SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Analysis of Student Activities Based on Log Files in E-Learning Using Clustering Algorithm

Indra Maulana[1,*], Metta Mariam[1]

[1] Graduate Program in Technology and Vocational Education, Universitas Negeri Yogyakarta, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia

ARTICLE INFO

Article history:
Received 3 March 2024
Received in revised form 9 July 2024
Accepted 4 September 2024
Available online 22 September 2024

Keywords:

E-learning; Log files; MOOC

ABSTRACT

This study was conducted to analyse log data generated by e-learning platforms such as Moodle or similar platforms. The main objective of this research is to identify patterns and insights that can help improve the student learning experience and the efficiency of platform management. This study identifies the most effective clustering algorithms for grouping students based on their behaviour and achievements in the e-learning environment, namely K-means and K-Medoids. The methods used to determine the optimal number of clusters are the Silhouette Coefficient method and the Elbow method, utilizing both methods to determine the best clustering algorithm results. Based on the analysis using K-means and K-Medoids clustering methods on the log data of the programming algorithm course, the total number of action logs over a one-month period is 95,461, with an average of 892 action logs per participant. The distribution of action logs based on class (A, B, and C) shows variation in the number and average of action logs per class. Types of activities such as assignments, forums, video views, and course views have different average frequencies, with video views being the most frequently visited activity. Pearson correlations between activity types show strong relationships between activities, with the highest correlation between visits and course views. The optimal number of clusters based on the Elbow and K-Medoids methods is three clusters. Cluster 3 in K-Means has the best performance with the smallest DBI value (0.424) and the smallest centroid distance (21,168.534).

1. Introduction

The application of e-learning in the world of education continues to develop and has even become the key to the implementation of learning during the COVID-19 pandemic [1]. E-learning platforms not only provide a variety of teaching materials and learning resources but also help students in independent learning [2]. Most MOOC systems, i.e. e-learning, store data about teaching and learning actions in log files, which provide us with detailed information about learner behaviour [3]. Log data, or log files, is a collection or list of various actions that have been performed by users [4]. Some people view it as the time that learners devote to the subject matter [5], others argue that

* *Corresponding author.*
E-mail address: Indramaulana.2021@student.uny.ac.id

https://doi.org/10.37934/araset.53.1.115

the time learners spend watching video lectures, answering quizzes, submitting assignments, and participating in forum discussions [6]. This arises especially because the involvement has four components: cognitive, behavioural, emotional (affective), and social [7]. The second school of thought argues that MOOC learning will become a passive activity if only tracking student activity through clicks is required [8,9].

Data logs or log files are a collection or list of actions that have been taken by users [10]. The data recorded in the Moodle data log can be in the form of activity data, assignment timestamp, and ranking value or final grade [11]. Student learning outcomes are indicated by test scores, and student participation can be seen from the e-learning system logfile [12]. A learning management system (LMS) as a distance learning platform automatically records log file data, which is the number of clicks or minutes spent by students on a specific task. Such individual log files provide objective information about the use of learning strategies that go beyond self-reporting, which may be susceptible to memory distortion or social desire [13].

Logfile data analysis in e-learning using data mining is a crucial research field in the context of modern education. The main goal is to identify patterns and insights that can help improve the student learning experience and the efficiency of managing the platform. Data mining can be used to identify trends in learning behaviours, such as the most active times, the types of content that are most in demand, and interaction patterns between students. The results of the analysis of logfile data can assist educational institutions and instructors in taking relevant actions to improve the effectiveness of e-learning and improve student learning outcomes. Most e-learning systems store data about teaching and learning actions in log files, which provide us with detailed information about learner behaviour [14]. Log data, or log files, is a collection or list of various actions that have been performed by a user [15].

Analysing student behaviour using logfiles in e-learning is very important because it has a major impact on learning effectiveness and progress, because log files record the sequence of user actions when learning online [16]. E-learning platforms not only provide various teaching materials and learning resources but also help students in independent learning [17]. With the help of logfile data, the study can identify areas that need improvement in curriculum design, teaching methods, or interactions between instructors and students. One of the techniques used in the analysis of logfile data is cluster analysis using the K-Means and K-Medoids algorithms. The researchers investigated features such as resource usage, frequency of actions, average latency, login frequency, number of module accesses, login time, login regularity, total learning time, and learning interval regularity [17-19]. One of the techniques used in log data analysis is cluster analysis. Cluster analysis is the process of grouping data into groups whose members have similar characteristics [20].

Most previous studies have focused on the use of clustering algorithms or individual log data analysis, but not much has discussed direct comparisons between various clustering algorithms in the context of e-learning. In addition, many studies have only looked at one method of determining the number of clusters, so there is no comprehensive approach to determine the most effective algorithms and methods. This study offers an in-depth comparative analysis between two clustering algorithms (K-Means and K-Medoids), which has not been widely discussed in the context of e-learning. The use of two methods for determining the optimal number of clusters, namely the Silhouette Coefficient and Elbow Method, provides a more accurate and reliable approach.

This study aims to identify the most effective clustering algorithm in grouping students based on their behaviour and achievements in an e-learning environment. By using these two methods, it is hoped that the best clustering algorithm results can be determined. In addition, the results of this study can be widely applied in the context of online education, with the potential to provide real benefits in improving the effectiveness of digital learning, as well as provide valuable insights to

optimize the learning experience in an e-learning environment with the best clustering algorithms. This allows educational institutions and teachers to develop more effective strategies in ensuring student learning success within the platform.

1.1 Logfile

Log data or log files are collections or lists of various actions that have been carried out by users [21]. The data recorded in the Moodle data log can include activity data, assignment timestamps, and ranking or final grades (grade) [22]. For more details, please refer to Table 1.

Table 1
Approach used for logfile-based learning evaluation

Studies	Evaluation approach categories	Evaluation approach sub category	Techniques/ description of techniques used
Labarthe *et al.,*	Modelling	Engagement patterns	Cluster analysis
Shi & Cristea	Modelling	Engagement patterns	Cluster analysis
Deng *et al.,*	Modelling	Engagement patterns	Cluster analysis
Sun *et al.,*	Modelling	Structural equations modelling	Partial least squares
Anutariya	Modelling	Machine learning models	Cluster analysis
Maniriho & Effendi	Modelling	Machine learning models	Cluster analysis
Kadoic & Oreski	Modelling	Machine learning models	Cluster analysis

Logfiles have a very important role in the context of learning theory, especially in the online learning era. Logfiles record student activities and interactions with learning materials in detail. This allows for a deep understanding of how students learn, what their learning styles are like, how often they participate in discussions, and how they process information. This research is an important foundation for understanding how logfile analysis can be used to personalize learning [23,24]. A deep understanding of student learning patterns, which can be gained through logfile analysis, helps in designing responsive curricula. Data Mining in analysing data is very important to understand how to extract valuable information from logfiles. This includes techniques such as clustering, classification, and pattern recognition that can be used to identify student learning patterns and trends in logfile data. Data mining has a crucial role in analysing e-learning logfiles, one of which is using clustering algorithms [25].

1.2 Clustering Algorithms

Clustering is the process of partitioning a set of data objects into subsets called clusters. Objects in a cluster have similar characteristics to each other and are different from other clusters. Partitioning is not done manually but with a clustering algorithm. The clustering algorithm divides the population or data points with the same characteristics into several small groups for grouping. And so, on that the cluster will form like a tree, where there is a clear hierarchy between objects, from the most similar to the least similar [17,18]. For more details, please refer to Figure 1.

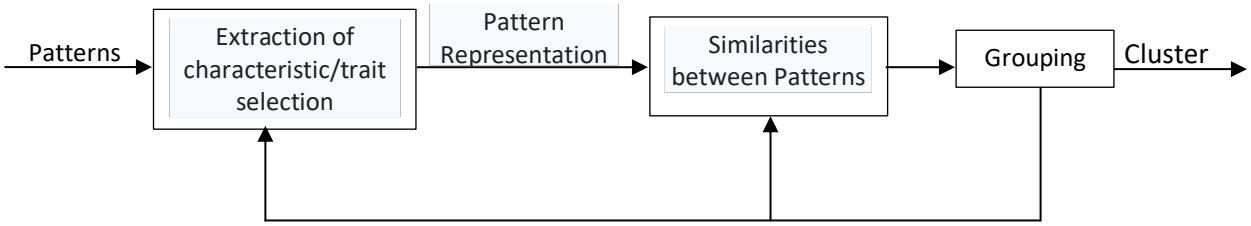


Fig. 1. Clustering stages

1.3 K-Means Clustering

K-means clustering is a non-hierarchical cluster analysis method that attempts to partition existing objects into one or more clusters or groups of objects based on their characteristics, so that objects that have the same characteristics are grouped in the same cluster and objects that have the same characteristics. different groups are grouped into other clusters. The K-Means Clustering method attempts to group existing data into several groups (Figure 2), where the data in one group has the same characteristics as each other and has different characteristics from the data in other groups.

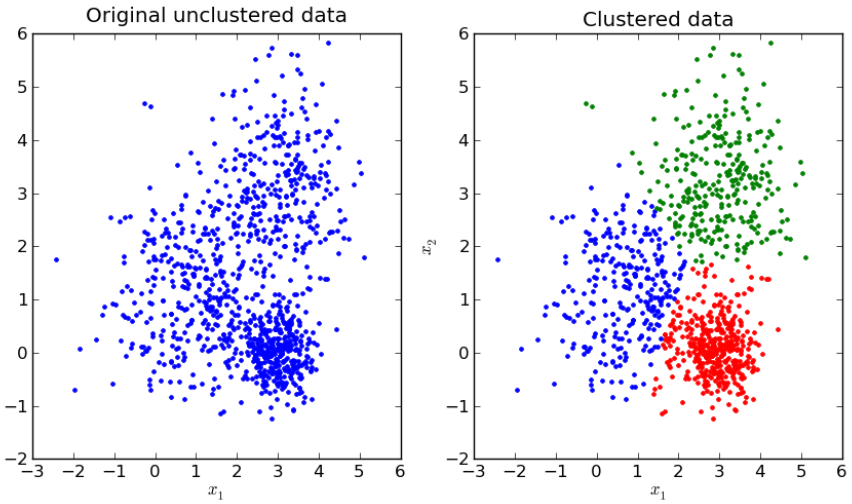


Fig. 2. Cluster Visualization

In other words, the K-Means Clustering method aims to minimize the objective function set in the clustering process by minimizing variations between data in one cluster and maximizing variations with data in other clusters. It also aims to find groups in the data, with the number groups represented by variable K. Variable K itself is the number of desired clusters. Divide data into groups. This algorithm accepts input in the form of data without class labels. This is different from supervised learning which receives input in the form of vectors (x1, y1), (x2, y2), ..., (xi, yi), where xi is data from training data and yi is the class label for xi.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{1}$$

where:
dij = distance between data i to data j
xik = i-th testing data
xjk = i-th training data

Updating a centroid point can be done with Eq. (2):

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \tag{2}$$

where:
μk = centroid point of the K-th cluster
Nk = the number of data in the K-th cluster
xq = qth data in the Kth cluster

1.4 K-Medoids

K-Medoids Clustering, also known as Partitioning Around Medoids (PAM), is a variant of the K-Means method. This is based on the use of medoids instead of observing the mean belonging to each cluster, with the aim of reducing the sensitivity of the resulting partition with respect to the extreme values present in the dataset [19]. The first step taken is to determine the most representative point (medoid) in the data group by calculating the distance within a cluster from all combinations of medoids so that the distance between points in a group is small, while the distance between points between groups is large [20]. K-Medoids are objects that represent their reference points, not taking a value as the mean of an object in each group. The algorithm will take parameters from input k, with the number of groups that will be separated between one part of n objects [21]. The steps of the K-Medoids algorithm are as follows:

i. Initiate a cluster centre (number of clusters)
ii. Allocate each data (object) to the closest cluster using the Euclidian Distance equation as in Eq. (1):
iii. Select objects randomly in each cluster as candidates for the new medoid.
iv. Then compute the distance to each object in each cluster with a new candidate field.
v. Next, compute the total deviation (S) by computing the new total distance value with the old total distance value. When S<0, then swap objects with cluster data to form a new set of k objects as medoids.
vi. Then repeat the steps 3 to 5 until there is no change in the field so that the clusters and members of each cluster are obtained.

$$d(x,y) = \| x - y \| = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \tag{3}$$

where :
d(x,y) = distance between the ith data and the jth data

xi = value of the first attribute of the ith data
yi = value of the first attribute of the jth data
n = number of attributes used

1.5 Elbow Method

The elbow method is to choose a k value that is small and still has a low within value. This method and analysis are used to select the optimal number of clusters or groups. Below is the elbow algorithm for determining the number of groups formed [31]. Namely based on the sum of square error (SSE).

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} \|X_i - C_k\|^2 \quad (4)$$

1.6 Davies Bouldin Index

The Davies Bouldin Index (DBI) was introduced by David L. Davies and Donald W. Bouldin in 1979 as a metric for evaluating the results of clustering algorithms. Evaluation using the Davies Bouldin Index has an internal cluster evaluation scheme, where whether the cluster results are good or not is seen from the quantity and closeness between them. cluster result data. This measurement with the Davies-Bouldin Index maximizes the inter-cluster distance between clusters Ci and Cj and at the same time tries to minimize the distance between points in a cluster. If the inter-cluster distance is maximal, it means that the similarities in characteristics between each cluster are small so that the differences between clusters are more clearly visible. If the intra-cluster distance is minimal, it means that each object in the cluster has a high level of similarity in characteristics. The formula for calculating the Davies-Bouldin Index (BDI) with Eq. (5):

$$DBI = \frac{1}{k} . \sum_{i=1}^{k} R_i$$

With

Ri = max Rij (5)

and

$$R_{ij} = \frac{var(C_i)+var(C_j)}{\|C_i - C_j\|} \quad (6)$$

Where:
Ci: cluster I and Ci is the centroid of cluster i.

2. Methodology

This research uses a quantitative approach to look at the picture of log data and see the grouping patterns of training participants based on training activities. The data used in this research comes from data from training participants (user participants) and data logs for the course "Algorithms and Programming" which was held on 11 September - 10 October 2022. The number of participants who took this course was 124 participants apart from the instructors and course admins who were obtained from the user participants table in the LMS. The data log obtained from the LMS contains

19027 rows and 9 columns. In this research, e-learning experiences are tried to be presented as interaction patterns. For this purpose, LMS log data is used. The above dataset includes user activity in an e-learning system. The following is a description of the dataset, see Table 2:

Table 2
Description of the Dataset

Timestamp	User_ID	Activity_Type	Module_ID
09:00:00	ALP002	Login	
09:05:00	ALP 003	View_Module	Module_1
09:15:00	ALP 004	Submit_Assignment	Assignment_1
09:30:00	ALP 005	Forum_Discussion	Discussion_1
10:00:00	ALP 006	Login	
10:10:00	ALP 007	View_Module	Module_1
10:30:00	ALP 008	View_Module	Module_2
10:45:00	ALP 009	View_Module	Module_2
14:10:00	ALP 010	View_Module	Module_1
14:30:00	ALP 011	View_Video	Video_1
14:45:00	ALP 012	View_Module	Module_2
15:00:00	ALP 013	Submit_Assignment	Assignment_2
12:00:00	ALP 014	Submit_Assignment	Assignment_1
12:30:00	ALP 015	Forum_Discussion	Discussion_2
….	….	….	….

i. Timestamp: This column records the time when the activity occurred, including the date and time.
ii. User_ID: This is a unique identification for each user in the e-learning system. This dataset includes 50 different User_IDs.
iii. Activity_Type: This column indicates the type of activity performed by the user. Activities can be "Login" (enter the system), "View_Module" (view learning modules), "Submit_Assignment" (send assignments), "Forum_Discussion" (participate in forum discussions), and "View_Video" (watch learning videos).
iv. Module_ID: This is a unique identification for each learning module, assignment, discussion, or video. These modules are part of the learning experience in the e-learning system.

This dataset records various types of user interactions in e-learning systems, from logging into the system to participating in discussions, watching videos, and submitting assignments. This data can be used to analyse student learning patterns, evaluate user performance, and derive insights to improve the learning experience in e-learning platforms.

Data on training participants (user participation) consisting of 116 participants contains attributes regarding name, email, agency, education and gender. This participant data will later be subjected to descriptive analysis to see a picture of the training participants in the "Algorithms and Programming" course as a whole and based on the characteristics of educational level and institution of origin. In addition, correlation analysis was carried out between variables to see the relationship between the activities carried out by participants on the course. Correlation analysis in this research is not connected as a basic assumption in using k-means clustering because log data is data on a series of LMS user activities and each activity is always connected to other activities.

The event log data used contains 15402 rows and 4 columns. The log data used in this research is focused on the activities of course participants so that admin or teacher activities are not included in

the data log lines. Next, from the existing data, a data preprocessing process is carried out with the following steps:

i. Eliminate Admin and Teacher activities
ii. Eliminate logs containing CLI
iii. Perform conversions for the time attribute
iv. Eliminate duplicate logs.

The available data logs are then carried out in the preprocessing stage using the help of RapidMiner. This stage includes the process of mutation, aggregation and data transformation to produce derivative data with the attributes in Table 3 as follows:

Table 3
Description of Log Data Attributes from Moodle After Data Preprocessing

Attribute	Description
Username	Student name
Assignments	Number of assignment submissions
Forum	Number of question-and-answer forums created
URLs	Number of links visited
CourseViews	The number of activities seen/read by students

3. Results

Before entering data mining analysis using the k-means clustering method, an exploratory analysis of the data is first carried out in the form of descriptive statistics. The results of the analysis through descriptive statistics discuss summary statistics of actions carried out by course participants, distribution of actions based on agency characteristics, distribution of actions based on educational level characteristics, and summary statistics based on type of activity. Actions are actions carried out by users in the LMS or you could say the number of clicks made by course participants on course activities in the LMS. See Table 4.

Table 4
Action Log Statistics

Size	Mark
Number of Learned Participants	107
Number of Action Logs	95461
Average participant action log	892

Furthermore, the results of descriptive statistics on participant data and log data generated from 107 students in the Programming Algorithms with Tableau course are produced in Table 3 to Table 5. In Table 3 for the time period September 11 – October 10, 2022.or for 1 month has occurred95461activities carried out by 107 course participants. The average action taken per course participant was 31,820 actions with a size of distribution calculated through a standard deviation (SD) of 39. See Table 5 and Table 6.

Table 5
Action Log by Class

Class	The number of students	Number of Action logs	Average action log
A	32	29958	936
B	38	33683	886

C	35	31820	909

Table 6
Action Log by Type

Type of action log	Number of action logs	Average
Assignment	2186	20.4
Forum	16609	155.2
ViewVideo	58248	544.4
ViewCourse	20604	192.6

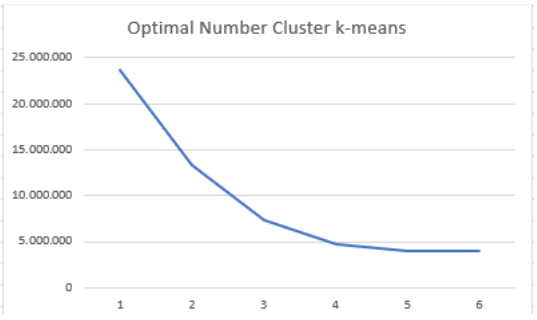
Based on the summary of activity statistics in the Algorithms and Programming course data log in Table 5, the frequency of actions can be seen based on the type of activity Assignment, Forum, View Video, and View Course. On average (mean) as seen in Table 5, the most frequently visited action is visiting courses (View Video) with an average of 544.4, followed by opening modules/activities in the course (View Course) with an average of 192.6, followed by forums (Forum) with an average of 155, and assignments 20.4.

Apart from Exploratory Data Analysis (EDA) which is presented through statistical summaries in tabular form in Table 3 to Table 5, EDA is also presented through images of the Pearson correlation (r) between types of activity which can be seen in Figure 1. Based on Figure 1, the Pearson correlation (r) Between all actions, each activity is classified as strong. The largest correlation value between activities is the correlation between Visit and CourseViews, which is 0.98, and the smallest correlation is the correlation between Quizes and Forum Created, which is 0.25. Based on the Pearson correlation significance test, these course activity variables are all significant for further use in clustering analysis using the k-means clustering method. See Table 7.

Table 7
Pearson Correlation Analysis of Participants' Activities

Attribute	Assignment	Forum	View Course	View Video
Assignment	1	-0.008	0.015	0.161
Forum	-0.008	1	0.792	0.017
View Course	0.015	0.792	1	0.211
View Video	0.061	0.017	0.211	1

Data mining analysis using k-means clustering begins with determining the optimal number of clusters using 3 methods for determining the number of clusters (k). Figure 3 shows that the optimal number of clusters to be formed is 3 clusters so that later using the K-Means clustering method 3 clusters of participants in the Algorithms and Programming course will be formed based on the type of action activity they carry out [30-32]. The Elbow method in this research was carried out 6 times. Testing the K value starts from K= 2 to K= 6, and produces an optimal number of clusters of 3 clusters. See Figure 3.



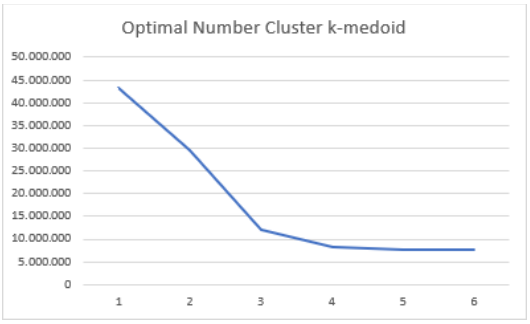


Fig. 3. Optimal Number of Clusters using the Elbow Method

In more detail, a comparison of the Elbow Score values from the clustering results using the method specified in Table 8 is shown.

Table 8
Comparison of the Elbow Score Values

Number of Clusters	Elbow scores	
	K-Means	K-Medoid
2	23,657,782	43,362,682
3	13,330,175	29,439,604
4	7,297,903	11,974,617
5	4,725,393	8,387,948
6	3,966,381	7,773,968

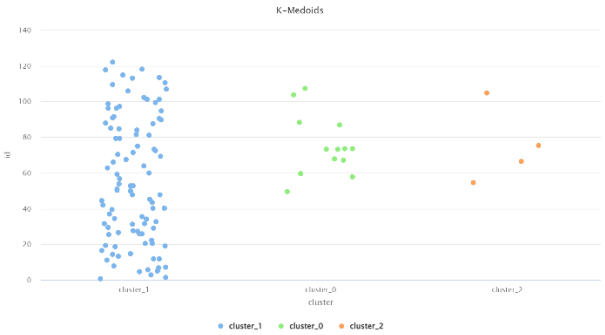


Fig. 4. K-Medoids

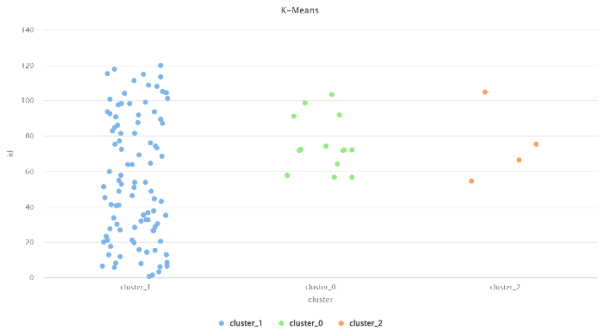


Fig. 5. K-Means

The parameters used to measure the performance of the K-Means and K-Medoids algorithms are done by calculating avg._within_centroid_distance and davies_bouldin. avg._within_centroid_ distance is the average cluster distance calculated from the average distance between the centroid and all cluster samples. Meanwhile, Davies_bouldin is an algorithm that produces clusters with low intra-cluster distance (high level of intra-cluster similarity) and high inter-cluster distance (low level of inter-cluster similarity) which will have a low Davies-Bouldin index [29]. The clustering algorithm that produces clusters with the smallest Davies-Bouldin index is considered the best algorithm based on the criteria [33,34]. The results of measuring the performance of the K-Means cluster algorithm in grouping student learning activity data in Padi e-learning log files using the operator cluster performance distance can be seen in Table 8.

From the cluster results in Figure 6, K-Medoids shows that there are 13 members of the activity cluster or student logs in cluster 1 or 11.21%, cluster 2 has 99 members or 85.34%, and cluster 3 has 4 or 3.448%. Meanwhile, in the K-Means algorithm, there are differences in the number of each cluster, namely Cluster 1 is 98 or 84.48%. Cluster 2 was 14 or 12.07%, and finally Cluster 3 or 4 or 3,448%.

Table 9
Davies Bouldin Index (DBI)

	K-Means	K-Medoid
Davies Bouldin Index	0.424	1,358

Table 10
Centroid distance values

	K-Means	K-Medoid
centroid distance	21168.534	15314.791

From the comparison of DBI values, it can be seen that the smallest DBI value is in cluster 3 using the K-Means algorithm, namely 0.424, this shows the performance of the best clustering algorithm chosen. The number of elements for each group resulting from the clustering process with 3 clusters using the K-Means algorithm can be seen in Figure 6.

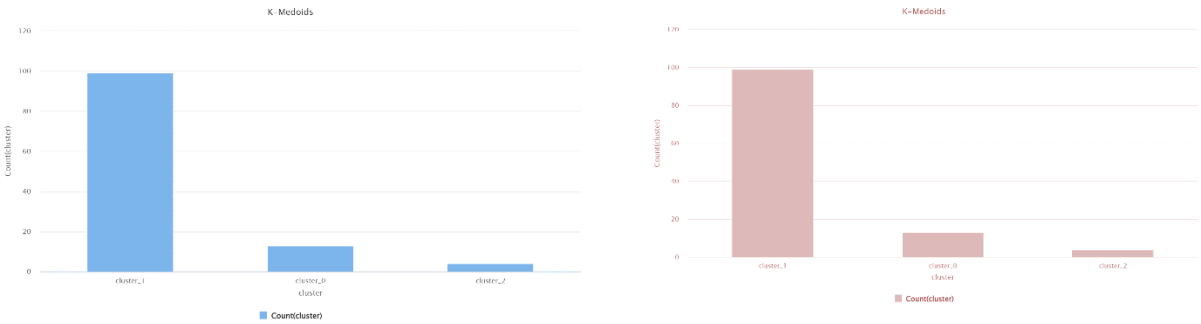


Fig. 6. Number of members of each cluster

Based on the results of the data mining process, it shows that the elements in the groups formed are separated into each group, thus there are no group elements that fall into two groups. A cluster can be declared convergent if there is no movement or change in elements from one cluster to another [28]. This proves that the elements resulting from clustering can represent each cluster.

Learner engagement in an e-learning platform is difficult to define. Some view it as the time the learner devotes to the course material [32], there are also those who argue that students spend time watching video lectures, answering quizzes, submitting assignments, and participating in forum discussions [33]. This arises especially because such involvement has four components: cognitive, behavioural, emotional (affective), and social [17]. The second school of thought argues that MOOC learning would be a passive activity if it only required tracking learner activity via clicks [8,9].

As shown in Table 5, the frequency of action can be seen based on the type of activity Assignment, Forum, ViewVideo, and ViewCourse. On average (mean) as seen in Table 5, the most frequently visited action is visiting courses (View Video) with an average of 544.4, followed by opening modules/activities in the course (View Course) with an average of 192.6, followed by forums (Forum) with an average of 155, and assignments 20.4. The participants' engagement patterns are not attached to their personality traits but only reflect a person's behaviour in the learning experience on the e-learning platform. This is then reinforced by the correlation results shown in Table 7 that there is a significant correlation between the action logs.

Research on logfile data clustering in e-learning has become an increasingly important research subject in the context of digital learning. cluster analysis method with data in the form of log files as done by [29-32]. And to find out the best algorithm for grouping student log files, the researchers looked for the best among the K-Means and K-Medoids algorithms which are shown in Table 5 to Table 8.

From the results of experiments using the elbow method 6 times, and resulting in the optimal number of clusters being 3 clusters, the K-Means elbow value was 13,330,175 and K-Medoids 29,439,604 It can be seen that the smallest DBI value is in cluster 3 using the K-Means algorithm,

namely 0.424, this shows the performance of the best clustering algorithm chosen. The number of elements for each group resulting from the clustering process with 3 clusters using the K-Means algorithm is in accordance with the results [34,35]. The results of this research show that the K-Means algorithm is better than K-Medoids based on figures from DBI, so K-Means is suitable for clustering student activities in e-learning platforms based on log files. Since log files record a series of user actions during online learning, using log files to analyse learning behaviour in e-learning is very important because it has a significant impact on learning effectiveness and progress [36]. E-learning platforms not only provide various teaching materials and learning resources but also help students in independent learning [33].

This study proves that data mining can be used to identify trends in learning behaviour, such as the most active time of day, the most popular types of content, and interaction patterns between students. The results of log file data analysis can help educational institutions and teachers take relevant steps to increase the effectiveness of e-learning and improve student learning outcomes. Most e-learning systems store data about teaching and learning activities in log files, which provide detailed information about learner behaviour [18]. Log data or log files are a collection or list of different actions carried out by users [20].

4. Conclusions

Based on data mining analysis using the k-means and K-Medoids clustering method on the data log of the Programming Algorithms course, there were 124 participants in the Programming Algorithms course. The total number of action logs over a period of one month was 95,461, with an average of 892 action logs per participant. The division of action logs by class (A, B, and C) shows variations in the number and average of action logs per class:

i. Activity types such as Assignment, Forum, View Video, and View Course have different average frequencies, with View Video being the most frequently visited activity.
ii. Pearson correlations between activity types showed a strong relationship between activities, with the highest correlation between Visit and CourseViews.
iii. The optimal number of clusters based on the elbow and k-medoids methods is 3 clusters.
iv. Cluster 3 in K-Means has the best performance with the smallest DBI value (0.424) and the smallest centroid distance (21,168.534).
v. The distribution of cluster members shows that Cluster 1 has the largest number of members (84.48%).

The clustering results can be used to provide more appropriate adjustments or recommendations according to the cluster characteristics of course participants. Thus, this analysis provides useful insights in understanding the behaviour patterns of course participants and can be used to improve the learning experience by providing recommendations that are more focused on the needs of each group of participants and also the performance of the best clustering algorithm in this grouping is K-means because it has the smallest DBI value (0.424).

Acknowledgement
The authors would like to express their deepest gratitude to the Center for Higher Education Funding (BPPT) and the Education Fund Management Institution (LPDP) of the Republic of Indonesia, which have provided the Indonesian Education Scholarship (BPI) with number 202101122150, sponsoring

the author's doctoral study and supporting the completion of this research study and the publication of this article.

References

[1] Hani, Amjad Bani, Yazan Hijazein, Hiba Hadadin, Alma K. Jarkas, Zahraa Al-Tamimi, Marzouq Amarin, Amjad Shatarat, Mahmoud Abu Abeeleh, and Raed Al-Taher. "E-Learning during COVID-19 pandemic; Turning a crisis into opportunity: A cross-sectional study at The University of Jordan." *Annals of Medicine and Surgery* 70 (2021): 102882. https://doi.org/10.1016/j.amsu.2021.102882

[2] Jansen, Renée S., Anouschka van Leeuwen, Jeroen Janssen, Rianne Conijn, and Liesbeth Kester. "Supporting learners' self-regulated learning in Massive Open Online Courses." *Computers & Education* 146 (2020): 103771. https://doi.org/10.1016/j.compedu.2019.103771

[3] Studiawan, Hudan, Ferdous Sohel, and Christian Payne. "A survey on forensic investigation of operating system logs." *Digital Investigation* 29 (2019): 1-20. https://doi.org/10.1016/j.diin.2019.02.005

[4] Lam, Pham Xuan, Phan Quoc Hung Mai, Quang Hung Nguyen, Thao Pham, Thi Hong Hanh Nguyen, and Thi Huyen Nguyen. "Enhancing educational evaluation through predictive student assessment modeling." *Computers and Education: Artificial Intelligence* 6 (2024): 100244. https://doi.org/10.1016/j.caeai.2024.100244

[5] Nishitani, Kimitaka, Jeffrey Unerman, and Katsuhiko Kokubu. "Motivations for voluntary corporate adoption of integrated reporting: A novel context for comparing voluntary disclosure and legitimacy theory." *Journal of Cleaner Production* 322 (2021): 129027. https://doi.org/10.1016/j.jclepro.2021.129027

[6] Riestra-González, Moises, Maria del Puerto Paule-Ruíz, and Francisco Ortin. "Massive LMS log data analysis for the early prediction of course-agnostic student performance." *Computers & Education* 163 (2021): 104108. https://doi.org/10.1016/j.compedu.2020.104108

[7] Canay, Özkan, and Ümit Kocabıçak. "An innovative data collection method to eliminate the preprocessing phase in web usage mining." *Engineering Science and Technology, an International Journal* 40 (2023): 101360. https://doi.org/10.1016/j.jestch.2023.101360

[8] Dehury, Chinmaya Kumar, Satish Narayana Srirama, and Tek Raj Chhetri. "CCoDaMiC: a framework for coherent coordination of data migration and computation platforms." *Future Generation Computer Systems* 109 (2020): 1-16. https://doi.org/10.1016/j.future.2020.03.029

[9] Garaialde, Diego, Anna L. Cox, and Benjamin R. Cowan. "Designing gamified rewards to encourage repeated app selection: Effect of reward placement." *International Journal of Human-Computer Studies* 153 (2021): 102661. https://doi.org/10.1016/j.ijhcs.2021.102661

[10] Sasi, Tinshu, Arash Habibi Lashkari, Rongxing Lu, Pulei Xiong, and Shahrear Iqbal. "A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges." *Journal of Information and Intelligence* (2023). https://doi.org/10.1016/j.jiixd.2023.12.001

[11] Sun, Leilei, Guoqing Chen, Hui Xiong, and Chonghui Guo. "Cluster analysis in data-driven management and decisions." *Journal of Management Science and Engineering* 2, no. 4 (2017): 227-251. https://doi.org/10.3724/SP.J.1383.204011

[12] Landauer, Max, Florian Skopik, Markus Wurzenberger, and Andreas Rauber. "System log clustering approaches for cyber security applications: A survey." *Computers & Security* 92 (2020): 101739. https://doi.org/10.1016/j.cose.2020.101739

[13] Dascalu, Maria-Dorinela, Stefan Ruseti, Mihai Dascalu, Danielle S. McNamara, Mihai Carabas, Traian Rebedea, and Stefan Trausan-Matu. "Before and during COVID-19: A Cohesion Network Analysis of students' online participation in moodle courses." *Computers in Human Behavior* 121 (2021): 106780. https://doi.org/10.1016/j.chb.2021.106780

[14] Kandia, I. Wayan, Ni Made Suarningsih, Wahdah Wahdah, Arifin Arifin, Jenuri Jenuri, and Dina Mayadiana Suwarma. "The strategic role of learning media in optimizing student learning outcomes." *Journal of Education Research* 4, no. 2 (2023): 508-514.

[15] Jumrio, Edy. "The Function Of Online Learning In Creating Human Resources In The Digital Age." *Stipas Tahasak Danum Pambelum Keuskupan Palangkaraya* 1, no. 2 (2023): 170-185.

[16] Chi, Cai, Melor Md Yunus, Karmila Rafiqah M. Rafiq, Hamidah Hameed, and Ediyanto Ediyanto. "A Systematic Review on Multidisciplinary Technological Approaches in Higher Education." *International Journal of Advanced Research in Future Ready Learning and Education* 36, no. 1 (2024): 1-10. https://doi.org/10.37934/frle.36.1.110

[17] Suprayogi. "Data Mining Clustering." *Economic Studies, 2001.* (2018).

[18] Madhulatha, T. Soni. "An overview on clustering methods." *arXiv preprint arXiv:1205.1117* (2012). https://doi.org/10.9790/3021-0204719725
[19] Bakkelund, J., Karlsen, R., Bjørke, Ø., Suryakumar, S., Karunakaran, K. P., Bernard, A., Chandrasekhar, U., Raghavender, N., Sharma, D., Çelik, A., Yaman, H., Turan, S., Kara, A., Kara, F., Zhu, B., Qu, X., Tao, Y., Zhu, Z., Dhokia, V., … Dutta, D. "Analysis of the co-dispersion structure of health-related indicators, the center of the subject's sense of health, and the elderly people living at home." *Journal of Materials Processing Technology*, 1(1), (2018): 1–8.
[20] Kadoić, Nikola, and Dijna Oreški. "Analysis of student behavior and success based on logs in Moodle." In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0654-0659. IEEE, 2018. https://doi.org/10.23919/MIPRO.2018.8400123
[21] Labarthe, Hugues, François Bouchet, Rémi Bachelet, and Kalina Yacef. "Does a peer recommender foster students' engagement in MOOCs?." In *9th international conference on educational data mining*, pp. 418-423. 2016.
[22] Maniriho, Pascal, and Ari Effendi. "Examining the performance of k-means clustering algorithm." *Journal of Research in Engineering, Science and Management* 1 (2018): 1-5.
[23] Rabiaa, Elkamel, Baccar Noura, and Cherif Adnene. "Improvements in LEACH based on K-means and Gauss algorithms." *Procedia computer science* 73 (2015): 460-467. https://doi.org/10.1016/j.procs.2015.12.046
[24] Shi, Lei, and Alexandra I. Cristea. "In-depth exploration of engagement patterns in MOOCs." In *Web Information Systems Engineering–WISE 2018: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part II 19*, pp. 395-409. Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-02925-8_28
[25] Singh, K. K., U. Kumar, and K. Anurupam. "Activity based students' performance measure using moodle log files." *International Journal of Research in Advent Technology* 7, no. 7 (2019): 2-4. https://doi.org/10.32622/ijrat.77201901
[26] Sun, Yongqiang, Linghong Ni, Yiming Zhao, Xiao-Liang Shen, and Nan Wang. "Understanding students' engagement in MOOCs: An integration of self-determination theory and theory of relationship quality." *British Journal of Educational Technology* 50, no. 6 (2019): 3156-3174. https://doi.org/10.1111/bjet.12724
[27] Syakur, Muhammad Ali, B. Khusnul Khotimah, E. M. S. Rochman, and Budi Dwi Satoto. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." In *IOP conference series: materials science and engineering*, vol. 336, p. 012017. IOP Publishing, 2018. https://doi.org/10.1088/1757-899X/336/1/012017
[28] Sukmawati, Wati, Asep Kadarohman, Omay Sumarna, and Wahyu Sopandi. "Investigation of the independence of pharmacy students in blended learning." In *AIP Conference Proceedings*, vol. 2734, no. 1. AIP Publishing, 2023. https://doi.org/10.1063/5.0157700
[29] Wahjusaputri, Sintha, Tashia Indah Nastiti, Bunyamin Bunyamin, and Wati Sukmawati. "Development of artificial intelligence-based teaching factory in vocational high schools in Central Java Province." *Journal of Education and Learning (EduLearn)* 18, no. 4 (2024): 1234-1245. https://doi.org/10.11591/edulearn.v18i4.21422
[30] Sukmawati, Wati, Asep Kadarohman, Omay Sumarna, and Wahyu Sopandi. "Investigation of the independence of pharmacy students in blended learning." In *AIP Conference Proceedings*, vol. 2734, no. 1. AIP Publishing, 2023. https://doi.org/10.1063/5.0157700
[31] Mohamed, Rosmawati, Mohd Zaid Mamat, and Anuar Ab Razak. "Using GeoGebra with Van Hiele's Model in Geometry Classroom: An Experience with Prospective Teacher." *Semarak International Journal of STEM Education* 1, no. 1 (2024): 1-19. https://doi.org/10.37934/sijste.1.1.119
[32] Ramli, Noraini, and Mohd Ekram Al Hafis Hashim. "Interactive AR Textbook Application For 3M Orang Asli Students in Primary School." *Semarak International Journal of Innovation in Learning and Education* 2 (2024): 1-24. https://doi.org/10.37934/sijile.2.1.124
[33] Sidhu, Pramita, Fazlin Shasha Abdullah, and Mohamad Sirajuddin Jalil. "Awareness and Readiness of Malaysian Generation Z Students towards the Fourth Industrial Revolution (IR4. 0)." *Semarak International Journal of STEM Education* 1, no. 1 (2024): 20-27. https://doi.org/10.37934/sijste.1.1.2027
[34] Sukmawati, Wati, Asep Kadarohman, Omay Sumarna, and Wahyu Sopandi. "The use of conceptual change text (CCT) based teaching materials to improve multiple ability of pharmaceutical chemical representation students." In *AIP Conference Proceedings*, vol. 2468, no. 1. AIP Publishing, 2022. https://doi.org/10.1063/5.0102578
[35] Hishamuddin, Fatimah, Khalidah Ahmad, Halina Kasmani, Nur Bahiyah Abdul Wahab, Mohd Zulfahmi Bahaudin, and Elme Alias. "Empowering Leaders: A Work in Progress on Promoting Leadership Roles in Online Learning through Project-Based Learning (PBL)." *Semarak International Journal of Innovation in Learning and Education* 2, no. 1 (2024): 65-73. https://doi.org/10.37934/sijile.2.1.6573
[36] Chi, Cai, Melor Md Yunus, Karmila Rafiqah M. Rafiq, Hamidah Hameed, and Ediyanto Ediyanto. "A Systematic Review on Multidisciplinary Technological Approaches in Higher Education." *International Journal of Advanced Research in Future Ready Learning and Education* 36, no. 1 (2024): 1-10. https://doi.org/10.37934/frle.36.1.110