Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Modelling User Behaviour from Log Information in an Online Learning Management System

Siti Khairuna Hamdiah[1], Arif Bramantoro[1*], Serina Mohd Ali[1], Ahmad A. Alzahrani[2]

[1] School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei Darussalam
[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
E-mail: arif.bramantoro@utb.edu.bn

Received: 8 June 2025; **Revised:** 24 June 2025; **Accepted:** 24 July 2025

Abstract: An online Learning Management System (LMS) is a web-based technology used to organise, implement, and evaluate a learning process. In this context, the use of automated techniques is required to cluster the activities of students, allowing for better identification of their behaviours from large data sets. The initial phase in developing such an automated approach is to collect log information. An online LMS is suitable for this purpose, as it contains a large user behaviour data set. In this study, activity and process information are extracted from such a data set, followed by analysis of user behaviours to obtain various insights, such as the correlation between the learning process and the academic performance of students. By discovering processes, checking performance, and analysing engagement from the log data obtained from an online LMS, user behaviours can be successfully identified. In addition, the results demonstrate that the behaviours of students slightly affect their grades; for example, students with strong engagement in the online LMS tend to achieve higher marks. To support the significance of this study, a proposed scheme is provided along with examples and results. The key goal of this study is to identify student behaviours. By comparing different algorithms that can be applied in the context of this study, the results indicate that the k-Means clustering algorithm and Naïve Bayes classification algorithm are suitable methods for this purpose.

***Keywords*:** learning management system, process mining, k-means clustering, Naïve Bayes classification

MSC: 68T01, 68P01, 62F15

Abbreviation

LMS	Learning Management System
CRISP-DM	Cross-Industry Standard Process for Data Mining
IBK	Three letter acronym

1. Introduction

The use of the k-Means clustering and Naïve Bayes classification methods can be expected to improve the analysis of system user behaviour based on log data providing, for example, the times at which each user logs in and out, changes

DOI: https://doi.org/10.37256/cm.6520257400

made to the data, and other information. In this context, collecting log information from the systems that students and instructors use can facilitate relevant analyses. In this study, we examine an online Learning Management System (LMS) that contains a data set potentially capable of describing user behaviour extract appropriate activity and process information and then analyse the behaviours of system users to obtain various insights, such as the correlations between the learning process and student academic performance.

This study aims to analyse the academic performance of students based on data obtained from an online LMS. To achieve this aim, there are several objectives that must be accomplished. The first is to identify user behaviours based on their activities using process mining. The second is to use *k*-Means clustering and Naïve Bayesian classification to analyse student behaviours. The third is to obtain insights regarding the academic performance of students based on the data and process mining results.

Given the varied usage patterns among students and instructors, understanding how LMS engagement correlates with academic performance is essential. Such insights can inform teaching strategies and encourage more effective use of the platform. Understanding of these relationships is important for the instructor, in order to evaluate the effectiveness of the course flow, as well as for the students, in order to motivate them to use the LMS effectively based on their overall and individual performance in the course. Hence, there is a need to assess student behaviours from data obtained from LMSs, using analysis techniques to determine the relevant processes and exploring their relationships with academic performance through correlation analysis.

In this research, it is argued that technological developments can overcome the issues inherent to online learning, thus increasing its quality. The associated behavioural process is affected by both the materials provided to the students and their individual learning patterns. Moreover, once determined, they can be utilised to motivate them to benefit from such technology.

2. Background study

This section presents an overview of related studies that have identified user behaviours based on their log data, the implementation of machine learning algorithms.

There are two key research works that have considered the implementation of the *k*-Means algorithm for user behaviour prediction [1, 2]. Based on network log data from an educational institution, the first related work [1] focused on the actions of internet users. Their goal was to discover what resources were commonly accessed, such that they could identify behaviours in the internet activities of users. The data utilised in this study were obtained from a one-week observation period at an educational institution in Indonesia. The *k*-Means technique was used to examine the data using two software tools: Statistical Package for the Social Sciences (SPSS) and RapidMiner. A pre-processing stage was also conducted, in order to remove unnecessary data. The results indicated that the educational institution's internet was mainly used to access search engines, information websites, and social media websites. It is interesting to note that constraints were imposed on the source data during the profiling process. To achieve the best results in the profiling process, the data should include information about computer-related activities.

The purpose of the second related work [2] was to broaden the understanding of identifying user behaviours in collaborative online social networks by analysing the benefits of a quantitative technique based on clustering algorithms. This approach was assessed through comparison with the user behaviours assessed according to traditional qualitative methodology. This procedure is crucial as, although there are many tools available to make the process simpler, analysis based on user observations remains time-consuming. By gathering enough observational data, it was possible to fulfil the research objective. The benefit of utilizing quantitative approaches was represented by the statistical certainty of the analysis on all users with a minimum confidence level. The techniques used in the first paper [1] to examine data (i.e., the Means algorithm in SPSS and RapidMiner) were more accurate, as the unnecessary data were identified and screened using the pre-processing technique.

Another two key research papers have focused on the implementation of different algorithms for user behaviour prediction [3, 4]. The first relate work [3] proposed a strong classifier to predict buying intentions based on user activity data from a large e-commerce website. The results of this study are significant as online merchants may gain a deeper knowledge of the actions and intentions of consumers, allowing them to expand their market by tracking relevant search patterns. Two classifiers-logistic regression and random forest-were used to assess the data. The first result of the study was that logistic regression did not present the same advantages as the random forest algorithm. Another finding was that networks pre-trained with Stacked Denoising Autoencoders reached the highest accuracy for deep learning, due to the sparseness of the considered data set. As the results clearly indicated the gap in the performance improvement when compared with a larger data size, it would be interesting to determine the effect when using a much larger volume of training data.

The second related work [4] was to focus on the detection and identification of user behaviours through web usage mining with the main goal being to identify the most and least popular web pages. It is important to understand how website users interact with a site, as this may provide guidance on how to best restructure it and reduce confusion among the users. The technique utilised to gather log data was web usage mining, and the sub-tasks included resource searching, information selection, and pre-processing. The data were analysed using a novel technique called visitor online behaviour analysis. Their findings indicated that the proposed algorithm can efficiently classify the preferences of users for various categories of websites. In addition, web usage mining has shown considerable promise and is commonly used for activities such as web personalization, pre-fetching, and rearrangement. In this paper, we adapt this finding due to its demonstrated efficacy in categorizing people in several types of website users.

A further two research works [5, 6] have considered implementation of the *k*-Means algorithm to achieve other objectives. The purpose of the first related work [5] is to get the capacity to track the academic development of students. This approach was designed in order to develop methods to improve learning and academic performances by tracking the progress of the students. The data were analysed using the *k*-Means clustering algorithm and the Euclidean distance measure. The paper produced several key findings, as follows: First, the *k*-Means clustering algorithm serves as a good benchmark for monitoring the progression of student performance; second, it improves the decision-making of academic planners by monitoring the performance of students every semester, allowing for improvement of future academic results in subsequent academic sessions.

The second related work [6] discussed the development of an application to assist universities in categorising the IT skills of first-year students, based on their IT proficiency and experience. The aim of this method was to prevent any favouritism in the university. In the results of this study, the Python program achieved the best outcome (of 68% accuracy) when using 50 data samples to predict the grade range based on the IT experience of the students. The division of classes in the institution demonstrated the success of the study, as the results obtained in practice had high accuracy. The steps elaborated in the paper were presented in detail and can be considered useful for future research.

Two key research works have considered the use of several types of log information [7, 8]. In the first related work [7], it was discussed how a database system is typically located deep within the business network, which gives insiders the opportunity to attack, breach, and eventually steal the data. The goal of the paper was to develop a technique to reduce the likelihood of a successful attack on the organisation's database while simultaneously improving its performance. Three steps were followed to analyse the data: Using software on database server that allows for efficient log collection and secure data transfer, analysing the collected log data based on reports and searches, and retaining log data that are directly mandated by several regulations.

It is important to possess a logging system that meets current operational, security, and compliance standards. This results in the flooding of audit log data and the adoption of log management solutions. While vendor-specific database tools might be useful, a complete log management solution is a better option in most cases.

The second related work [8] investigated the log analysis capability of major commercial products. The significance of the study is that the context-aware result obtained through structured analysis allows analysts to automate operations by implementing result selection according to context information. Therefore, structured analysis is advantageous for automation. The findings in this research included existing log data analysers that solve a subset of the problems encountered by log analysts, a new framework that provides effortless structural extraction of log data by expressing the format of a log file in a declarative language, and a strong platform that is required for log data processing, but not for the user interface engine. This work provided relevant results by dividing them into several successful outcomes; however, it was not as accurate as the first paper, and no solution was presented.

3. Learning management systems

A LMS is a software program or web-based technology that is used to organise, implement, and evaluate a particular learning process. An LMS is important for helping users to better organise the educational process [9] and to ensure that they are developing transferrable skills. In this research, we focus on the students: the system must be designed with a user-friendly interface such that the students can easily figure out how to work with the system. Applying intelligence components to the system can provide a better experience that suits the needs of users.

Student engagement in a traditional classroom tends to have a positive effect on grades, with students achieving higher grades as their participation increases. In terms of LMSs, a previous study [10] has examined student behaviours when participating in both LMSs and face-to-face courses. This study aimed to identify two factors: the characteristics of LMS participation that the student accomplishes, measured by the course grade, and the range of participation profiles that are used to describe the LMS activity of students.

The collected data described the activity of 48 senior-level students enrolled in 400-level courses. SPSS was used to generate descriptive statistics and explore the correlations between the LMS activity and course grades of students. The k-Means clustering method was chosen, as all of the standardised input variables were in a continuous format [11]. The outcome of this approach was consistent with the expectation that there exists a positive relationship between LMS participation and student achievement.

Traditionally, institutions group students according to their grade points. One study [12] aimed to develop a system that identifies students based on their academic performance with a fair preference. A total of 20 records relating to students from the Department of Information and Computer Science were used as the data set, and k-Means clustering was used to strengthen the prediction of student performance. The findings were considered effective in categorising the performance into three clusters, and the result was expected to overcome the difficulty in identifying students.

Other works have focused on classification to predict student behaviours in learning management systems [13, 14]. The authors of [13] claimed that monitoring user behaviour is necessary. Two pre-processing techniques were used: A fuzzy rule-based system and linear regression. Fuzzy rules were proposed for network user web behaviour classification while linear regression was used for prediction. From the simulation results, it was inferred that the proposed technique has high potential for the classification of users. The logs were retrieved from a central server and the conclusion was that the algorithm used was a success.

Another study stated that the measurement of student performance is an important part of the education system [4]. The use of data mining in an LMS relies on the use of data to enhance teaching and learning methods, as well as to predict the achievement level of students. The data used in this study were collected from various courses at Eindhoven University of Technology from 2014 to 2015.

The authors of [14] investigated factors affecting students' readiness towards using the Blackboard LMS in the context of the University of Ha'il in Saudi Arabia. The study employs the Student Online Learning Readiness model to assess technical, social, and communication competencies' impact on students' readiness. The research findings suggest that all three competencies significantly influence students' readiness to use Blackboard LMS. This research is exclusively centred on a singular university, offering a rationale for our investigation aimed at conducting a comprehensive examination of our institution. Similar findings were obtained by examining differences in student characteristics at the beginning of the course, as they had a positive influence on predicting their final grades. The grade became less relevant as the semester progressed, and the primary elements became more important in predicting a better grade.

Although both of the abovementioned research papers provided interesting predictions, the outcome from the second paper was not finalised as the data shifted at the end of the study. This was due to the unclear result of the decision tree J48 algorithm; therefore, only results obtained using the ID3 algorithm were presented.

A growing body of literature highlights the importance of cloud-based infrastructures and immersive technologies in enhancing educational outcomes. For example, the use of smart technologies in open education settings has been found to improve student autonomy and learning personalization [15]. Moreover, combining cloud platforms with augmented reality fosters interactive learning environments that increase motivation and deepen understanding [16, 17]. These findings support the rationale for mining LMS log data to better understand and predict learner behaviors, aligning with the goals of our study.

4. Methodology

In this research, the Cross-Industry Standard Process for Data Mining (CRISP-DM) [18] is modified and used as a methodology to carry out process mining. There are eight phases in this model.

In detail, the first phase is business understanding, which takes place before proceeding to the other phases. Four tasks are conducted in this phase: Determining the objectives, assessing the situation, determining process mining goals, and producing a project plan.

The second phase is data understanding, which describes the details of the log information used, the algorithms, and the user behaviour. There are also four tasks in this phase: Data collection, describing the data, exploring the data, and verifying the quality of the data. This aligns with prior work on data-centered service composition, where system log information was central to automating and refining service behavior analysis [19].

In the data preparation phase, pre-processing is carried out to make sure that the datasets are ready for use. Pre-processing facilitates better monitoring and improvement of the modelling process. The data preparation tasks include selecting and cleaning the data, the latter of which is achieved by constructing, integrating, and formatting the values in the data sets.

The fourth phase is activity discovery which involves finding events from the log files. The main steps in this phase are preparation, visualisation, and analysis. The fifth phase involves the visualisation of the data (e.g., using charts). This process starts with specifying the number of clusters k to be assigned, followed by initialising k centroids randomly. These processes are repeated, following which each point is assigned to its closest centroid. The repetition is completed once there are no more centroids to be positioned. The k-Means clustering algorithm was chosen due to its efficiency, interpretability, and proven performance in educational data mining contexts where behavioral log data are high-dimensional and unlabeled. We set the number of clusters to $k = 3$, based on domain knowledge anticipating three primary user groups: students, lecturers, and admins. The initial centroids were randomly selected, and the algorithm was run until convergence was achieved (i.e., no change in cluster assignment). To validate the clustering, we examined the coherence of activities within each cluster and compared cluster assignments with known user roles where possible.

Once the clusters are produced, the Naïve Bayesian classification phase is applied. The steps for the proposed classification approach include splitting the data into training and testing sets, followed by training the model and, finally, evaluating the classifier to be used. The Naïve Bayesian algorithm was selected because of its simplicity, fast training speed, and strong performance with large categorical datasets where independence assumptions are reasonably valid. Moreover, Naïve Bayes is robust to noise and requires relatively small training data for effective learning. We used the Weka tool to implement the algorithm, performing a 70/30 train-test split.

The evaluation phase is carried out to verify that the process conducted in the previous phase meets the relevant objectives. The first task in this phase is evaluating the performance results and reviewing the process. We measured the results using precision, recall, and F-measure. We also compared Naïve Bayes with Instance-Based learner (IBK) and decision tree J48 classifiers as baselines. The best performance was achieved by Naïve Bayes, with a nearly perfect F-measure, confirming its suitability for this type of behavioral classification.

The final phase starts with the evaluation results and concludes with a strategy for deployment of the data mining result in the associated context. This final phase includes four tasks: Planning the deployment, monitoring and maintenance, producing the final report, and reviewing the project. Figure 1 depicts the phases of the proposed research.

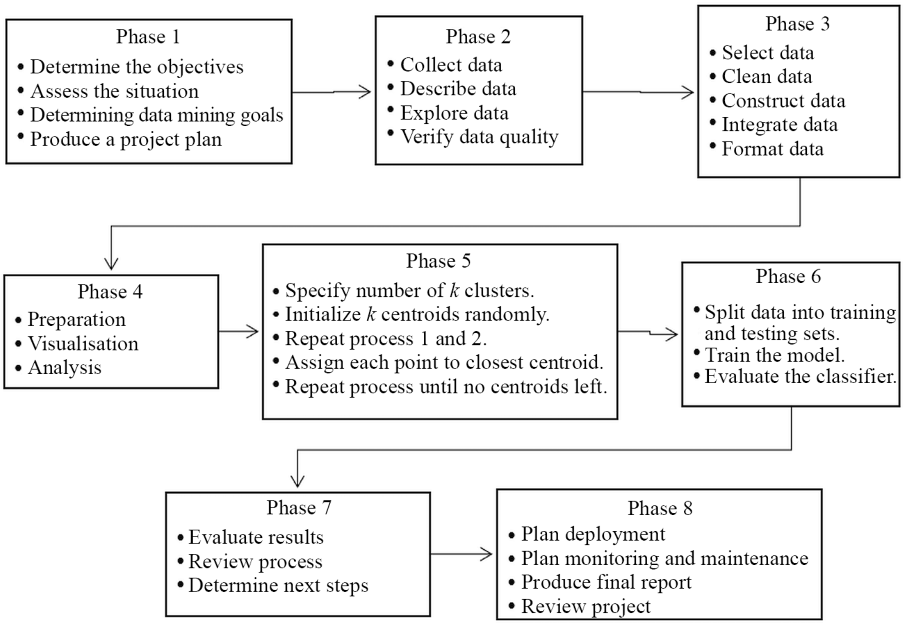


Figure 1. Overview of research phases

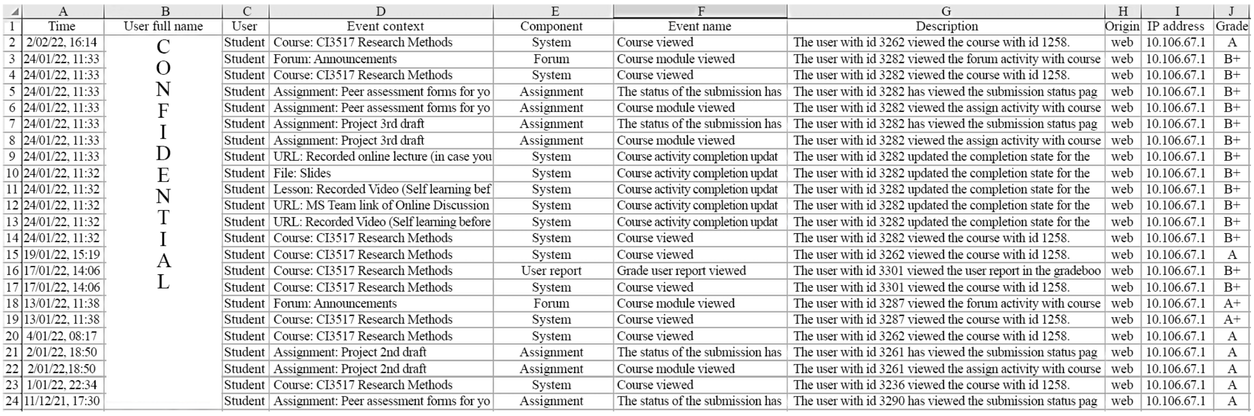
5. Analysis

In this section, we detail the mining and analysis of the LMS logs obtained from a real course module in a particular undergraduate course at our university. Specifically, log data were collected from one undergraduate course module, involving a total of 69 students enrolled during the semester. The module was conducted over a 14-week period, during which students engaged with various LMS activities such as assignment submissions, quiz attempts, content viewing, and grade tracking. The data comprised approximately 26,480 log instances associated with student interactions, which provided a substantial foundation for identifying behavioural patterns and correlating them with final academic performance. This dataset size offers a meaningful representation of student engagement in a typical LMS-enabled course environment at our institution.

We investigate the logs and their contribution to LMS, in order to identify the frequencies of the activities related to the final grades of each student. Our key objective is to identify the effect of a given activity in the learning management system on a student's final grade.

Data pre-processing was first conducted to discover the best outcome from the acquired data. The data used were extracted from the module log files. There were nine features in the original dataset: Time, user full name, affected user, event component, component, event name, description, origin, and IP address. Two additional columns were included in the file: User and grade. These columns were utilised after the process mining phase. For comprehensive understanding of the final data set used for the analyses, Figure 2 shows sample data for one module.

The collected events were grouped according to the activities of each student in the LMS, in order to determine the grades of each student according to their contributions. The events that most often accrued in the LMS were extracted by observing the number of views, submitted files and uploaded files by the students.



	A	B	C	D	E	F	G	H	I	J
1	Time	User full name	User	Event context	Component	Event name	Description	Origin	IP address	Grade
2	2/02/22, 16:14	CONFIDENTIAL	Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3262 viewed the course with id 1258.	web	10.106.67.1	A
3	24/01/22, 11:33		Student	Forum: Announcements	Forum	Course module viewed	The user with id 3282 viewed the forum activity with course	web	10.106.67.1	B+
4	24/01/22, 11:33		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3282 viewed the course with id 1258.	web	10.106.67.1	B+
5	24/01/22, 11:33		Student	Assignment: Peer assessment forms for yo	Assignment	The status of the submission has	The user with id 3282 has viewed the submission status pag	web	10.106.67.1	B+
6	24/01/22, 11:33		Student	Assignment: Peer assessment forms for yo	Assignment	Course module viewed	The user with id 3282 viewed the assign activity with course	web	10.106.67.1	B+
7	24/01/22, 11:33		Student	Assignment: Project 3rd draft	Assignment	The status of the submission has	The user with id 3282 has viewed the submission status pag	web	10.106.67.1	B+
8	24/01/22, 11:33		Student	Assignment: Project 3rd draft	Assignment	Course module viewed	The user with id 3282 viewed the assign activity with course	web	10.106.67.1	B+
9	24/01/22, 11:33		Student	URL: Recorded online lecture (in case you	System	Course activity completion updat	The user with id 3282 updated the completion state for the	web	10.106.67.1	B+
10	24/01/22, 11:32		Student	File: Slides	System	Course activity completion updat	The user with id 3282 updated the completion state for the	web	10.106.67.1	B+
11	24/01/22, 11:32		Student	Lesson: Recorded Video (Self learning bef	System	Course activity completion updat	The user with id 3282 updated the completion state for the	web	10.106.67.1	B+
12	24/01/22, 11:32		Student	URL: MS Team link of Online Discussion	System	Course activity completion updat	The user with id 3282 updated the completion state for the	web	10.106.67.1	B+
13	24/01/22, 11:32		Student	URL: Recorded Video (Self learning before	System	Course activity completion updat	The user with id 3282 updated the completion state for the	web	10.106.67.1	B+
14	24/01/22, 11:32		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3282 viewed the course with id 1258.	web	10.106.67.1	B+
15	19/01/22, 15:19		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3262 viewed the course with id 1258.	web	10.106.67.1	A
16	17/01/22, 14:06		Student	Course: CI3517 Research Methods	User report	Grade user report viewed	The user with id 3301 viewed the user report in the gradeboo	web	10.106.67.1	B+
17	17/01/22, 14:06		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3301 viewed the course with id 1258.	web	10.106.67.1	B+
18	13/01/22, 11:38		Student	Forum: Announcements	Forum	Course module viewed	The user with id 3287 viewed the forum activity with course	web	10.106.67.1	A+
19	13/01/22, 11:38		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3287 viewed the course with id 1258.	web	10.106.67.1	A+
20	4/01/22, 08:17		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3262 viewed the course with id 1258.	web	10.106.67.1	A
21	2/01/22, 18:50		Student	Assignment: Project 2nd draft	Assignment	The status of the submission has	The user with id 3261 has viewed the submission status pag	web	10.106.67.1	A
22	2/01/22,18:50		Student	Assignment: Project 2nd draft	Assignment	Course module viewed	The user with id 3261 viewed the assign activity with course	web	10.106.67.1	A
23	1/01/22, 22:34		Student	Course: CI3517 Research Methods	System	Course viewed	The user with id 3236 viewed the course with id 1258.	web	10.106.67.1	A
24	11/12/21, 17:30		Student	Assignment: Peer assessment forms for yo	Assignment	The status of the submission has	The user with id 3290 has viewed the submission status pag	web	10.106.67.1	A

Figure 2. Module log information

6. Data pre-processing and data analysis

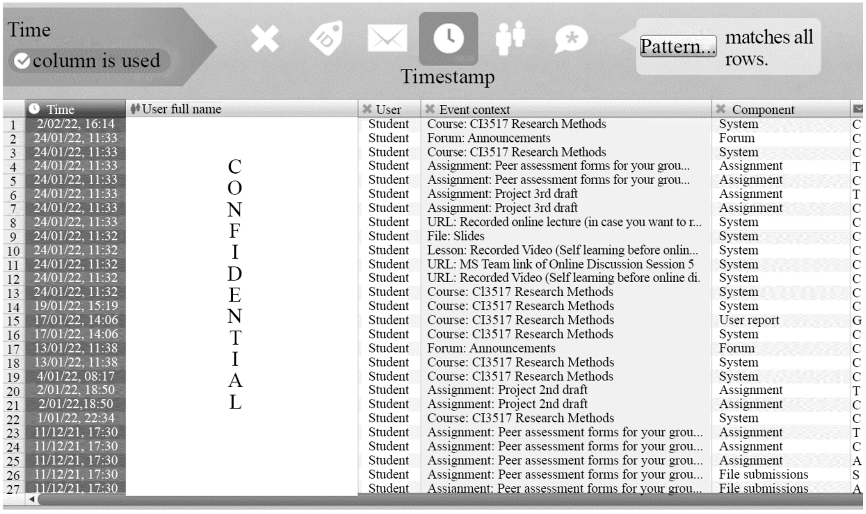
This section details the characteristics of the data. The first step was data cleaning, followed by process mining, data clustering and data classification.

Data cleaning is a strategy used to enhance the quality of data, ensuring that there are no missing data and that no inappropriate data types are used. Data cleaning helps to increase productivity and to produce the best result. The 'search' function was applied to each column, in order to guarantee that there were no missing values. The result was that no missing values were identified; however, an alteration was applied, as the inclusion of a symbol to indicate the user ID in the 'Description' column meant that it could not be read by the process mining tool. Hence, these symbols were deleted before proceeding to the next phase.

Next, feature selection and engineering were conducted. From the original nine raw columns (e.g., Time, User Full Name, Event Component), we extracted and synthesized new features to support behaviour analysis. For example, the columns 'User' and 'Grade' were added. These columns were required for clustering and classification. The 'User' column was divided into three attributes: Student, lecturer, or admin. The 'Grade' column contained the final grade obtained by each student. The 'User full name' column was copied and pasted at the end of the other columns, and the student's name was replaced according to their grades. This was carried out using the 'Replace and Find' function. For the lecturers and admin staff, the grade value was set to 'Null'.

For clustering, the cleaned and structured data were normalized prior to applying the k-Means algorithm. The value of k was set to 3 based on the expected role-based grouping (students, lecturers, admins). Features contributing to clustering included frequency of logins, types of activities (e.g., quiz attempts, file submissions), and session durations. To assess clustering performance, we analysed cluster cohesion, role consistency within clusters, and inter-cluster separation. This structured pre-processing pipeline ensured that only high-quality, behaviourally relevant data were used for downstream analysis.

Process mining helps us to discover the processes in a data set and produce an automated flow diagram according to the processes in the supplied data. The labels that we observed were information related to the activities of students in the LMS. After uploading the data set, the labelled attributes must be identified according to the properties of the LMS. Figure 3 shows the attributes and sample data, whereas Table 1 details the conversion of the attributes. The visualization in Figure 3 was generated using Disco software [20].



	Time	User full name	User	Event context	Component	
1	2/02/22, 16:14		Student	Course: CI3517 Research Methods	System	C
2	24/01/22, 11:33		Student	Forum: Announcements	Forum	C
3	24/01/22, 11:33		Student	Course: CI3517 Research Methods	System	C
4	24/01/22, 11:33	CONFIDENTIAL	Student	Assignment: Peer assessment forms for your grou...	Assignment	T
5	24/01/22, 11:33		Student	Assignment: Peer assessment forms for your grou...	Assignment	C
6	24/01/22, 11:33		Student	Assignment: Project 3rd draft	Assignment	T
7	24/01/22, 11:33		Student	Assignment: Project 3rd draft	Assignment	C
8	24/01/22, 11:33		Student	URL: Recorded online lecture (in case you want to r...	System	C
9	24/01/22, 11:32		Student	File: Slides	System	C
10	24/01/22, 11:32		Student	Lesson: Recorded Video (Self learning before onlin...	System	C
11	24/01/22, 11:32		Student	URL: MS Team link of Online Discussion Session 5	System	C
12	24/01/22, 11:32		Student	URL: Recorded Video (Self learning before online di.	System	C
13	24/01/22, 11:32		Student	Course: Cl3517 Research Methods	System	C
14	19/01/22, 15:19		Student	Course: CI3517 Research Methods	System	C
15	17/01/22, 14:06		Student	Course: CI3517 Research Methods	User report	G
16	17/01/22, 14:06		Student	Course: Cl3517 Research Methods	System	C
17	13/01/22, 11:38		Student	Forum: Announcements	Forum	C
18	13/01/22, 11:38		Student	Course: CI3517 Research Methods	System	C
19	4/01/22, 08:17		Student	Course: Cl3517 Research Methods	System	C
20	2/01/22, 18:50		Student	Assignment: Project 2nd draft	Assignment	T
21	2/01/22,18:50		Student	Assignment: Project 2nd draft	Assignment	C
22	1/01/22, 22:34		Student	Course: CI3517 Research Methods	System	C
23	11/12/21, 17:30		Student	Assignment: Peer assessment forms for your grou...	Assignment	T
24	11/12/21, 17:30		Student	Assignment: Peer assessment forms for your grou...	Assignment	C
25	11/12/21, 17:30		Student	Assignment: Peer assessment forms for your grou...	Assignment	A
26	11/12/21, 17:30		Student	Assignment: Peer assessment forms for your grou...	File submissions	S
27	11/12/21, 17:30		Student	Assignment: Peer assessment forms for your grou...	File submissions	A

Figure 3. Data labelling in the process mining tool

Table 1. Conversion from attribute to labelled data

Attribute	Label	Attribute	Label
Time	Timestamp	Event name	Timestamp
User full name	Resource	Description	Resource
User	Resource	Origin	Resource
Event context	Activity	IP address	Activity
Component	ID	Grade	ID

Once the data were completely labelled, they could be imported to automatically generate a visualisation map of the mined activities. This map is essential for discovering the flow of each activity from the logs that cannot be analysed directly. Figure 4 presents a simplified flowchart of the map. The first activity carried out by a student is to examine the module in the LMS after it is assigned by the administration. The next step is to attempt quizzes or assignments from the lecturer, followed by submitting files, viewing the status of a submission and, finally, verifying their grade for the module.

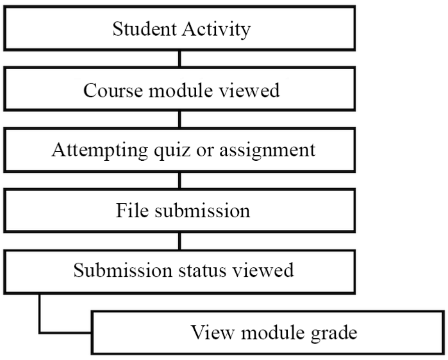


Figure 4. Student activity flowchart

Statistical information were produced based on the outcomes, as illustrated in Figure 5. Six random students were selected as examples, together with their associated events and activities. The students were selected with one associated with each grade (A+, A, B+, B, C+, C), in order to determine the different amounts of data. Figure 6 presents the mean and median case duration (in days) for the abovementioned activities.

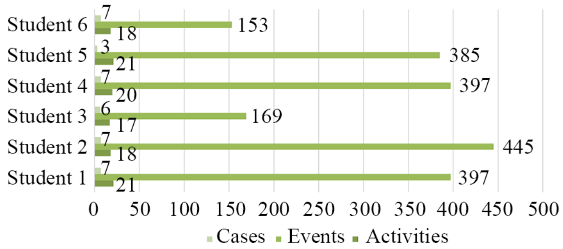


Figure 5. Student statistics

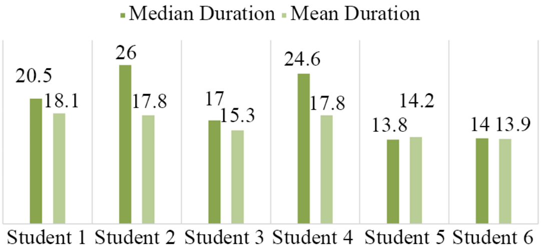


Figure 6. Mean and median case duration for each student's activity

k-Means clustering was conducted before proceeding to the final phase of data analysis, (i.e., classification), in order to discover groups that had not been labelled properly in the data. We expected to obtain three clusters on the basis of activities related to the user role: Admin, lecturer, or students.

Based on the outcome of the clustered instances, cluster 0 contained 67% of instances, cluster 1 contained 26%, and the remaining 7% was attributed to cluster 2. Table 2 summarises the instances.

Table 2. Clustering instances and percentages

Cluster	Instances	Percentage
0	17,967	67%
1	6,959	26%
2	1,782	7%

From these results, we focused on the first cluster, which was initially expected to be activities related to the students. The obtained clusters were used to support the classification of the grades of the students. The number of instances for the students in each cluster differed. Cluster 0 contained 14,829 instances, cluster 1 had 6,919 instances, and the remaining 1,444 instances were in cluster 2.

The three clusters identified in the *k*-Means analysis reflect distinct user behavior patterns. Cluster 0 has the highest number of instances, as this was the group of students that were intensely active in using the LMS, carrying out activities such as attempting the assignments, answering quizzes, submitting files, and viewing the submitted assignments. Clusters

1 and 2 were associated with various activities. Cluster 1 seemed to involve any materials created by the lecturer, whereas cluster 2 was related to the allocation of students to their courses at the beginning of LMS access by the administration.

This segmentation not only validates the effectiveness of the clustering configuration ($k = 3$), but also demonstrates that behaviorally distinct user groups can be automatically identified based on LMS logs. Such profiling is valuable for LMS administrators to understand the distribution of activity across user roles and to detect anomalies or under-engagement within expected groups. Moreover, this clustering lays the groundwork for tailored interventions. For example, students with activity patterns that diverge from Cluster 0 could be flagged for additional support or encouragement.

Figure 7 represents the clusters in distinct colours. Cluster 0 is shown in blue, cluster 1 is in red, and cluster 2 is in green. The admin activity-related cluster was quite homogeneous, with the colour being 99% blue. The cluster for the students was colourful, signifying that the students completed a range of activities from various resources. The classification algorithms utilized for this analysis were executed through Weka software [21].

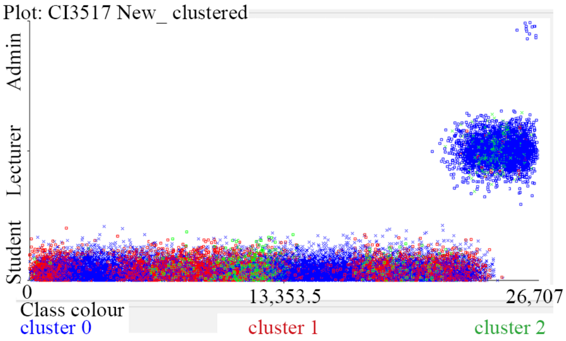


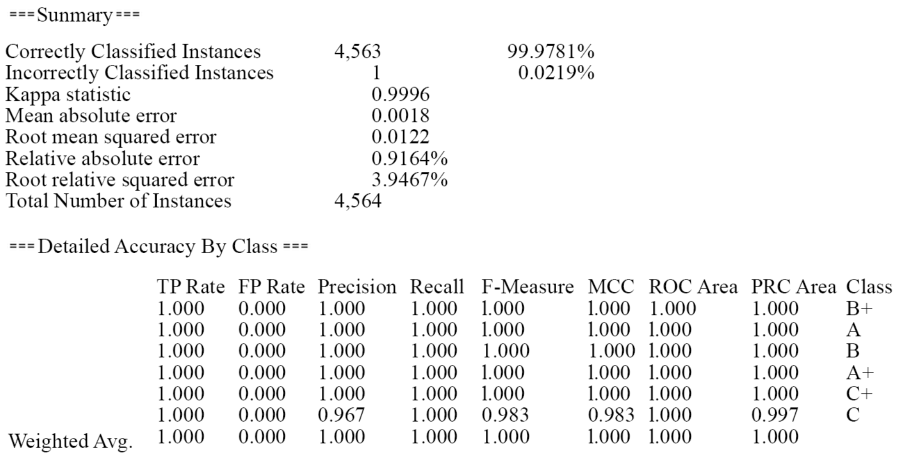
Figure 7. Representation of k-Means clustering result

For the classification, different algorithms were tested in order to determine the one with the best result. The classification algorithms chosen for this study were Naïve Bayes, IBK, and J48. The student's grade was appointed as the targeted class.

The Naïve Bayes algorithm is well-known to perform well with categorical input variables, in comparison to numerical input variables [22] and is based on a probabilistic model that involves strong independence assumptions. Based on the output shown in Figure 8, the result of the weighted average F-measure is almost 100%. Overall, 4,563 of the instances were correctly classified, with only 1 incorrectly classified instance.

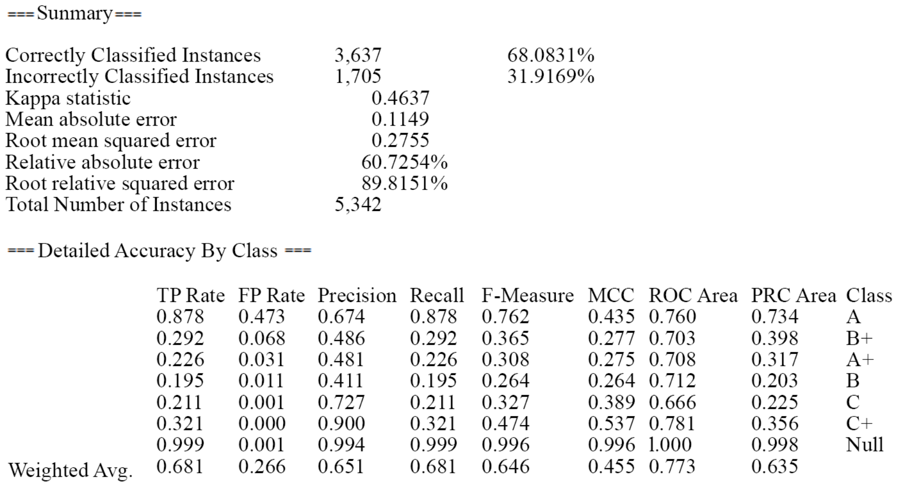
While the Naïve Bayes classifier achieved an exceptionally high weighted average F-measure of nearly 100%, this result may raise concerns about potential overfitting. Overfitting occurs when a model performs extremely well on the training data but fails to generalize to unseen data. In our case, the result may be influenced by the nature of the dataset and the possibility that some features strongly correlate with the target labels, making them easier to classify. However, to ensure that the model performance is not inflated due to overfitting, we acknowledge that future studies should apply additional validation techniques such as k-fold cross-validation or holdout methods. This would provide more rigorous insights into the model's generalization capabilities and stability across different data splits.

IBK is an algorithm that does not develop a model; instead, it produces a prediction for a test instance. This approach aims to measure the position of k instances in the data and utilises the selected instances to make a prediction [23]. The result of this algorithm, as illustrated in Figure 9, was 64% of its weight average with 68% correctly classified instances.



```
===Sunmary===

Correctly Classified Instances        4,563              99.9781%
Incorrectly Classified Instances          1               0.0219%
Kappa statistic                           0.9996
Mean absolute error                       0.0018
Root mean squared error                   0.0122
Relative absolute error                   0.9164%
Root relative squared error               3.9467%
Total Number of Instances             4,564

===Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     B+
                1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     A
                1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     B
                1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     A+
                1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     C+
                1.000    0.000    0.967      1.000   0.983      0.983  1.000     0.997     C
Weighted Avg.   1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000
```

Figure 8. Naïve Bayes algorithm results

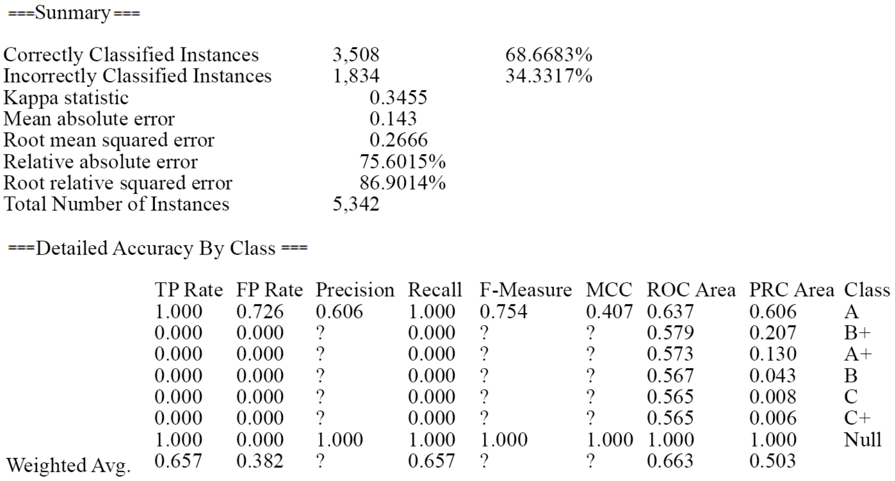


```
===Sunmary===

Correctly Classified Instances        3,637              68.0831%
Incorrectly Classified Instances      1,705              31.9169%
Kappa statistic                           0.4637
Mean absolute error                       0.1149
Root mean squared error                   0.2755
Relative absolute error                  60.7254%
Root relative squared error              89.8151%
Total Number of Instances             5,342

===Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                0.878    0.473    0.674      0.878   0.762      0.435  0.760     0.734     A
                0.292    0.068    0.486      0.292   0.365      0.277  0.703     0.398     B+
                0.226    0.031    0.481      0.226   0.308      0.275  0.708     0.317     A+
                0.195    0.011    0.411      0.195   0.264      0.264  0.712     0.203     B
                0.211    0.001    0.727      0.211   0.327      0.389  0.666     0.225     C
                0.321    0.000    0.900      0.321   0.474      0.537  0.781     0.356     C+
                0.999    0.001    0.994      0.999   0.996      0.996  1.000     0.998     Null
Weighted Avg.   0.681    0.266    0.651      0.681   0.646      0.455  0.773     0.635
```

Figure 9. IBK algorithm results

One of the most widely used algorithms for the evaluation of categorical and continuous data is J48. This algorithm is defined using decision trees and, thus, produces a visualisation that includes nodes associated with attributes in a branching strategy. The outcome of the J48 classification method is shown in Figure 10, from which it can be seen that 65% of the instances were correctly classified. A weighted average is a technique used to combine multiple evaluation metrics by assigning different weights to each metric. It allows us to prioritise certain metrics over others based on their importance in the evaluation process. However, the F-measures for the A+, A, B+, B, C and C+ classes could not be fetched properly, resulting in an error when producing the percentage of the weighted average. The F-measure could only be read for the classes A and Null.

The classification results further support the role of LMS behavior in predicting student performance. The Naïve Bayes classifier outperformed both IBK and J48, achieving an F-measure close to 100%. Although Naïve Bayes assumes feature independence-which may not fully hold in behavioral data-it remains a strong baseline due to its efficiency and robustness to noise. In contrast, the IBK algorithm, which is sensitive to the curse of dimensionality, yielded a lower F-measure (64%). J48 also struggled, likely due to sparse categorical combinations and an imbalanced label distribution. These results reinforce our choice to prioritize Naïve Bayes, while also highlighting the limitations of instance-based and tree-based classifiers in this context. From a practical standpoint, the classification model could be deployed in an early-warning system to predict academic risk levels based on LMS usage patterns.

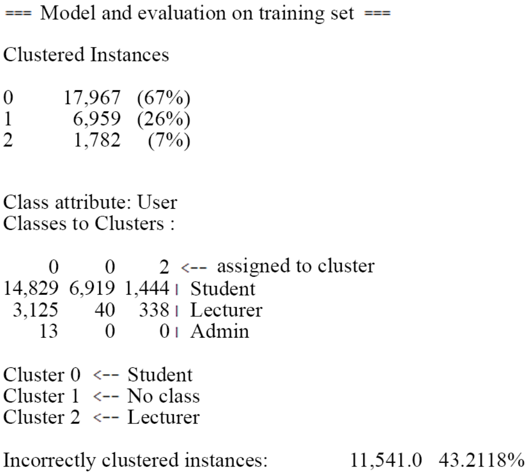


```
===Sunmary===

Correctly Classified Instances        3,508             68.6683%
Incorrectly Classified Instances      1,834             34.3317%
Kappa statistic                          0.3455
Mean absolute error                      0.143
Root mean squared error                  0.2666
Relative absolute error                75.6015%
Root relative squared error            86.9014%
Total Number of Instances             5,342

===Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               1.000    0.726    0.606      1.000   0.754      0.407  0.637     0.606     A
               0.000    0.000    ?          0.000   ?          ?      0.579     0.207     B+
               0.000    0.000    ?          0.000   ?          ?      0.573     0.130     A+
               0.000    0.000    ?          0.000   ?          ?      0.567     0.043     B
               0.000    0.000    ?          0.000   ?          ?      0.565     0.008     C
               0.000    0.000    ?          0.000   ?          ?      0.565     0.006     C+
               1.000    0.000    1.000      1.000   1.000      1.000  1.000     1.000     Null
Weighted Avg.  0.657    0.382    ?          0.657   ?          ?      0.663     0.503
```

Figure 10. J48 algorithm results

7. Results and discussions

According to the clustering results shown in Figure 11, the clusters were not misleading and the number of clusters was optimal. There were three clusters, and the best cluster was cluster 0, as it provides more activities for the students. During the evaluation, whether the clusters matched the data set or not was identified by monitoring the number of incorrectly clustered instances.



```
=== Model and evaluation on training set ===

Clustered Instances

0       17,967   (67%)
1        6,959   (26%)
2        1,782    (7%)


Class attribute: User
Classes to Clusters :

      0       0       2  <-- assigned to cluster
 14,829  6,919  1,444 | Student
  3,125     40    338 | Lecturer
     13      0      0 | Admin

Cluster 0  <-- Student
Cluster 1  <-- No class
Cluster 2  <-- Lecturer

Incorrectly clustered instances:          11,541.0   43.2118%
```

Figure 11. Result of the clustering algorithm

Clustering techniques are used to group any labelled data that are too sophisticated to cluster manually. They are used to acquire cluster labels from each test data set before proceeding to the next phase. The performance of the data is strengthened by clustering, which has a strong impact on providing a highly accurate classification outcome. Clustering improved the precision, recall, and accuracy values of the classification approach used in this study.

The clustering results indicated that there were 11,541 wrongly clustered instances, which is approximately 43%. This led to a sparse number of correctly clustered instances. As listed in Table 3, the best result, in terms of the F-measure, was for Naïve Bayes. J48 was only able to make predictions on numeric attributes to produce its pattern. These results were calculated from the weighted average percentage under the F-measure.

Table 3. Summary of the classification results

Algorithm	F-measure
Naïve Bayes	100%
IBK	64%
J48	Null

Naïve Bayes classification was found to provide the best classification result in the context of this study. The most likely reason for this is that it enables us to model quantitative and qualitative information about the behaviour of students [24]. Naïve Bayes is an ideal classification technique, as the datasets used were huge and it can avoid over-fitting. Some other benefits of Naïve Bayes are that it is easy to build as the structure is given priority, there are no learning procedures, and it is an efficient classification process.

In addition, the activities conducted by the admins were comparable to the activities conducted by the students and lecturers during clustering, as they consisted of similar activities. Another issue was that sometimes the data set was not always classified appropriately, according to the grades and the activity resources.

The findings of this study imply that the activities in the LMS may significantly affect the grades of the students. Table 4 presents summary data for the six randomly selected students indicating that the selected students engaged in a varying number of activities and events in their learning phase in LMS. The data indicate that the students who were more often engaged in the LMS could be predicted to achieve a higher grade at the end of their studies. Although no students failed the class, there were students with lower grades (i.e., C+ and C) who were still observed to be active in the LMS.

Table 4. Summary of each student's activities

Student	Activities	Events	Cases	Mean duration	Median duration	Grade
1	21	397	7	20.5 days	18.1 days	B
2	18	445	7	26 days	17.8 days	A+
3	17	169	6	17 days	15.3 days	C
4	20	397	7	24.6 days	17.8 days	A
5	21	385	3	13.8 days	14.2 days	B+
6	18	153	7	14 days	13.9 days	C+

By reviewing the log data, we observed that most of the students were highly active and used the LMS occasionally. Furthermore, our objective was to identify the performance of students and the simplest approach that may be followed to encourage students to achieve better results in their studies. Hence, from the results of the study, we can consider the provided approach to have successfully achieved this goal.

The findings of this study offer practical value for both instructors and LMS administrators. For instructors, understanding how student activities in the LMS relate to academic performance allows them to identify students who may be at risk early in the semester based on their interaction patterns. Instructors can use this information to implement timely interventions, such as providing additional support or encouraging greater LMS engagement for students who are less active. For LMS administrators, clustering and classification results can guide the development of intelligent notification systems or dashboards that flag students with potentially lower engagement. These features could be integrated into the LMS to assist teaching staff in real-time decision-making. Additionally, tracking engagement trends over time can help inform instructional design, course sequencing, and the allocation of learning resources, eventually supporting improved learning outcomes.

Overall, the combined clustering and classification results not only offer a diagnostic lens into student engagement patterns but also present opportunities to develop predictive analytics tools that enhance student retention and success through data-informed teaching interventions.

8. Conclusion

The goal of this study was to identify the behaviours of students based on their activities recorded in an online Learning Management System (LMS). The *k*-Means clustering and Naïve Bayes classification techniques were utilised to examine the student behaviours. These behaviours were identified using process mining, with the aim of obtaining insights into the academic performance of the students. The purpose of carrying out the clustering and classification steps was to ensure better performance. We proposed the use of data from an online LMS, as it contains a huge data set reflecting user behaviours. We extracted certain activity and process information from the system, and then analysed the user behaviours in the system to obtain insights such as the correlation between the learning process and academic performance of students. By discovering processes, checking performance, and analysing engagement from the log data in a learning management system, user behaviours can be identified, and it can be determined whether they affect academic performance or not. In conclusion, based on the outcomes of the analysis, the engagement of students in the learning management system was found to affect their final grades; however, it should be noted that the students who obtained lower grades were also active in the online LMS. Overall, our main objective was to identify both the performance of students and the simplest approach which can be followed to encourage students to achieve better results in their studies.

While the study provides valuable insights, several limitations should be considered to balance the conclusions. First, the data set is limited to 79 students from a single module at Universiti Teknologi Brunei, which may not fully represent diverse learning contexts or larger student populations. Second, the analysis relies solely on LMS log data, potentially overlooking external factors such as offline learning activities, personal circumstances, or instructor interventions that could influence academic performance. Third, the institutional-specific context may limit the generalizability of the findings to other educational settings with different LMS implementations or cultural factors. Future research should aim to expand the sample size, incorporate multi-source data (e.g., surveys or classroom observations), and conduct cross-institutional studies to validate and extend these results.

Ethics statement

This research was approved by the Universiti Teknologi Brunei (UTB) Research Ethics Committee under the School of Computing and Informatics. All procedures involving data collection and analysis complied with UTB's Research Ethics Policy and Procedures (Version 3, April 2021). The study made use of historical LMS log data collected during routine academic operations, and no direct interaction with students occurred. All user data were fully anonymized before analysis to remove any personally identifiable information such as names, student IDs, and IP addresses. Access to the dataset was restricted to the research team and conducted under strict confidentiality agreements. Informed consent was deemed not applicable as no personal identifiers were retained and the study involved secondary use of de-identified data in accordance with institutional policy.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Zulfadhilah M, Riadi I, Prayudi Y. Log classification using *k*-means clustering for identify internet user behaviors. *International Journal of Computer Applications*. 2016; 154(3): 34-39.

[2] UmaMaheswari S, Srivatsa S. Algorithm for tracing visitors' on-line behaviors for effective web usage mining. *International Journal of Computer Applications*. 2014; 87(3): 22-28.

[3] Veluvali P, Surisetti J. Learning management system for greater learner engagement in higher education-A review. *Higher Education for the Future*. 2022; 9(1): 107-121.

[4] Shayan P, van Zaanen M. Predicting student performance from their behavior in learning management systems. *International Journal of Information and Education Technology*. 2019; 9(5): 337-341.

[5] Vieira A. Predicting online user behaviour using deep learning algorithms. *arXiv:151106247*. 2015. Available from: https://doi.org/10.48550/arXiv.1511.06247.

[6] Rahman A, Alrashed SA, Abraham A. User behaviour classification and prediction using fuzzy rule based system and linear regression. *Journal of Information Assurance & Security*. 2017; 11: 86-93.

[7] Parveen Z, Alphones A, Naz S. Extending the student's performance via *k*-means and blended learning. *International Journal of Engineering and Applied Computer Science*. 2017; 2(4): 133-136.

[8] Oyelade O. Application of *k*-means clustering algorithm for prediction of students' academic performance. *International Journal of Computer Science & Information Security*. 2010; 7(1): 292-295.

[9] Bradley VM. Learning management system (LMS) use with online instruction. *International Journal of Technology in Education*. 2021; 4(1): 68-92.

[10] Johansson A. *Clustering User-Behavior in a Collaborative Online Social Network: A Case Study on Quantitative User-Behavior Classification*. Master's thesis. School of Computer Science and Communication, KTH Royal Institute of Technology; 2016.

[11] Nelson K. Using *k*-means clustering to model students' LMS participation in traditional courses. *Issues in Information Systems*. 2015; 16(4): 102-110.

[12] Hidayat N, Wardoyo R, Sn A, Surjono HD. Enhanced performance of the automatic learning style detection model using a combination of modified *k*-means algorithm and naive bayesian. *International Journal of Advanced Computer Science and Applications*. 2020; 11(3): 638-648.

[13] Jayathilake D. Towards structured log analysis. In: *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*. Bangkok, Thailand: IEEE; 2012. p.259-264.

[14] Alshammari MH, Alshammari SH. Examining students' readiness toward using learning management system at University of Ha'il: A structural equation modelling approach. *Sustainability*. 2022; 14(22): 15221.

[15] Papadakis S, Kiv AE, Kravtsov HM, Osadchyi VV, Marienko MV, Pinchuk OP, et al. Unlocking the power of synergy: The joint force of cloud technologies and augmented reality in education. In: *Joint Proceedings of the 10th Workshop on Cloud Technologies in Education, and 5th International Workshop on Augmented Reality in Education (CTE+AREdu 2022)*. Aachen, Germany: CEUR-WS.org; 2023. p.1-23.

[16] Osadchyi V, Papadakis S, Kiv A, Kravtsov H, Marienko M, Pinchuk O, et al. Revolutionizing education: using computer simulation and cloud-based smart technology to facilitate successful open learning. In: *Joint Proceedings of the 10th Illia O. Teplytskyi Workshop on Computer Simulation in Education, and Workshop on Cloud-based Smart Technologies for Open Education (CoSinEi and CSTOE 2022) Co-Located with ACNS Conference on Cloud and Immersive Technologies*. Aachen, Germany: CITEd 2022, Ukraine; 2023. p.1-18.

[17] Bramantoro A, Alzahrani A, Bahaddad A, Alfakeeh AS. Cloud-based learning service platform for multilingual smart class. *International Journal of Advanced and Applied Sciences*. 2020; 7(7): 83-91.

[18] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0: Step-by-step data mining guide. *SPSS Inc*. 2000; 9(13): 1-73.

[19] Murakami Y, Tanaka M, Bramantoro A, Zettsu K. Data-centered service composition for information analysis. In: *2012 IEEE International Conference on Services Computing (SCC)*. USA: IEEE Computer Society; 2012. p.602-608.

[20] Rozinat A, Günther C. *Disco User Guide-Process Mining for Professionals*. Eindhoven, The Netherlands: Fluxicon BV; 2012. Available from: https://fluxicon.com/disco/files/Disco-User-Guide.pdf [Accessed 5th June 2025].

[21] Frank E, Hall MA, Witten IH. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. 4th ed. Burlington, MA, USA: Morgan Kaufmann (Elsevier); 2016. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf [Accessed 5th June 2025].

[22] Wickramasinghe I, Kalutarage H. Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing*. 2021; 25(3): 2277-2293.

[23] Brownlee J. *Machine Learning Mastery with Weka*. Australia: Machine Learning Mastery; 2016. Available from: https://machinelearningmastery.com/machine-learning-mastery-weka/ [Accessed 5th June 2025].

[24] As'Sidiq A, Mandala R. Implementation of *k*-means algorithm for information technology freshman class division. *IT for Society*. 2020; 4(1): 1-6.