approach", *Expert Systems with Applications*, Vol. 213, p. 118868.

Wang, X., Guo, B., Shen, Y., *et al.* (2022), "Predicting the at-risk online students based on the click data distribution characteristics", *Scientific Programming*, Vol. 2022.

Wong, T.-T. and Yeh, P.-Y. (2019), "Reliable accuracy estimates from k-fold cross validation", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 32 No. 8, pp. 1586-1594.

Wu, B., Qu, S., Ni, Y., Zhou, Y., Wang, P. and Li, Q. (2019), "Predicting student performance using weblogs", *2019 14th international conference on computer science and education (iccse)*, pp. 616-621, doi: 10.1109/ICCSE.2019.8845440.

Yang, T.-Y., Brinton, C.G., Joe-Wong, C. and Chiang, M. (2017), "Behavior-based grade prediction for moocs via time series neural networks", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11 No. 5, pp. 716-728.

Zang, X., Du, J. and Song, Y. (2023), "Early prediction of heart disease via lstm-xgboost", *Proceedings of the 2023 9th international conference on computing and artificial intelligence*, pp. 631-637.

Zerkouk, M., Mihoubi, M., Chikhaoui, B. and Wang, S. (2024), "A machine learning based model for student's dropout prediction in online training", *Education and Information Technologies*, pp. 1-20.

Zhou, Q., Zheng, Y. and Mou, C. (2015), "Predicting students' performance of an offline course from their online behaviors", *2015 fifth international conference on digital information and communication technology and its applications (dictap)*, pp. 70-73, doi: 10.1109/DICTAP.2015.7113173.

Zou, X., Hu, Y., Tian, Z. and Shen, K. (2019), "Logistic regression model optimization and case analysis", *2019 ieee 7th international conference on computer science and network technology (iccsnt)*, pp. 135-139.

Corresponding author

Zakaria Khoudi can be contacted at: zakaria.khoudi@usms.ma