UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Multidimensional Context Clustering to Analyse Student Engagement in Online Learning Environment

Chong Ke Ting[1*], Noraini Ibrahim[2], Sharin Hazlin Huspi[3], Wan Mohd Nasir Wan Kadir[4]
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
Email: keting-1997@graduate.my[1], noraini_ib@utm.my[2], sharin@utm.my[3], wnasir@utm.my[4]

Submitted: 10/9/2024. Revised edition: 29/10/2024. Accepted: 29/10/2024. Published online: 25/11/2024
DOI: https://doi.org/10.11113/ijic.v14n2.487

***Abstract*—Educational data mining is the application of data mining technology in an educational environment to indicate and resolve various types of issues faced in education. COVID-19 pandemic in 2020 accelerated the shift to emergency remote learning and online learning, which has continued to grow due to its flexibility and cost-effectiveness. However, several challenges exist when adopting the online learning, and which include the access to technology, technology attitude, psychological questions, teacher contact, and quality of assessment. Thus, it becomes important to focus on student engagement as a determinant of success during online learning. Student engagement is a complex construct which is comprised of four aspects which are behavioural, cognitive, emotional and social. In this study, the K-Means clustering approach is chosen to categorise students into clusters according to their active participation in the online learning process. The experiment used produces a silhouette coefficient of 0.71 clustering the datasets into three clusters. Cluster 0 are the disengaged learners who were observed to be least active across all the dimensions, while cluster 1 is composed of passive learners. The cluster 2 comprises the most engaged students with the high level of time management. These results provide information regarding the distinct engagement profiles that may be helpful when the lecturers attempt at student's interventions.**

***Keywords*—Educational data mining, student engagement, clustering, online learning, k-means, data preprocessing, normalisation**

I. Introduction

Educational Data Mining (EDM) is the implementation of data mining technology in educational environments that integrates knowledge of the multiple disciplines, which are pedagogy, computer science, psychology, statistics, and machine learning (ML), to identify the hidden patterns in the huge educational data [1]. EDM had been implemented to indicate and resolve various types of issues that were faced in education, which included assisting the managers in making decisions, helping teachers to improve the course, improving the students' learning efficiency, and developing more effective online learning tools [1].

Over time, the rapid technological development in recent decades has changed both human life and the learning process [2-4]. Online learning has become a popular learning approach where the students are allowed to learn using mobile platforms or devices, which include smartphones [2]. Education has changed and evolved from teacher-based instruction to modern technology-based learning to encourage active learning [3]. The Ministry of Education Malaysia aims to promote online learning for improving teaching and learning quality while boosting the cost effectiveness for Malaysian higher education [5].

The development of technology went even faster and significantly affected society during the COVID-19 pandemic [3]. It was highlighted that there is a huge increase of users for digital devices and the internet to emphasise the significance of the modern technologies in society to overcome the challenges that faced during the COVID-19 pandemic [3]. During the COVID-19 pandemic, higher education institutions are forced to convert from face-to-face (F2F) and blended learning to completely online delivery to continue the teaching and learning (T&L), which raises a hidden challenge for both instructors and students [6].

Student engagement is getting more important, and it becomes the essential element to indicate the academic success of the students during the online learning as compared to F2F learning [7]. Furthermore, it needs to be updated from time to time to assist the student in adjusting their engagement level to achieve the desired learning outcomes [8]. Therefore, it is needed to keep track of students' progress throughout the learning process to give continuous educational support [9]. Student engagement is a multidimensional context that consists of behavioural, cognitive, and emotional dimensions.

The clustering technique is one of the popular techniques used in EDM to indicate the students learning strategies or engagement [10]. It is a type of unsupervised learning technique that is used to find and group the objects based on their similar characteristics to deduce the hidden patterns and structures that consist in the large dataset [11, 12]. Among all the machine learning clustering techniques, the K-means algorithm is the most famous method that is used in EDM.

In this research, K-means clustering is implemented to cluster the student engagement into different levels. Other than that, the Spearman Correlation Coefficient is implemented to identify and eliminate those highly correlated attributes from the dataset before carrying out the K-Means clustering. Besides, this research used the Elbow method, starting from k = 2 to k = 10, to identify the most suitable number of clusters that give the best result. The silhouette coefficient is utilised to evaluate the performance of the clustering result.

II. Literature Review

EDM is the process of converting the raw data that is collected from different educational systems in higher education institutions into useful information that can have a potential impact on the practice and research in education [11, 13]. It is focused on developing effective techniques to mine the educational data to understand the student's learning behaviour and environment setting [11] and identify the students who are facing learning difficulties at the earlier stage [14]. Classification, clustering, relationship mining, and pattern discovery are the most popular methods used in EDM [15]. Besides, EDM is also focused on identifying the factors that cause student failure [15].

The Ministry of Education Malaysia has made online learning a crucial component of HE and lifelong learning that focus on student-centred learning and outcome-based education, which aim to develop students' knowledge through active student engagement and collaborative learning [5]. The implementation of online learning can encourage students to participate actively, build interaction, and increase performance and engagement during the learning process. [3]. With the exposure of the COVID-19 pandemic in 2020, almost all the areas of activities globally are affected, especially for the teaching and learning process. Therefore, Emergency Remote Learning (ERL) was introduced to continue the F2F and blended learning to completely online delivery [6]. Learning Management System (LMS) are used for all the access of learning resources, submission of assignments and exams, and communication with instructors and peers that are carried out online during the pandemic period [16]. The implementation of ERL has made online learning more popular even after the COVID-19 pandemic. There are few challenges that are faced by both the teachers and students during ERL as well as online learning, such as technology readiness [17], psychological issues [18, 19], teacher support [20], and assessment quality [21].

Even though modern technologies are good at providing support for online delivery, the active participation of students with the platform remains a challenge for successful online learning [22]. Therefore, higher education is increasing their effort to measure student engagement and participation level in an online learning environment to indicate students' success [6]. However, the evaluation and prediction of student engagement level in the online learning environment is still a challenge for education research [8]. Student engagement is not a monolithic but multidimensional construct, which includes behavioural, cognitive, and emotional [23]. Behavioural engagement is the effort of the student to complete the given task and follow the rules, while cognitive engagement is the student's effort and initiative to learn new knowledge in depth. Emotional engagement is the student's satisfaction with learning behaviour, participation in learning activities, and sense of self-worth among the peer interaction.

Student engagement does not have clear consensus on how to identify the level of student engagement; labelling methods need to be developed to label the data [16]. Therefore, clustering plays an important role in EDM research for identifying learning strategies or student engagement [10]. The K-Means clustering algorithm is the most famous clustering machine learning approach that was implemented to cluster educational data. K-Means clustering is the process of partitioning an N-dimensional population into *k* sets based on the characteristics of the sample [24]. It is used to divide the total numbers of observations into *k* clusters, where each observation belongs to the cluster with the nearest mean of the cluster centres [1].

In the research of Tamba, *et al.* [25], k-means clustering is implemented to arrange the student into groups based on their engagement in LMS. This research only used four attributes which measuring students' total contribution in forum, average page access, average activity per session, average time spent of LMS from three different courses [25]. The student engagement level is divided into active, quite active and less active. This result only achieves with 0.54 silhouette coefficient. This might be due to the limited attribute implemented to measure student the student engagement in this research. Furthermore, the data preprocessing is not carried out after transforming data into the indicator of student engagement, while the standardization and distribution of the data is playing important role and impact of the clustering result.

Benabbes, *et al.* [26] implemented four different clustering approaches which include K-Means, agglomerative, Birch, and DBSCAN clustering algorithm. Furthermore, elbow method is implemented in this research to get the optimal number of clusters. In this research, student engagement is measured in four different dimensions which are behavioural, cognitive, emotional and social engagement. In order to improve performance and generalization of clustering, Principal Component Analysis (PCA) is implemented to reduce the number of attributes [26]. PCA is sensitive to noise in the data, which can lead to misleading results [27]. However, students exhibit varied engagement pattern, influenced by personal circumstances, learning styles, and external factors such as the COVID-19 pandemic, which led to abrupt shifts in learning modalities [28]. Therefore, in an LMS content, the noise from can distort the analysis, potentially leading to incorrect interpretation of student interactions.

There are few challenges that faced are in the clustering approach, which include the curse of dimensionality that the clustering approach can perform well with low-dimensional datasets but faces difficulties to deal with high-dimensional datasets [29]. Furthermore, the presence of highly correlated attributes can cause redundancy issues that consequently affect the effectiveness of clustering by skewing the result and making the distinct pattern of the dataset hard to identify [30]. The presence of highly correlated attributes will increase the computational cost and increase the risk of overfitting, which consequently reduces the clustering performance and the model generalisation [29]. Other than that, the high correlation attributes can lead to biased clustering results, where they may dominate the similarity measures, causing clusters to form based on redundant information rather than meaningful distinction [31].

Spearman correlation coefficients is more favoured for the feature reduction in student interaction dataset due to its less sensitivity to outliers as compared to PCA, which can skew the results significantly [32]. Furthermore, Spearman correlation coefficients can capture both linear and non-linear associations [33]. This flexibility allows for a more accurate representation of complex interactions within student interaction data. Moreover, Spearman correlation approaches evaluate the strength and direction of associations between attributes based on their ranks, providing a clearer insight into relationships [34]. While PCA is a powerful tool in linear data reduction, its limitation in handling non-linear relationships and sensitivity to outliers make Spearman Correlation coefficient a more suitable choice for handling with student interactions dataset.

III. Research Methodology

This research is mainly made up of three phases, which are data collection, data preprocessing and data clustering. The research framework is as shown in Fig. 1.

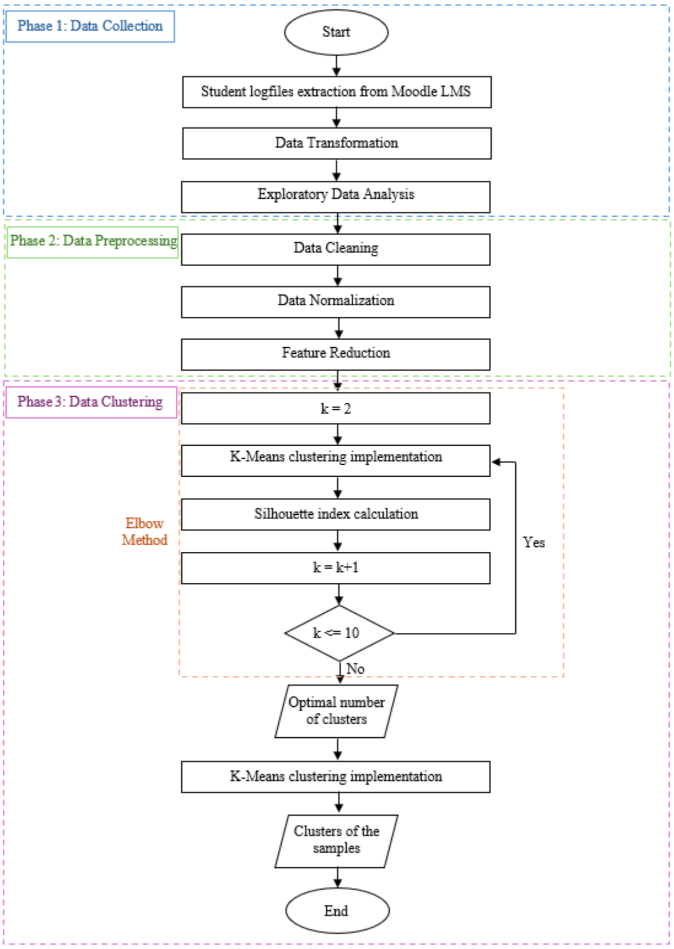
Phase 1: Data Collection
Start
Student logfiles extraction from Moodle LMS
Data Transformation
Exploratory Data Analysis
Phase 2: Data Preprocessing
Data Cleaning
Data Normalization
Feature Reduction
Phase 3: Data Clustering
k = 2
K-Means clustering implementation
Silhouette index calculation
k = k+1
k <= 10
Yes
No
Elbow Method
Optimal number of clusters
K-Means clustering implementation
Clusters of the samples
End

Fig. 1. Research Framework

Fig. 1 shows that the overall research framework of this paper which is mainly made up three phases including data collection, data pre-processing and the data clustering phase. The details for each of the research phases are illustrated and discussed in the following subsection.

A. Data Collection

The dataset that was utilised in this research is extracted from the Faculty of Computing at Universiti Teknologi Malaysia Moodle LMS. The data are collected from the Moodle LMS logfiles from 2018 to 2021. A total of eleven general courses that need to be taken by all the students in this faculty are collected. The courses include Technology & Information System, Discrete Structures, Programming Techniques I, Digital Logic, Database, System Analysis & Design, Data Structure & Algorithm, Network Communication, Human Computer Interaction, Artificial Intelligence, and Application Development. A total of 297 logfiles from different sections are being collected from the Moodle. The dataset is collected since the beginning of the semester until the end of the semester. The raw logfiles collected from the LMS are made up of nine attributes that trace the students' footprint during online learning. The data used in this study is not allowed to be shared publicly due to confidentiality agreements established with UTM Digital Centre, ensuring the privacy and protection of sensitive information. The details of attributes used in this research are illustrated in Table I.

TABLE I. Details of Attributes Used

Attributes	Data Type	Description
Time	Datetime	The date and time when the event occurred.
User full name	String	The full name of the user who triggered the event. 1. LMS Admin – 'admin elearning', 'Pengurusan IT UTMLead' 2. Lecturer – Lecturer name followed by staff ID 3. Student – Student name followed by matric card number
Affected user	String	The user who is impacted by the event or action.
Event context	String	The specific part or context within Moodle where the event took place. For example, the name of a course or activity
Component	String	The part of Moodle responsible for the event. This could be a specific module, plugin, or tool, such as "Assignments", "Quizzes", "Forum", "File" or "Folder"
Event name	String	A label or title for the event that took place. This gives a brief description of the type of action, such as "Course viewed," "File uploaded," or "Quiz attempted."
Description	String	A more detailed explanation of what exactly happened during the event.
Origin	String	The source of the event, indicating where or how the event was initiated, such as "Web" (browser), "CLI" (command line), or "ws" (web service).
IP Address	String	The IP address of the device used by the user when the event occurred. This can be useful for tracking access locations or identifying potential unauthorized access.

Based on the data collected in this research, the usage of Moodle LMS for online learning can be categorised into four different levels. The level of usage with the amount of each category and description is illustrated in Table II.

TABLE II. Level of Moodle LMS Usage

Level	No. of classes	Description
0	10	Does not use Moodle LMS
1	21	Use for uploading resources only
2	148	Use for uploading resources, assignment and quiz submission
3	118	Use for uploading resources, assignment and quiz submission, and forum announcement

Based on Table II, there are 10 classes that do not use Moodle LMS for any activity, and 21 of the classes use Moodle LMS only for uploading the resources to the student. These 31 courses are excluded from this research study, since it does not give much information for the purpose of this research. Therefore, in this study, 266 classes are implemented for further analysis.

Even though the raw logfiles that were collected from Moodle LMS are rich in detailed data, there is a lack of the structure and context needed to provide meaningful insights on student engagement during online learning. Therefore, further data cleaning and transformation are required to extract the knowledge in the dataset. Since this research is only focused on the student interaction with LMS during online learning, only the Time, User full name, Event context, Component, Event name are being studied in this research. The columns are then utilized and transformed into dataset that record the interaction of each student with the LMS. The details of the transformed data are shown in the Table III.

Table III shows the 16 attributes that were extracted from the Moodle logfiles to measure the four dimensions of student engagement, which include behavioural, cognitive, emotional, and social engagement. After data transformation, a total of 9406 samples are extracted from the logfiles collected. Exploratory Data Analysis (EDA) is carried out to gain an initial understanding of the data key characteristics, which include mean, standard deviation, minimum value, maximum value, distribution, and correlation. The details of the key characteristics are recorded in Table II. Furthermore, the distribution of each attribute is shown in Fig. 2.

TABLE III. DETAILS OF ATTRIBUTES USED

Attributes	**Dimension**	**Data Type**	**Mean**	**STD**	**Min.**	**Max.**
No of login	Behavioural	Integer	153.6	106.6	1	1012
No of quiz completed	Behavioural	Integer	3.2	4.7	0	38
No of individual assignment submitted	Behavioural	Integer	5.5	5.4	0	44
No of group assignment submitted	Behavioural	Integer	1.8	2.9	0	24
Total no of assignment submitted	Behavioural	Integer	7.3	5.8	0	49
No of access to course material	Cognitive	Integer	81.3	61.0	0	635
Total time spend individual assignment	Cognitive	Float	581.2	1288.4	0	16192.8
Average time spend individual assignment	Cognitive	Float	109.5	193.4	0	2277.9
Total time spend group assignment	Cognitive	Float	207.1	565.3	0	13093.8
Average time spend group assignment	Cognitive	Float	71.4	168.7	0	2769.0
No of forum viewed	Social	Integer	18.8	67.1	0	1115
No of individual assignment ontime	Emotional	Integer	4.8	4.9	0	36
No of individual assignment late	Emotional	Integer	0.7	1.4	0	17
No of group assignment ontime	Emotional	Integer	1.6	2.7	0	23
No of group assignment late	Emotional	Integer	0.2	0.6	0	11
Total no of clicks	Behaviour	Integer	267.2	189.8	1	2337

According to Fig. 2, each of the attributes is mostly right skewed, showing that most of the values are distributed around the lower end, with a long tail extending towards the higher end. This skewness suggests that only a small number of samples are having significantly larger values compared to the rest of the dataset. The distribution of attributes that relate to the forum and group assignment is more right-skewed. This is because the forum component is not popularly utilised during online learning for the fifclass discussion, causing that the higher value is very limited. On the other hand, not all the classes are having the group assignment assessment. Moreover, only one representative is required to submit the group assignment.

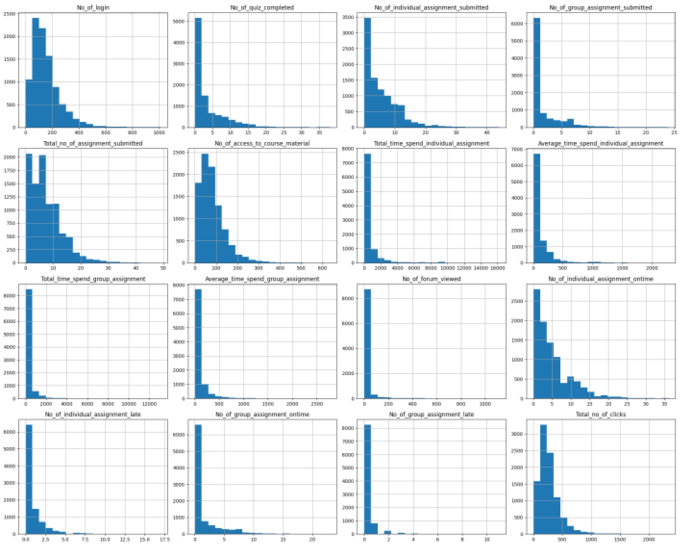

Fig. 2. Distribution of student interactions attribute

Furthermore, the Spearman Correlation Coefficient is carried out in the EDA phase of this research to study the correlation between attributes. The correlation between attributes is shown in the heatmap as shown in Fig. 3.

Fig. 3 shows that there few of the attributes are highly correlated with others attribute. Among all the attributes No_of_login and Total_no_of_clicks, Total_time_spend_individual_assignment and Average_time_spend_individual_assignment, Total_time_spend_grou_assignment and Average_time_spend_group_assignment, No_of_individual_assignment_submitted and No_of_individual_submitted_ontime, and No_of_group_assignment_submitted and No_of_group_submitted_ontime are highly correlated with each other with more than 0.90 correlated.

Then, the transformed data is then implemented for clustering to identify the hidden pattern of student engagement inside the dataset. Before clustering the data, the dataset is undergone data preprocessing to clean and improve the quality of the dataset.

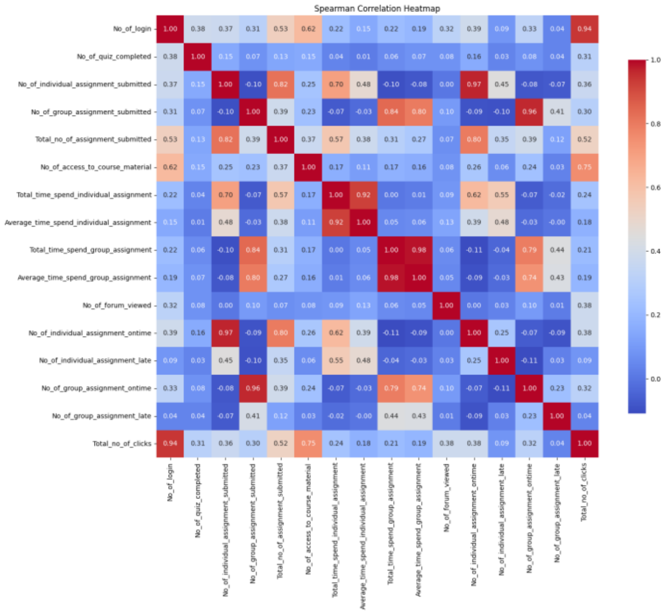

Fig. 3. Heatmap of correlation between attributes

B. Data Pre-processing

Data preprocessing activity is started with data cleaning to identify and remove invalid data, followed by data normalisation and feature reduction of the data as shown in Fig. 1. The irrelevant data where the student that does not leave any interaction footprints on Moodle LMS is eliminated. At this step, 320 samples are being eliminated. Besides, the 29 duplicates are also eliminated from the dataset. After the irrelevant sample elimination, 9057 of the samples are kept in this study for clustering.

Besides, data normalisation is carried out to normalise the features to obtain the normal distribution of the data. In this research, Robust Scaler is implemented to normalise the data because it is suitable for non-normally distributed data, and it can handle datasets that contain outliers. Robust Scaler is a process of implementing a statistical interquartile method to identify and remove the exception. It eliminates the median and scales the data based on the interquartile range, which is between the 25th quantile and 75th quantile. The centring and scaling processes occur independently on each attribute. The formula of Robust Scaler is as follows:

$$RS(X_i) = \frac{X_i - median(X)}{IQR_{1,3}(X)} \quad (3)$$

Where $IQR_{1,3}$ (X) represents the interquartile range between the 25th quantile and 75th quantile of X.

After data normalisation, feature reduction is carried out to eliminate the highly correlated attribute to reduce the complexity of the clustering process. In this step, the attributes that are more than 0.8 correlated by Spearman Correlation will be highlighted, and the attributes that have a higher sum of correlation with all other attributes will be eliminated. The attributes 'Total_no_of_assignment_submitted', 'No_of_individual_assignment_ontime', 'No_of_group_assignment_ontime', 'Total_time_spend_individual_assignment', 'Total_time_spend_group_assignment', 'Total_no_of_assignment_submitted', 'Total_no_of_clicks' are eliminated to reduce the number of attributes and complexity of clustering process.

C. Data Clustering

This research considers the K-Means clustering algorithm as the clustering model to group and cluster the data that was collected from Moodle logfiles, which represent the engagement of the student with online learning. Tomasevic, *et al.* [12] mentioned that the advantage of the clustering method is the ability to deduce the hidden patterns and structures from the dataset even with limited or even none of the background knowledge. The K-Means clustering algorithm from scikit-learn, which is based on Python, is utilised in this research. The parameter of K-Means clustering that implemented in this research is illustrated in Table IV.

TABLE IV. K-MEANS CLUSTERING PARAMETER

Parameter	**Value Setting**
n_clusters	2 to 10
init	'k-means++'
n_init	'auto'
max_iter	300
tol	1e-4
verbose	0
random_state	None
copy_x	True
algorithm	lloyd

Besides, the elbow method is implemented in this research to identify the optimal number of clusters, and the silhouette coefficient is employed to measure the performance of different numbers of clusters used. Then, the number of clusters that give the highest value of the silhouette coefficient is the optimal number of clusters. The number of clusters ranges from two to ten, and then the optimal number of clusters obtained from the elbow method is used to cluster the samples into different levels of student engagement. According to Kaufman and Rousseeuw [35] the assessment criteria for a clustering algorithm based silhouette coefficient is as shown in Table V.

TABLE V. KAUFMAN AND ROUSSEEUW ASSESSMENT CRITERIA

Silhouette Coefficient Value	Evaluation
0.71 – 1.00	Strong structure
0.51 – 0.70	Medium structure
0.25 – 0.50	Weak structure
≤ 0.25	No structure (bad structure)

IV. RESULT AND DISCUSSION

The result obtained from the K-Means clustering algorithm based on the data that is extracted from the Moodle logfiles will be discussed and analysed in this section. The performance of the clustering algorithm to cluster student engagement level is shown in Table VI.

TABLE VI. PERFORMANCE OF CLUSTERING ALGORITHM

k	**Silhouette Coefficient**
2	0.8630
3	0.7104
4	0.4334
5	0.4601
6	0.4614
7	0.4567
8	0.3270
9	0.3165
10	0.2495

The performance of the K-Means is illustrated in Table VI. The K-Means clustering algorithm performs with good structure based on Kaufman and Rousseeuw Assessment Criteria when k = 2 and k = 3. However, when k = 2 is not selected because the group of student engagement levels is too general, which is not detailed information for this research. Therefore, the result of k = 3 is selected since it gives an acceptable silhouette coefficient while providing more detail information for the cluster of student engagement levels. The cluster distribution of the attribute scatter plots is illustrated in Fig. 4.

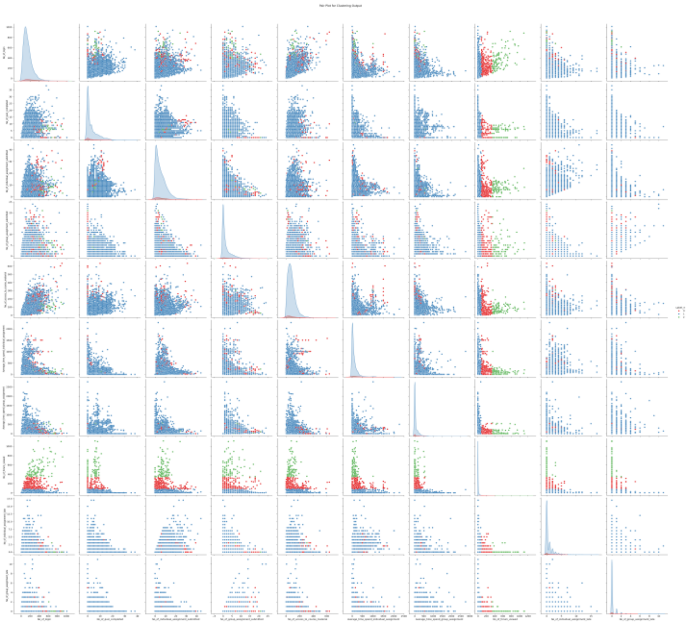

Fig. 4. Clustering result scatter plot

Fig. 4 shows the pair scatter plot of the clustering result based on k = 3. The three distinct clusters emerge are represented by blue, red, and green points. The first cluster (blue points) is located primarily in the lower left region of the plot, where both the X and Y variables exhibit lower values. In most of the scatter plots, this cluster is tightly packed, which means that the first cluster has high similarity between data points. In contrast, the second cluster (red points) is more widespread and occupies the central region of the plot. However, a few outliers are noticeable, especially around the edge of the cluster, this shows the presence of noise or exceptional cases that might need further investigation. Besides, the third cluster (green points) is also widespread and located on the upper portion of the plots.

Cluster 0 (blue points) is made up of 8918 students, which represents the students that are not actively engaged during online learning. The student at this cluster has a low number of logins into Moodle LMS, a low number of forums viewed, and a low number of assignments submitted, both individual and group assignments. More importantly, the students in this cluster have a high number of assignments that are submitted late. Even though the student in this cluster is having a high average time spent on group assignments, they are having very low average time spent on their individual assignments. In the context of this research, only the time spent on group assignments is measured by the time taken for the student to submit the assignment, but the participation of students in the assignment discussion is not measured. The high average time spent might be due the students in this cluster being 'sleeping partners' that enjoy the effort of their peers but do not participate in the assignment. Therefore, it can be summarised that the student in this cluster is not actively engaged during online learning since they are low in all four dimensions of student engagement.

Cluster 1 (red points) is made up of 383 students, representing the passively engaged during online learning. The students in this cluster are having high numbers of access to course materials and average time spent on individual assignments. In contrast, these students are spending minimal time on group assignments and having the minimal number of quizzes completed. In the meantime, they are having a higher number of logins, both individual and assignments submitted, and forum views as compared to cluster 0 but lower than cluster 2. On the other hand, the number of assignments submitted late for the students in this cluster is less than cluster 0 but more than cluster 2. The result shows the students in this cluster are more focused to giving effort on individual tasks rather than group tasks, and they are not active in social dimension engagement while moderately involved in the other three dimensions of engagement. According to Ayouni, *et al*. [16], passively engaged student participate less in group activities, however they are spending more time than the not engaged student. Passive engaged student performed few exercises or quizzes as compared to actively engaged student [26].

Cluster 2 (green points) is made up of 105 students that represent the actively engaged students during online learning. The students in this cluster actively log into the Moodle LMS, having a high number of quizzes submitted, assignments submitted, as well as the forum viewed. Even though the average time spent for both individual and group assignments of these students is lower than cluster 0 and cluster 1, respectively, the number of assignments submitted late is minimal. This result shows that the students in this cluster are good to time management and work effectively to make sure that they can complete the assignment on time without wasting unnecessary time and effort. In summary, the student in this cluster is active in all four dimensions of student engagement. Although they are having a lower number of accesses to course materials, this might be because they are looking for extra references from external sources rather than depending on the material provided. Actively engaged student are those participates in activities inside as well as inside the group activities [16]. This is because the actively engaged student are enthusiastically engaged in the learning process [26].

In conclusion, the student engagement is divided into three different levels that are not active, passive, and active by K-Means clustering in this research. This finding is similar to the research of Ayouni, *et al*. [16], Benabbes, *et al*. [26] and Tamba, *et al*. [25], where the student engagement level is divided into three levels. Students that are clustered as not actively engaged are having low participation in all four dimensions of student engagement and having a high tendency to submit assignments late. On the other hand, passively engaged students are moderately involved in the individual tasks but minimal participation in group activities and social interactions. Finally, actively engaged students are demonstrating strong time management, high participation across all engagement dimensions, and efficiency on task completion.

V. Conclusion

The proposed K-Means clustering approach was successfully used in this study to classify students into three groups based on their level of participation during online learning, achieving an acceptable result with a silhouette coefficient of 0.71. The results identified the following insights in three aspects of students' engagement regarding Cluster 0 as the least engaged students with low involvement in engagement aspects. The Cluster 1 refers to the student who is a passive learner and has low activity in terms of individual tasks, low participation in completing group tasks as well as few interactions with other students. Last of all, Cluster 2 refers to the participating students, who are well disciplined and organized within online learning and have excellent time management when participating in different activities. These results provide significant information on various approaches regarding student learning during online classes, helping instructors to focus on students who need reinforcement in their giftedness.

This paper adds to the current body of knowledge on student engagement since it offers a quantitative method to address the way students use Moodle LMS. Thus, by identifying student engagement clusters, the analysis reveals the potential range of activity and discusses which students are at risk of low academic performance, which can enhance the targeted approach to both educators and institutions. The result of clustering showing that most of the student are not actively engaged with LMS during online learning. This might be due to the reason that some of the courses are not highly using the features that in Moodle LMS, as shown in Table II out of the 288 courses that studied in this research only 118 are using LMS for forum features, while most of the courses are only using LMS for uploading the course resources and assignments submission. There are even some courses that used LMS only for uploading resources. The usage of forum is one of the important features to engaged student during class for discussion, but the usage is very limited.

Furthermore, the application of the clustering strategy also provides methodological contribution as the paper showcases that unsupervised learning approaches are feasible in educational data mining for investigating engagement level of students. Furthermore, the implementation of the Spearman Correlation Coefficient to reduce the attributes before clustering emphasises the importance of feature reduction in reducing the computational complexity of the clustering approach.

Despite this research providing a promising finding, this research has several limitations. First, the level of LMS usage varied significantly across the classes studied, with only 118 out of 266 classes utilising LMS forums for announcements and discussions, which could have influenced the clustering results. Besides, students' grades are not considered in clustering process, potentially limiting the understanding of how engagement correlates with academic performance. These limitations suggest that the findings may not fully capture the complete spectrum of student engagement level during online learning.

In the future, research should explore more alternative approaches, such as rule-based approaches that incorporate additional factors such as students' grade and the level of class LMS usage, to gain a more comprehensive understanding of student engagement patterns during online learning. Additionally, more different engagement and demographic attributes need to be explored, such as those related to social and collaborative interactions, and their impact on student academic success could further refine the categorisation of student engagement level. Lastly, in future the correlation between the level of lecturer utilization with LMS features during online learning and student engagement level in LMS need to be further explore.

ACKNOWLEDGMENT

The authors thank the Universiti Teknologi Malaysia (UTM) for giving an opportunity and providing support to carry out this research. Besides, the author would like to extend sincere gratitude to the Pusat Pengurusan Data Universiti (PPDU) UTM and the FC Academic Office UTM for their invaluable support and prompt approval of student data for this research.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] J. Luo and T. Wang. (2020). Analyzing students' behavior in blended learning environment for programming education. *Proceedings of the 2020 The 2nd World Symposium on Software Engineering,* Chengdu, China. https://doi-org.ezproxy.utm.my/10.1145/3425329.3425346.

[2] K. H. Zhe and N. A. Saadon. (2024). Revolutionizing English vocabulary learning for Malaysian university students through gamified mobile learning. *International Journal of Innovative Computing, 14*(1), 49-55.

[3] S. J. M. Shahrol, S. Sulaiman, and H. Mohamed. (2023). Acceptance factors towards mobile technologies in learning English among rural students. *International Journal of Innovative Computing, 13*(2), 37-47.

[4] C. K. Ch'ng. (2024). Hybrid machine learning approach for predicting e-wallet adoption among higher education students in Malaysia. *Journal of Information and Communication Technology, 23*(2), 177-210.

[5] Ministry of Education Malaysia. (2015). *Malaysia Education Blueprint 2015-2025 (Higher Education)*.

[6] R. Nand, A. Chand, and M. Naseem. (2020). Analyzing students' online presence in undergraduate courses using Clustering. Doi: 10.1109/CSDE50874.2020.9411534.

[7] T. Binali, C.-C. Tsai, and H.-Y. Chang. (2021). University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification. *Computers & Education, 175*, 104315. Doi: https://doi.org/10.1016/j.compedu.2021.104315.

[8] V. A. Nguyen, Q. B. Nguyen, and V. T. Nguyen. (2018). A model to forecast learning outcomes for students in blended learning courses based on learning analytics. *Proceedings of the 2nd International Conference on E-Society,* E-Education and E-Technology, Taipei, Taiwan. https://doi-org.ezproxy.utm.my/10.1145/3268808.3268827.

[9] A. Kumar Veerasamy, D. D'Souza, M. V. Apiola, M. J. Laakso, and T. Salakoski. (2020). Using early assessment performance as early warning signs to identify at-risk students in programming courses. Doi: 10.1109/FIE44824.2020.9274277.

[10] Y. Yang, D. Hooshyar, M. Pedaste, M. Wang, Y. M. Huang, and H. Lim. (2020). Prediction of students' procrastination behaviour through their submission behavioural pattern in online learning. *Journal of Ambient Intelligence and Humanized Computing.* Doi: 10.1007/s12652-020-02041-8.

[11] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences (Switzerland), 10*(11), 3894. Doi: 10.3390/app10113894.

[12] N. Tomasevic, N. Gvozdenovic, and S. Vranes. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers and Education, 143*, 103676. Doi: 10.1016/j.compedu.2019.103676.

[13] S. Altaf, W. Soomro, and M. I. M. Rawi. (2019). Student performance prediction using multi-layers artificial neural networks: A case study on educational data mining. *Proceedings of the 2019 3rd International Conference on Information System and Data Mining,* Houston, TX, USA. https://doi-org.ezproxy.utm.my/10.1145/3325917.3325919.

[14] D. Hooshyar and Y. Yang. (2021). Predicting course grade through comprehensive modelling of students' learning behavioral pattern. *Complexity, 2021*, 7463631. Doi: 10.1155/2021/7463631.

[15] M. Bucos and B. Dragulescu. (2020). Student cluster analysis based on Moodle data and academic performance indicators. Doi: 10.1109/ISETC50328.2020.9301061.

[16] S. Ayouni, F. Hajjej, M. Maddeh, and S. Al-Otaibi. (2021). A new ML-based approach to enhance student engagement in online environment. *PLOS ONE, 16*(11), e0258788. Doi: 10.1371/journal.pone.0258788.

[17] M. Abu Talib, A. M. Bettayeb, and R. I. Omer. (2021). Analytical study on the impact of technology in higher education during the age of COVID-19: Systematic literature review. *Education and Information Technologies, 26*(6), 6719-6746.

[18] E. Hehir, M. Zeller, J. Luckhurst, and T. Chandler. (2021). Developing student connectedness under remote learning using digital resources: A systematic review. *Education and Information Technologies, 26*(5), 6531-6548.

[19] G. N. Santos *et al.* (2021). The scope of dental education during COVID‐19 pandemic: A systematic review. *Journal of Dental Education.*

[20] A. Gewerc, D. Persico, and V. Rodés-Paragarino. (2020). Guest editorial: Challenges to the educational field: Digital competence the emperor has no clothes: The COVID-19 emergency and the need for digital competence. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 15*(4), 372-380.

[21] A. E. R. Prof. Ghaleb, A. Kaba, and S. Eletter. (2021). The impact of demographic characteristics on academic performance: Face-to-face learning versus distance learning implemented to prevent the spread of COVID-19. *International Review of Research in Open and Distributed Learning, 22*(1), 91-110. Doi: http://dx.doi.org/10.19173/irrodl.v22i1.5031.

[22] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb. (2021). Intelligent decision support system for predicting student's e-

learning performance using ensemble machine learning. *Mathematics. 9*(17), 2078. Doi: 10.3390/math9172078.

[23] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research, 74*(1), 59-109.

[24] J. MacQueen. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281-297.

[25] A. R. Tamba, K. Lumbantoruan, A. Pakpahan, and S. Situmeang. (2023). A cluster and association analysis visualization using Moodle activity log data. *Int J Inf & Commun Technol., 2252*(8776), 8776,

[26] K. Benabbes, K. Housni, B. Hmedna, A. Zellou, and A. El Mezouary. (2023). A new hybrid approach to detect and track learner's engagement in e-learning. *IEEE Access.*

[27] T. Konishi. (2024). Means and issues for adjusting principal component analysis results.

[28] A. Abdullah, A. Kamaludin, and A. Romli. (2020). Uncovered user engagement pattern in lms during covid-19 pandemic using temporal visualization matrix. *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, IEEE, 1-5.

[29] S. S. Mustapha. (2024). High-dimensional data analysis using parameter free algorithm data point positioning analysis. *Applied Sciences, 14*(10), 4231,

[30] M. Sumalatha, M. Ananthi, A. Arvind, N. Navin, and C. Siddarth. (2014). Highly correlated feature set selection for data clustering. *2014 International Conference on Recent Trends in Information Technology*, IEEE, 1-4.

[31] D. Liu. (2022). Better private algorithms for correlation clustering. *Conference on Learning Theory*. PMLR, 5391-5412.

[32] N. H. Zainuddin and M. T. Abdullah. (2023). Hybrid correlation coefficient of spearman with mm-estimator.

[33] M. S. Khokhar, K. Cheng, M. Ayoub, and N. E. Rub. (2019). Data driven processing via two-dimensional spearman correlation analysis (2D-SCA). *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, IEEE, 1-7.

[34] S. Nogueira, K. Sechidis, and G. Brown. (2017). On the use of Spearman's rho to measure the stability of feature rankings. *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*, Springer, 381-391.

[35] L. Kaufman and P. J. Rousseeuw. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.