Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 3 (2022) 448 - 456 ISSN Media Electronic: 2580-0760

Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining

Safira Nury Safitri[1], Haryono Setiadi[2], Esti Suryani[3]
[1,2,3]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret
[1]safira.safitri04@student.uns.ac.id, [2]hsd@staff.uns.ac.id*, [3]estisuryani@staff.uns.ac.id

Abstract

Higher education institutions store data keep growing every year. The data has important information, but it still not optimized into knowledge. Data Mining (DM) can be used to process existing data in universities in order to obtain knowledge that can be utilized further. Educational Data Mining (EDM) often appears to be applied in big data processing in the education sector. One of the educational data that can be further processed with EDM is activity log data from an e-learning system used in teaching and learning activities. The log activity can be further processed more specifically by using log mining. The purpose of this study was to process log data from the Sebelas Maret University Online Learning System (SPADA UNS) to determine student learning behavior patterns and their relationship to the final results obtained. The data mining method applied in this research is cluster analysis with the K-means Clustering and Decision Tree algorithms. The clustering process is used to find groups of students who have similar learning patterns. While the decision tree is used to model the results of the clustering in order to enable the analysis and decision-making processes. Processing of 11,139 SPADA UNS log data resulted in 3 clusters with a Davies Bouldin Index (DBI) value of 0.229. The results of these three clusters are modeled by using a Decision Tree. The decision tree model in cluster 0 represents a group of students who have a low tendency of learning behavior patterns with the highest frequency of access to course viewing activities obtained accuracy of 74.42% . In cluster 1, which contains groups of students with high learning behavior patterns, have a high frequency of access to viewing discussion activities obtained accuracy of 76.47%. While cluster 2 is a group of students who have a pattern of learning behavior that is having a high frequency of access to the activity of sending assignments obtained accuracy of 90.00%.

Keywords: Analysis Cluster, Decision Tree, Educational Data Mining (EDM), Log, SPADA

1. Introduction

Currently, higher education institutions have millions of data that can be used for decision-making. Unfortunately, these data are still not used as knowledge. Further data processing techniques are needed in order to obtain important information that can assist the academic community in understanding the educational data. This understanding is important for evaluating and improving the quality of future teaching and learning activities[1]. A data processing technique that can be used to do this is data mining. Data mining can extract important information or important patterns in big data [2]. The type of data mining technique that appears in the execution of data processing in the field of education is Educational Data Mining (EDM)[3][4]. One of the data in education that can be used is logs from an online learning system (e-learning).

A log is a file extension that is created to automatically store information on execution or user command from certain software or operating systems in the form of semi-structured data or event text[5][6]. Log documentation can show a user's access history on a day, time, and heterogeneous context with the volume and scale of complexity increasing every year [7][8]. There are various kinds of log data analysis automation techniques, including using log mining. This log mining technique can be applied to perform data processing from an online learning system (e-learning).

The e-learning system provides various kinds of user log data, especially student history in a course that is ready to be further processed with log mining. In this era, many e-learning systems are packaged in a web-based Moodle Learning Management System (LMS) as learning support [9]. LMS Moodle has a feature that allows users to modify the appearance of e-learning as needed, such as providing material, developing assignments, and monitoring student learning processes [10]. Sebelas Maret University (UNS) is one of the

Accepted: 12-03-2022 | Received in revised: 16-06-2022 | Published: 30-06-2022

higher education institutions that has implemented an e-learning system based on the Moodle LMS in its learning process with the Online Learning System (SPADA). The Online Learning System (SPADA) is a program of the Directorate General and Student Affairs of the Ministry of Research, Technology and Higher Education (Kemenristekdikti) with the aim of increasing student access to high-quality education by using an online learning system or e-learning system[11]. The log data in the LMS Moodle SPADA Sebelas Maret University is one example of a hidden knowledge data source in the field of higher education that can be processed using the Educational Data Mining (EDM) technique. The implementation of EDM techniques can be categorized according to the target user of the processed data. The target that can be determined is one of the stakeholders involved, such as students, teacher courses, admin courses, and researchers themselves [12]. The EDM technique that will be applied to the SPADA UNS log data will focus on finding patterns of student learning behavior which will then be evaluated for their relationship to the final score obtained.

Research on the Educational Data Mining (EDM) technique has been carried out by previous researchers in analyzing students' Self Reported Learning (SLR) patterns using Self Reported Surveys and log data using the K-means algorithm to perform cluster analysis [13]. This study results in Self Reported Surveys data having poor performance in predicting student achievement. While the log data that is processed with data mining techniques is effectively able to do the same thing with better performance, in the calculation of many variables the log data at n = 17 and the Self Reported Survey at n = 4 shows that both are still strong in making predictions. Self Reported Learning pattern in e-learning. This study states that data mining techniques in education (EDM) are effective for developing opportunities for further processing of educational data in identifying online learning patterns.

The Educational Data Mining (EDM) technique for processing log data has also been carried out by other researchers who compared the K-means, Hierarchical, and Louvain clustering algorithms [14]. This study shows that the Louvain algorithm has a better performance when compared to the Hierarchical and K-means clustering algorithms. Louvain clustering is able to detect groups of data accurately which cannot be done by the K-means algorithm. Hierarchical clustering shows an interesting performance where this algorithm can display anomalies in the data (outliers) that are used to create a grouping. This study also explains that the K-means algorithm can get a high silhouette coefficient value in several groups that have a number of adjacent data.

In another similar study, an analysis of the behavior of the student learning process was carried out in improving the online learning system by using the K-means clustering algorithm for cluster analysis and using a decision tree [5]. Data processing with the applied method found that online learning instructors can more easily identify learning content by referring to the student learning behavior patterns that have been obtained. The use of decision trees is able to represent decision-making that displays the identification of groups of objects for further analysis of student learning patterns.

Further related research was carried out by Ieannoal Vhallah, et al, were in this study used the K-means Clustering method to group potential dropout students. The stage of forming the data into clusters is done by calculating the Euclidean Distance which produces a number of 3 clusters in each batch of students. The results of the implementation of the K-means Clustering method with attributes of total semester credit units (SKS), Grade Point Average (GPA), and semesters that have been taken can show potential students dropping out early so that educators can immediately take directive action to improve academic achievement students to avoid dropping out [15].

Based on related research that has been described previously, the cluster analysis method with the K-means clustering algorithm and decision tree is the right method to be applied to this research problem. The clustering analysis process is tasked with grouping data into clusters based on indicators that are focused or observed so that the data entered into groups have a high similarity compared to data in different groups[16]. The K-means clustering algorithm which is a type of distance-based clustering (Euclidean Distance) has the advantage of fast-paced big data clustering [17]. While the decision tree algorithm, which is one type of classification algorithm, has the advantage of being able to represent decisions that are easy to analyze and efficient in handling discrete and numerical attributes [18]. In this study, the author will apply the cluster analysis method to map student learning behavior patterns in the SPADA UNS online learning system (e-learning) and apply a decision tree to represent the relationship between student learning behavior patterns and the final results obtained by students.

2. Research Methods

This research was conducted in six stages consisting of dataset collection, data cleaning, data partitioning, data clustering, decision tree analyzing, and interpretation of the results. The stages of this research are represented in Figure 1.

DOI: https://doi.org/10.29207/resti.v6i3.3935

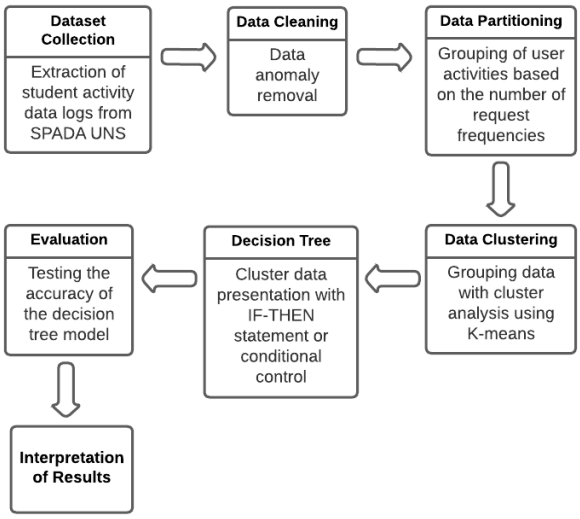

Figure 1. The Flow of Research Methods

2.1 Dataset Collection

The dataset collected from the university's e-learning system is the Sebelas Maret University Online Learning System (SPADA UNS). The UNS SPADA log data is managed by the Information and Communication Technology Technical Implementation Unit (UPT TIK) Sebelas Maret University.

The data taken is a course activity log containing interactions between course instructors and students such as discussions, quizzes, task collection, and the provision of material in one semester.

2.2 Data Cleaning

Data cleaning is done to improve the quality of the data to be processed. The data cleaning carried out in this study consisted of deleting duplicate data, deleting teacher course activities to focus more on student activities, deleting data anomalies or abnormal data, deleting activities that were not relevant in describing student learning patterns, and anonymous data was carried out to maintain the privacy of the lecturer, course name and the student concerned.

2.3 Data Partitioning

Data partitioning is a process for distributing data in several tables with the aim of improving the data query process. The data partitioning process is done by separating the main table into smaller individual tables. This is done so that the query process runs faster because the query only accesses a small part of the data to be scanned. In this data partitioning, student activities will be combined in one table according to the frequency with which the user performs a requested task.

2.4 Data Clustering

Data clustering or data grouping is a technique for dividing datasets into separate groups of sets called clusters[19]. Clustering is done to find groups of students who have similar learning behavior patterns. The similarity of the group of students is based on student behavior in the e-learning system which is accessed by students for one semester on the same course with the same teacher course. This data grouping is done by applying the cluster analysis method with the K-means clustering algorithm. The K-means clustering algorithm maps the log data into unsupervised clusters. Before doing the grouping, the data is normalized first using the Z transformation. Data normalization is one of the stages of forming the data attribute value scale to create the same weight value for each feature[20]. Data normalization is done to handle data sets that have an unequal range or range values. In addition, data normalization can also improve the data quality and performance of the applied algorithm. After normalization, it is continued with the implementation of the K-means Clustering algorithm. The K-means clustering algorithm stage begins with analyzing the need for k centroid points (midpoints) and calculating the distance of each data point to the centroid. These calculations can be done using the Euclidean Distance formula in equation (1).

$$d(a,b) = \sqrt{(x-a)^2 + (y-b)^2} \qquad (1)$$

Where x and y are the coordinates of the data object. While a and b are the coordinates of the centroid (midpoint). After obtaining the distance of each data point to the centroid, the new centroid value (midpoint) is calculated using equation (2).

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \qquad (2)$$

Where N_i is the amount of data in the i cluster, x_{kj} is the value of the x data in the cluster for the j variable, i and k are clusters and j is a variable.

The calculation of the new centroid point will stop if the calculated result has no difference from the new centroid value calculated earlier. If there is still a difference, then the calculation will continue from the Euclidean Distance calculation until the new centroid value is no longer changing [21].

If the appropriate new centroid results have been obtained, then the number of student groups formed from the K-means clustering algorithm needs to be tested in order to obtain a quality cluster group. One technique that can be used to test the accuracy of the number of clusters formed is the Davies Bouldin Index (DBI). This technique is able to find out how accurately an object is categorized in a cluster. The Davies Bouldin Index (DBI) is a combination of the separation and cohesion methods where this method separately functions in calculating how far the center point is from one cluster to another and cohesively functions as a measure of how close the relationship between objects in a cluster is. The Davies Bouldin Index (DBI)

DOI: https://doi.org/10.29207/resti.v6i3.3935

calculation is done by calculating the Sum of squares within the Cluster (SSW) of a cluster to determine its cohesion. SSW calculation is in accordance with equation (3) below:

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{mi} d(x_j, c_i) \quad (3)$$

Where m_i is the number of data in the cluster, $d(x_j, c_i)$ is the distance of the j data with the i cluster centroid. The calculation of the Davies Bouldin Index (DBI) value can be done by the formula (4).

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad (4)$$

Where $(R_{i,j})$ or the ratio of the i cluster and j cluster is calculated using equation (5).

$$R_{i,j} = \frac{SSW_i + SSW_j}{d(x_j, c_i)} \quad (5)$$

The smaller the value of the Davies Bouldin Index cluster, the more clearly the data partition or differences in data groups from one another [22]. This means that the better the quality of the number of clusters formed. Conversely, if the value of the Davies Bouldin Index (DBI) is smaller, the level of similarity of group data is getting closer.

2.5 Decision Tree

The data that has been grouped will be analyzed for the formation of a more structured data model using the decision tree method. This method classifies the data described by nodes in the form of a hierarchical tree [23]. The use of this method in data modeling aims so that each group object can be identified more specifically to facilitate decision-making. Decision tree modeling begins by determining the root or root attribute tree to get the appropriate decision tree size. Attribute branches in the decision tree are declared pure if they come from the same class. The size of the purity of the branch can be calculated by the entropy formula in equation (6) below:

$$E(S_A) = \sum_{i=1}^{n} - p_i \, log_2(p_i) \quad (6)$$

Where A is an attribute, n is the number of partitions S and pi is the proportion of Si to S. The higher the branch purity value, the better the quality. The impurity criteria of a branch need to be tested by calculating the information gained to find out whether the selected attribute is correct or not. The selection of the right attributes can produce the greatest information gain. The information gain value can be determined by using the information gain calculation in equation (7).

$$G(S, A) = E(S_A) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * E(Si) \quad (7)$$

Where n is the number of attribute partitions A, $|Si|$ is the number of cases in the partition I, and $|S|$ is the number of cases in S. The calculation process will continue to repeat itself starting from the calculation of entropy and information gain until all branch nodes get the same class.

2.6 Evaluation

The evaluation stage aims to test the performance of the decision tree model. Performance testing at the evaluation stage is an analytical step of the level of accuracy of the system that has been built [24]. Evaluation is carried out using a confusion matrix consisting of a column of prediction class (Pred) and actual class (True)[25]. The base class that composes the confusion matrix consists of Yes and No. The confusion matrix can be seen in Table 1.

Table 1. Confusion Matrix

	Actual Yes	Actual No
Predicted Yes	True Positive (TP)	False Positive (FP)
Predicted No	False Negative (FN)	True Positive (TN)

Where True Positive (TP) is a positive class that is correctly predicted as a positive class or correct result, False Positive (FP) is a negative class that is incorrectly predicted as a positive class or unexpected result, and True Negative (TN) is a negative class and is correctly predicted as a class. negative, False Negative (FN) is a positive class that is wrongly predicted as a negative class [26][27].

Based on the confusion matrix table, it can be evaluated from the decision tree model by calculating the values of accuracy, recall, and precision. Accuracy is the percentage of closeness between the predicted value and the actual value [28]. The calculation of accuracy can be seen in equation (8). The recall is the actual positive class that is predicted to be correct compared to the number of correct positive classes in the overall data [29]. Recall can be calculated according to equation (9). Precision is the level of prediction accuracy with the actual class [30]. The formula for precision can be seen in equation (10). The higher the value of accuracy, recall, and precision, the better the classification performance of the model [31].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FN} \quad (10)$$

2.7 Interpretation of Results

Result analysis contains the presentation of data, to obtain useful information and is composed of logical and important facts. Submission of the analysis of the results is carried out using language that is easily understood by the reader, in other words, the submission is carried out no longer using statistical language.

DOI: https://doi.org/10.29207/resti.v6i3.3935

3. Results and Discussions

3.1 Results of Dataset Collection

The dataset collected is in the form of log data on an e-learning system in the Sebelas Maret University Online Learning System (SPADA UNS) for the 2020/2021 academic year starting from February 2021 to July 2021 (one semester). Obtained as many as 11,139 raw data logs of teaching and learning activities (KBM) with the same teacher course. The selection of courses is based on the number of interactions between the teacher course and students in the form of providing material, assignments, discussions, quizzes, mid-semester exams (UTS), and final semester exams (UAS). The dataset obtained is as shown in Table 2.

Table 2. Log Data of SPADA UNS

Time	Userid	...	Action	Target	…
21 Feb, 15:46	15094	…	created	course	…
…	…	…	…	...	…
29 Mar, 09:09	5411	…	assigned	role	…
29 Mar, 09:09	5411	…	viewed	course	…
29 Mar, 09:11	15094	…	viewed	course	…
…	…	…	…	…	…
07 Jul, 09:46	5407	…	viewed	course	…

Table 2 shows the SPADA UNS log data obtained from UPT TIK UNS. The collected log data contains much information in the form of event timestamps from user request, userid, course name, user full name, affected user, component, eventname, action, target, event context, origin, and IP address. This means that every activity carried out by the user is recorded automatically in a semi-structured form as shown in Table 2.

User activities that include interactions between the teacher course and students form the dimensions of the SPADA UNS feature can be seen in Table 3.

Table 3. User Activities based on SPADA UNS

Dimension	Meaning	Description
Uploaded Assessable	Sending Assignments	Assignments have been sent and can be graded by the teacher's course
Viewed Course	Viewed Course	The instance of the user accessing the course to read or download the course file
Viewed Discussion	Viewed Discussion	View discussion content that has been posted on the discussion forum
Created Post	Sending Discussion	Send a discussion with the default text content. Text style can be changed and added with image, sound, video attachments, and so on.
…	…	…
Created Submission	Sending Quiz	Allows users to sending a quiz

Table 3 shows the dimensions of the features formed in the interaction of teaching and learning activities (KBM) between the teacher course and students. The collection of student activities at the KBM is then carried out further processes to determine student learning patterns in the SPADA UNS e-learning system.

3.2 Data Cleaning Results

The data cleaning process begins with the search for duplicate data, which found as many as 2,213 rows of duplicate data from a total of 11,139 data. The duplication in question is a row of data that has the exact same contents in all of its column components. After deleting duplicate data, the next step is to delete the entire teacher course activity. In the collected dataset, 579 lines of deleted teacher course activity data were found.

The next data cleaning process is the removal of data anomalies or abnormal data entries. A total of 68 rows of anomalous data were found which were deleted and then continued with the selection of activities that reflect student learning behavior such as sending assignments, viewing quizzes, sending quizzes, viewing discussions, and sending discussions as well as anonymous data processing. This anonymous process is represented by the userid of each student instead of the label of each student. The datasets that have gone through the data cleaning process are 6,989 which are ready to be processed to the next stage.

3.3 Data Partitioning Results

All student activities that have gone through the data cleaning process are then partitioned by extracting and combining data based on the frequency of dimensions accessed by students. The partitioned data to see the pattern of student learning behavior are as shown in Table 4.

Table 4. Results of Data Partitioning

Userid	Sending Discussion	…	Viewed Course	Viewed Discussion
5411	1	…	43	13
5410	1	…	54	30
5413	3	…	52	10
…	…	…	…	…
	2		20	13

Based on Table 4, the results of the user activity partition are in accordance with the dimensions of the requested and the number of access frequencies. One example is user id 5411 carrying out the activity of sending 1 discussion, viewing the course 43 times, and viewing the discussion 13 times.

3.4 Results of Clustering Data

Data that has gone through the data partitioning process is then carried out in the data clustering stage or data grouping. Student activity data are grouped according to the similarity of their learning behavior patterns with userid as the label of the student index variable. Before applying the K-means clustering algorithm, the data was normalized using Z transformation. The grouping of log data with the K-means clustering algorithm used

DOI: https://doi.org/10.29207/resti.v6i3.3935

Euclidean Distance numerical measurements. In testing the validity level of the number of clusters formed, the Davies Bouldin Index (DBI) calculation is applied. In this case, 6 experiments were carried out for various possible numbers of clusters starting from iterations k=2 to k=7. The iteration calculation can be seen in Table 5.

Table 5. Calculation of Davies Bouldin Index (DBI)

Experiment	Iteration (k)	DBI Value
1	2	0.262
2	3	0.229
3	4	0.244
4	5	0.266
5	6	0.234
6	7	0.244

In Table 5, the results of the Davies Bouldin Index (DBI) calculation above show that the k=3 iteration test resulted in a good DBI value of 0.229. This means that the cluster with the number of k = 3 has a high level of validity because the value is the smallest or close to 0. The results of the clustering formed can be seen in Figure 2.

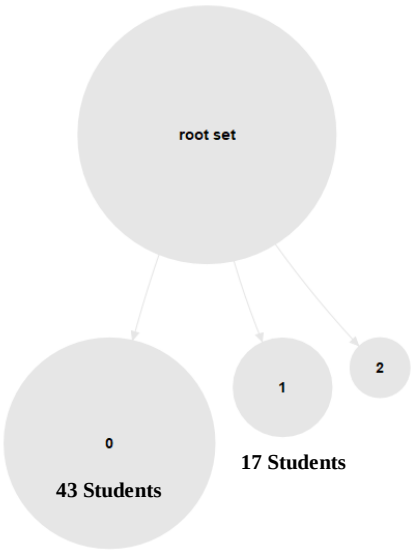

Figure 2. Clustering Data

Figure 2 is a presentation of the cluster formed in the form of a root and the results of the cluster. The implementation of the K-means clustering method on 70 students' data with a value of k=3 resulted in 3 clusters consisting cluster 0 containing 43 students, cluster 1 containing 17 students, and cluster 2 containing 10 students.

Figure 3 shows the dimensional distribution of each cluster. Cluster 0 represents student learning patterns with access to SPADA UNS activities which tend to be low on all dimensions. Cluster 1 contains students who have high learning patterns. Meanwhile, cluster 2 is students with moderate access to learning patterns. Details of the distribution of the dimensions of each cluster can be seen in Table 6.

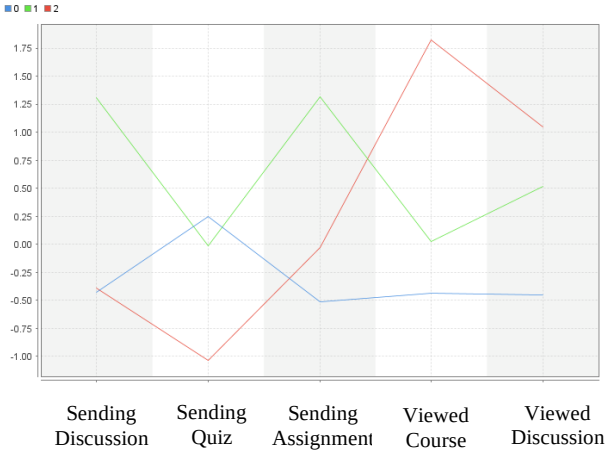

Figure 3. Dimensional Distribution of Each Cluster

Table 6. Centroid

Dimensional Spread	Cluster 0	Cluster 1	Cluster 2
Sending Discussion	-0.428	1.313	-0.392
Sending Quiz	0.246	-0.015	-1.034
Sending Assignments	-0.514	1.317	-0.027
View Course	-0.435	0.027	1.825
View Discussion	-0.448	0.517	1.047

Table 6 of the centroids (midpoint) shows that cluster 0 which consists of 43 students represents a low learning pattern in the activities of sending discussions, sending assignments, viewing courses, and viewing discussions. This cluster shows high activity in the activity of sending quizzes. Cluster 1 with 17 students has high access to the activities of sending discussions and sending assignments. Meanwhile, cluster 2 which contains 10 students has the highest access to the activity of viewing courses and viewing discussions. Cluster 2 has the lowest access to the activity of sending quizzes when compared to the other two clusters.

3.5 Decision Tree Interpretation

After going through cluster analysis using the K-means clustering method, each cluster group made a decision tree model. Figure 4 is a decision tree that represents student learning behavior patterns in cluster 0, Figure 5 represents a decision tree of learning behavior patterns in cluster 1, and Figure 6 represents a decision tree that represents student learning behavior patterns in cluster 2. The decision tree model in cluster 0 can be seen in Figure 4.

Based on Figure 4, the decision tree model for cluster 0 shows the highest frequency of access to course viewing activities. Students who often access activities to view the course and often see discussions produce high scores. A similar pattern from cluster 0 is that students who combine high access frequencies in viewing courses, viewing discussions, sending assignments, and sending discussions also get high scores.

DOI: https://doi.org/10.29207/resti.v6i3.3935

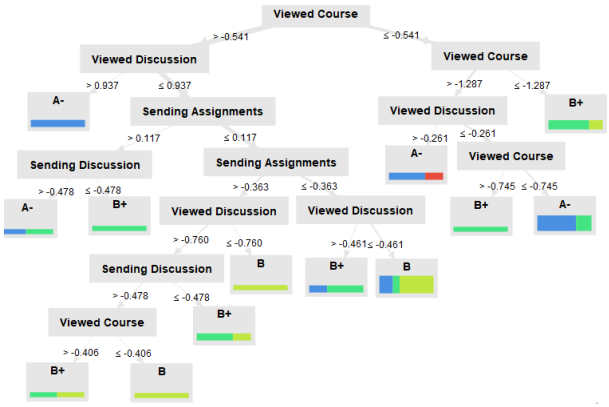

Figure 4. Decision Tree Cluster 0

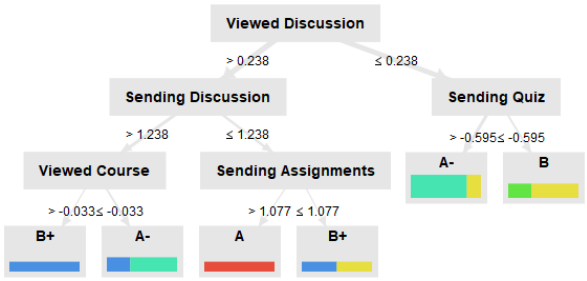

Figure 5. Decision Tree Cluster 1

According to Figure 5, which represents the decision tree model, cluster 1 shows the highest frequency accessed by viewing discussions. The combination of a high frequency of access to view discussions, low access to sending discussions, and high access to sending assignments resulted in the highest or maximum final score. Meanwhile, the low frequency of access to view discussions and send quizzes resulted in less than optimal or low scores.

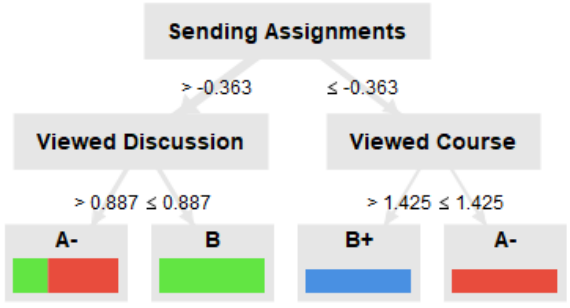

Figure 6. Decision Tree Cluster 2

The last decision tree model that represents cluster 2 in Figure 6, it has the highest frequency of sending tasks. High frequency of sending assignments and viewing discussions resulted in high final scores. However, students in this group with a combination of submitting assignments and viewing low-frequency courses also produced high final scores.

3.6 Evaluation Result

The evaluation stage carried out in this study was calculated using the accuracy, recall, and precision parameters from the confusion matrix table described previously. Based on these calculations, it produces an accuracy value of 74.42% in the decision tree cluster model 0. Details of the results of the recall and precision calculations along with the number of predictions (pred) and actual (true) from the decision tree cluster 0 models can be seen in Table 7.

Table 7. Evaluation of Decision Tree Cluster Model 0

	True A-	True B+	True B	True C+	Class Precision
Pred. A-	11	3	0	1	73%
Pred. B+	1	12	3	0	75%
Pred. B	2	1	9	0	75%
Pred C+	0	0	0	0	0%
Class Recall	79%	75%	75%	0%	

Table 7 shows that there are 11 records that are predicted to be correct in class A- with a recall value of 79% and a precision value of 73%. Class B+ predicted correctly as many as 12 records with recall and precision of 75%. A total of 9 records were predicted to be correct in class B with 0% recall and 75% precision. In class C+ there are 0 records that are predicted to be correct with a recall and precision value of 0%. In contrast to the decision tree cluster model 0, the accuracy resulting from the calculation of the decision tree cluster 1 model is 76.47% with details in Table 8.

Table 8. Evaluation of the Decision Tree Cluster 1 Model

	True A	True A-	True B+	True B	True C+	Class Precision
Pred. A	2	0	0	0	0	100%
Pred. A-	0	6	1	1	0	75%
Pred. B+	0	0	3	1	0	75%
Pred B	0	0	0	2	1	67%
Pred C+	0	0	0	0	0	100%
Class Recall	100 %	100 %	75%	50%	0%	

Table 8 shows that as many as 2 records were predicted to be correct in class A with a recall percentage and a precision of 100%. The number that is predicted to be correct in class A- is 6 records with 100% recall and 75% precision. There are 3 records that are predicted to be correct in class B+ with a recall value and a precision of 75%. A total of 2 records were predicted to be correct in class B with 50% recall and 67% precision. Class C+ contains 0 records that are predicted to be correct with a recall value of 0% and a precision of 100%. As for the decision tree cluster 2 model, the accuracy value reaches 90.00% with details in Table 9.

Table 9 details as many as 4 records that are predicted to be correct in class A- with a recall value of 100% and a precision of 80%. The predicted correct records in class B+ are 2 with recall and precision values of 100%.

DOI: https://doi.org/10.29207/resti.v6i3.3935

A total of 3 records were predicted to be correct for class B with a recall value of 75% and a precision of 100%.

Table 9. Evaluation of the Decision Tree Cluster 2 Model

	True A-	True B+	True B	Class Precision
Pred. A-	4	0	1	80%
Pred. B+	0	2	0	100%
Pred. B	0	0	3	100%
Class Recall	100%	100%	75%	

3.7 Results Interpretation

Based on the whole process of processing 11,139 raw data from the activity log of the Sebelas Maret University Online Learning System (SPADA UNS) the data cleaning process produced 6,989 data which was then pivoted on the data partitioning process. Then the results of the data partitioning are processed with the K-means clustering algorithm which produces 3 clusters that are modeled on each decision tree. Cluster 0 which represents a group of students who have a low tendency of learning behavior patterns shows the highest frequency of access to course viewing activities. In cluster 1, which contains groups of students with high learning behavior patterns, they have a high frequency of access to viewing discussion activities. While cluster 2 is a group of students who have a pattern of learning behavior that is having a high frequency of access to the activity of sending assignments.

4. Conclusion

In this study, it can be concluded that Data Mining (DM) can be used to explore information about the unique patterns of a number of big data. The use of data mining for big data processing is very popular because of the increasing accumulation of stored data, especially in higher education institutions with thousands of students. Therefore, Educational Data Mining (EDM) is used, which is part of Data Mining in the education sector. The implementation of cluster analysis using the K-means clustering algorithm can show the learning patterns of student groups formed based on access carried out in the Sebelas Maret University Online Learning System (SPADA UNS) for one semester with the same teacher course.

The data clustering process produces 3 clusters with a Davies Bouldin Index (DBI) value of 0.229. Each cluster performed data modeling using the decision tree method to facilitate the process of further analysis of student behavior patterns in the teaching and learning process related to the final score obtained. Cluster 0 consists of 43 students representing the low learning pattern group having the highest frequency of access to course viewing activities with the decision tree model accuracy value of 74.42%. Cluster 1 with 17 students representing the high learning pattern group has a high frequency of access to viewing discussion activities and has an accuracy decision tree value of 76.47%. While cluster 2 which consists of 10 students representing moderate learning patterns has a high frequency of access to the activity of sending assignments with an accuracy decision tree reaching 90.00%.

Reference

[1] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in edx MOOCs," *Proc. - 2016 IEEE 2nd Int. Conf. Multimed. Big Data, BigMM 2016*, pp. 440–443, 2016, doi: 10.1109/BigMM.2016.70.

[2] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[3] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Educ.*, vol. 51, no. 1, pp. 368–384, 2008, doi: 10.1016/j.compedu.2007.05.016.

[4] C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, "Computers & Education Data mining in educational technology classroom research : Can it make a contribution ?," *Comput. Educ.*, vol. 113, pp. 226–242, 2017, doi: 10.1016/j.compedu.2017.05.021.

[5] S. Križanić, "Educational data mining using cluster analysis and decision tree technique: A case study," *Int. J. Eng. Bus. Manag.*, vol. 12, pp. 1–9, 2020, doi: 10.1177/1847979020908675.

[6] M. Pettinato, J. P. Gil, P. Galeas, and B. Russo, "Log mining to re-construct system behavior: An exploratory study on a large telescope system," *Inf. Softw. Technol.*, vol. 114, no. May, pp. 121–136, 2019, doi: 10.1016/j.infsof.2019.06.011.

[7] T. Lerche and E. Kiel, "Predicting student achievement in learning management systems by log data analysis," *Comput. Human Behav.*, vol. 89, pp. 367–372, 2018, doi: 10.1016/j.chb.2018.06.015.

[8] M. Hussain, M. A. Sujith, and M. Abdullah, "Mining Educational Data for Academic Accreditation : Aligning Assessment with Outcomes," *Glob. J. Flex. Syst. Manag.*, 2016, doi: 10.1007/s40171-016-0143-3.

[9] Y. Park and I. Jo, "Assessment & Evaluation in Higher Education Using log variables in a learning management system to evaluate learning activity using the lens of activity theory," vol. 2938, no. April, pp. 0–17, 2016, doi: 10.1080/02602938.2016.1158236.

[10] N. Kerimbayev, N. Nurym, A. Akramova, and S. Abdykarimova, "Virtual educational environment: interactive communication using LMS Moodle," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1965–1982, 2020, doi: 10.1007/s10639-019-10067-5.

[11] U. Anis Chaeruman, B. Wibawa, and Z. Syahrial, "Determining the Appropriate Blend of Blended Learning: A Formative Research in the Context of Spada-Indonesia," *Am. J. Educ. Res.*, vol. 6, no. 3, pp. 188–195, 2018, doi: 10.12691/education-6-3-5.

[12] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Educ. Inf. Technol.*, vol. 23, no. 1, pp. 537–553, 2018, doi: 10.1007/s10639-017-9616-z.

[13] M. H. Cho and J. S. Yoo, "Exploring online students' self-regulated learning with self-reported surveys and log files: a data mining approach," *Interact. Learn. Environ.*, vol. 25, no. 8, pp. 970–982, 2017, doi: 10.1080/10494820.2016.1232278.

[14] C. Pradana, S. S. Kusumawardani, and A. E. Permanasari, "Comparison Clustering Performance Based on Moodle Log Mining," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 722, no. 1, 2020, doi: 10.1088/1757-899X/722/1/012012.

[15] I. Vhallah, S. Sumijan, J. Santony, and others, "Pengelompokan mahasiswa potensial drop out menggunakan metode Clustering K-Means," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 572–577, 2018.

[16] R. Ananda, A. Z. Yamani, and others, "Determination of Initial

DOI: https://doi.org/10.29207/resti.v6i3.3935

K-means Centroid in the Process of Clustering Data Evaluation of Teaching Lecturers," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 4, no. 3, pp. 544–550, 2020.

[17] T. Susilowati, D. Sugiarto, I. Mardianto, and others, "Validity Test of Self-Organizing Map (SOM) and K-Means Algorithm for Employee Grouping," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 4, no. 6, pp. 1171–1178, 2020.

[18] I. Romli, F. Kharida, and C. Naya, "Penentuan Kepuasan Pelanggan Terhadap Pelayanan Kantor Pelayanan Pajak Menggunakan C4. 5 dan PSO," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 296–302, 2020.

[19] M. Capó, A. Pérez, and J. A. Lozano, "Knowle dge-Base d Systems An efficient approximation to the K -means clustering for massive data," vol. 0, pp. 1–14, 2016, doi: 10.1016/j.knosys.2016.06.031.

[20] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput. J.*, vol. 97, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[21] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, "Two improved k-means algorithms," *Appl. Soft Comput. J.*, vol. 68, pp. 747–755, 2018, doi: 10.1016/j.asoc.2017.08.032.

[22] R. Ünlü and P. Xanthopoulos, "Estimating the number of clusters in a dataset via consensus clustering," *Expert Syst. Appl.*, vol. 125, pp. 33–39, 2019, doi: 10.1016/j.eswa.2019.01.074.

[23] D. C. Wickramarachchi, B. L. Robertson, M. Reale, C. J. Price, and J. Brown, "HHCART: An oblique decision tree," *Comput. Stat. Data Anal.*, vol. 96, pp. 12–23, 2016, doi: 10.1016/j.csda.2015.11.006.

[24] B. Irena, E. B. Setiawan, and others, "Fake news (hoax) identification on social media twitter using decision tree c4. 5 method," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 4, no. 4, pp. 711–716, 2020.

[25] A. Souri, M. Yassin, G. Aram, M. Ahmed, and F. Safara, "A new machine learning-based healthcare monitoring model for student ' s condition diagnosis in Internet of Things environment," *Soft Comput.*, vol. 6, 2020, doi: 10.1007/s00500-020-05003-6.

[26] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, and D. Tien, "Catena A comparative study of logistic model tree , random forest , and classi fi cation and regression tree models for spatial prediction of landslide susceptibility," *Catena*, vol. 151, pp. 147–160, 2017, doi: 10.1016/j.catena.2016.11.032.

[27] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Science of the Total Environment Performance evaluation of the GIS-based data mining techniques of best- fi rst decision tree , random forest , and naïve Bayes tree for landslide susceptibility modeling," *Sci. Total Environ.*, vol. 644, pp. 1006–1018, 2018, doi: 10.1016/j.scitotenv.2018.06.389.

[28] D. Delen, C. Kuzey, and A. Uyar, "Expert Systems with Applications Measuring firm performance using financial ratios : A decision tree approach," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 3970–3983, 2013, doi: 10.1016/j.eswa.2013.01.012.

[29] M. Hasan, M. Islam, I. I. Zarif, and M. M. A. Hashem, "Internet of Things Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches," *Internet of Things*, vol. 7, p. 100059, 2019, doi: 10.1016/j.iot.2019.100059.

[30] T. Reddy, G. Neelu, K. Sweta, and B. Saurabh, "Deep neural networks to predict diabetic retinopathy," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2020, doi: 10.1007/s12652-020-01963-7.

[31] Y. Yang, "The Evaluation of Online Education Course Performance Using," vol. 2021, 2021.

DOI: https://doi.org/10.29207/resti.v6i3.3935