ANALYSIS OF MALANG UNIVERSITY STUDENT ACHIEVEMENT GROUPING USING THE K-MEANS CLUSTERING METHOD

[1*]MOH. AQIL MUKHTAR ALFARERA, [2]ZAEHOL FATAH

Faculty of Science and Technology, Information Technology, Universitas Ibrahimy

Jl. KHR. Syamsul Arifin No. 1-2, Situbondo 68374, East Java

e-mail: [1]aqilfarel86@gmail.com, [2]zaehofatah@gmail.com

[*]Corresponding author

ABSTRACT

With the increasing number of students and variations in achievement, managing achievement data in higher education has become more complex, so manual methods are insufficient. K-means clustering was chosen because of its ability to group data based on specific attributes, which makes it easier to identify patterns and trends. This research aims to prove K-Means' effectiveness in analyzing achievement data and adding to the literature regarding the application of data mining in education. The dataset includes student achievement indexes from various study programs at the University of Malang from 2018 to 2022. The data is processed to group student achievements efficiently. The clustering model was built using one of the algorithms in the clustering method, namely K-Means. This research produced the best cluster with a total of 3 clusters. The process was conducted to determine the best grouping by testing six cluster models. The best cluster was selected using the Davies Bouldin index test. Based on research with the results, these three groups can be categorized as cluster 0 in the low category with a value of 100, cluster 1 in the high category with a value of 4.100, and cluster 2 in the middle category with a value of 1.900.

Keywords: *Data analysis, K-Means, Clustering, Student achievement, University of Malang*

1. INTRODUCTION

Data management and analysis are critical in the current information era, especially in higher education. Malang University (UM), as one of the leading educational institutions in Indonesia, has the responsibility to manage education and the learning process in higher education by focusing on students, using efficient learning methods, and maximizing the use of technology. A critical aspect of this achievement is student achievement, which includes academic results and student achievement motivation while is influenced by various internal and external factors [1]. Predictions of student achievement can provide an overview of the final results likely to be obtained in academic achievement. Apart from that, this prediction also functions as an early warning for students so they can take corrective steps to become quality graduates with academic achievements that meet the expected standards [2].

Student achievement results from a learning process completed based on individual potential or ability and by the competencies determined by the relevant educational institution. Students' success in achieving learning is reflected in their cumulative achievement index, which is classified into five categories based on the highest potential score for each predetermined indicator [3]. At Malang University (UM), student achievement data continues to increase every year; this data includes academic and non-academic achievements and individual and group achievements. Universities have made various efforts to improve student academic achievement to achieve national education standards. The results of this data processing are beneficial in the decision-making process for future planning and development [4].

In the context of grouping student achievements at the University of Malang, there are challenges in identifying and grouping students based on achievement levels objectively and systematically. Student achievements can be seen from various aspects, such as GPA (Cumulative Achievement Index), number of extracurricular activities participated in, achievements in competitions or competitions, as well as contributions to research or community service. However, this achievement grouping is often done subjectively and is not well structured, resulting in difficulties in providing

DOI: https://doi.org/10.54732/jeecs.v9i2.8

rankings or recommendations for students. This problem requires testing that involves a calculation system that applies data mining concepts to analyze and group student achievements in an objective and structured manner [5]. The K-Means Clustering method is a method in data mining that is popularly known for its speed in grouping data based on similar attributes. This method is able to identify patterns and trends in data that can be used for decision making and aims to divide objects into one or more groups based on similar characteristics which will be grouped in one cluster while objects with different characteristics will be in another cluster. In addition, K-Means Clustering has weaknesses in handling data with many dimensions, especially when dealing with data that cannot be separated non-linearly [6].

Data mining is defined as a set of techniques used automatically to thoroughly explore and bring to the surface complex relationships in very large data sets [7]. The main goal is to find patterns, relationships, or hidden information in data, which may not be immediately apparent. This process uses various statistical, mathematical and artificial intelligence techniques to analyze data automatically and systematically [8]. Data mining generally processes data from large databases. From this data, a search for patterns or trends is carried out in accordance with the objectives of applying data mining. The results of this process can then be used for decision making and the necessary analysis [9]. The K-Means algorithm is a clustering method that aims to divide data into one or more cluster groups. This method groups data into the same cluster, while data with different characteristics is put into another cluster [10].

Several studies have been conducted to apply the K-Means Clustering approach in analyzing student achievement, which results in grouping students based on achievement. The aim of this research is to identify and form clusters of student achievement data based on academic and non-academic aspects, so that they can be used as a reference for improving student academic achievement in the subsequent learning process [11]. The K-Means Clustering algorithm approach is used to produce a model that can group student achievements from various study programs in higher education [12]. This explains the concept of the K-Means algorithm and illustrates how the algorithm can be applied to evaluate and classify student academic performance [13].

The results of this research show that the K-Means clustering technique can be used to analyze the achievements of Malang University students from various study programs in the 2018 to 2022 period, by grouping students into low, medium, and high achievement categories [14]. Next, the best grouping was tested using the Davies-Bouldin method [15]. Therefore, the process of grouping student achievements is carried out by applying the K-Means Clustering method using the RapidMiner application [16].

2. RESEARCH METHODOLOGY

2.1 Data Source

This research uses secondary data teken from the online platform that provides it information about the achievements of Malang University students. The data source was obtained through the official portal universities and public sources that provide academic and non-academic achievement data students in a structured from. The data collected includes university student achievements Malang in Various areas of achievement in each study program, which have been published by related universities. This data is available online via the *Kaggle* repository. This dataset consists of 70 entries and 5 attributes. Results dataset processing is shown in Table 1.

2.2 Data Collection

This data collection is carried out to obtain and find out information about achievements academic and non-academic students at Malang University (UM). Student achievement data is accessed through the university achievement archive which contains official records of academic and non-academic achievements have been achieved by students. The documents used include annual student achievement reports in each study program.

Table 1. Malang University Student Achievement Dataset

No	Fakultas	Prodi	2018	2019	2020	2021	2022
1	FEB	S1 Pendidikan Ekonomi	0	0	396	36	120
2	FEB	S1 Akuntansi	1	1	1562	473	4152
3	FEB	S1 Pendidikan Akuntansi	0	0	722	285	1903
4	FEB	S1 Manajemen	0	0	707	61	339
5	FEB	S1 Pendidikan Perkantoran	0	0	586	86	549
6	FEB	S1 Pendidikan Tata Niaga	0	0	266	18	78
7	FEB	S1 Ekonomi Pembangunan	0	0	421	24	251
8	FEB	D3 Akuntansi	0	0	69	18	0
9	FEB	D3 Manajemen Pemasaran	0	0	35	7	7
10	FIK	S1 Pendidikan Olahraga	1	0	291	159	37
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
70	FT	D3 Tata Busana	0	0	0	1	0

2.3 Data Processing

In this research, data processing aims to group the achievements of Malang University student based on study programs based on study programs during the 2018 to 2022 period using the K-Means Clustering method. The data processing process includes the following steps:

a. Data Preparation
Data collection: student achievement data is taken from academic and non-academic records during the 2018 to 2020 period.

b. K-Means Clustering Process
Initialization: we determine the number of clusters, for example k = 3, or k = 4, to describe groups of students with high, medium, low, or other achievements.

c. Evaluation of Cluster Results
After the clusters are formed, the quality of the clustering is assessed using the Davies Bouldin index (DBI) method. The best clustering model is selected based on the smallest DBI value from several experiments with different numbers of clusters.

d. Grouping Based on Study Program
After the clusters are formed, analysis is carried out for each study program, identify cluster characteristics for each study program (for example, students with, high, medium, or low academic achievement).

e. Grouping Results
The results of data processing will display student groups based on lavel of achievement, for example cluster 1 includes students with high GPA, cluster 2 includes students with moderate academic achievement, and cluster 3 includes students with low achievement.

The results of data processing using K-Means Clustering and evaluation using the Davies Bouldin index provide an overview of student achievement patterns at the University of Malang based on study program. This analysis helps the university understand the distribution of student achievement and determine strategic steps to improve the quality of education and student development.

2.4 Data Mining

Data mining is a process that automatically searches for useful information in large data storage, data mining techniques are applied to analyze large databases as a method for discovering useful new patterns. But not all search activity information can be categorized as data mining.

Data mining is a technique for analyzing large databases to discover new and useful patterns. This process integrates concepts from various fields, such as:

- Statistics: sampling, estimation, and hypothesis testing.
- Artificial intelligence and machine learning: search algorithms, modeling techniques, and learning theory.
- Other fields: optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval.

Data mining architecture includes components such as databases, knowledge based, data mining engines, pattern evaluation modules, and graphical interfaces. The data mining process follows the CRISP-DM stages: Data Understanding, Data Preparation, Modeling, Evaluation, and Dissemination.

2.5 Clustering

Clustering is the process of dividing a set of data objects into groups called clusters. Objects within a cluster have characteristics that are similar to each other, but different from object within other clusters. This division is carried out automatically using a clustering algorithm, therefore clustering is useful for identifying previously unseen groups or patterns in data. Clustering is also known as data segmentation because it separates data into several groups based on certain similarities.

Clustering is differentiated based on group structure, membership, and data compactness. Based on structure clustering consists of hierarchical and partitioning. Hierarchical clustering is a method of grouping data based on a hierarchical structure. This method is divided into two types: agglomerative and divisive.

- Agglomerative Clustering
  Agglomerative clustering starts with each objects as a separate cluster, then the most similar objects are combined gradually until they form a single cluster.
- Divisive Clustering
  Divisive clustering is the opposite, starting with one large cluster which is then broken down gradually into smaller clusters.

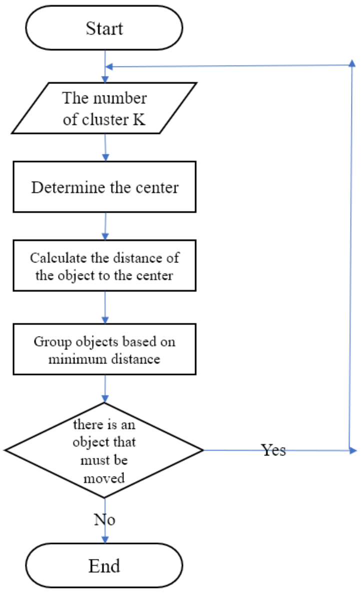

Start
The number of cluster K
Determine the center
Calculate the distance of the object to the center
Group objects based on minimum distance
there is an object that must be moved
Yes
No
End

Figure 1. K-Means flowchart

These two methods organize data based on a proximity matrix, and the results are visualized as a binary tree or dendrogram. The root of the dendrogram represents the entire dataset, while the branches represent data points. The final cluster is obtained by cutting the dendrogram at a certain level.

2.6 K-Means

K-Means is an unsupervised learning algorithm that functions for grouping data into clusters, this algorithm can work with data without category labels. K-Means clustering is where data is grouped based on characteristics that are similar in one cluster, but different from data in other clusters. The flow diagram of the K-Means method can be seen in Figure 1.

From Figure 1, you can see the flow diagram of the K-Means method which starts by determining the desired number of clusters. After the number of clusters is determined, the nex step is to determine the initial position of the cluster center (centroid). The the distance between each data object and each cluster center is calculated using the Euclidean distance formula, to determine the shortest distance between each data point and the existing centroid. Based on this minimum distance, objects are grouped into appropriate clusters. The cluster center is then updated by calculating the everage position of objects in that cluster (the new centroid). If any objects need to be moved to another cluster, this process is repeated. The process continues until there are no more changes in the clustering, indicating that convergence is achieved and the algorithm is complete.

The main formula used in the K-Means algorithm is the Euclidean distance formula, which is used to calculate the distance between data points and centroids. The Euclidean distance between two data points x = (x_1,x_2,…,x_n) and c = (c_1,c_2,…,c_n) is calculated using the formula :

$$d_{Euclidean}(x,c) = \sqrt{\sum_{i=1}^{n}(x_{i-}c_i)^2} \tag{1}$$

where :

- d (x,c) is the distance between data point x and centroid c.
- x_i and c_i are the values in the *i*-th dimension of the data point and centroid.
- n is the number of data dimensions (e.g the number of achievement criteria used).

After determining the distance, each data point (student) will be grouped into the cluster that has the closest distance to the centroid. After assignment of data to clusters, a new centroid is calculated based on the average position of the data in the clusters. For example, for the *k*-th cluster with data members x_1,x_2,…,x_m the new centroid c_k is calculated by the formula :

$$c_k = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2}$$

where :

- c_k is the new centroid of the kth cluster.
- m is the number of data in the kth cluster.
- x_i is the data position in the kth cluster.

2.7 Davies Bouldin

Davies Bouldin is a method for evaluating the value of the results of a clustering algorithm or grouping data, based on the quantity and characteristics contained in the data set. However this approach has weaknesses because the good grades produced do not always reflect quality information obtained.

The Davies Bouldin index (DBI) is a metric for achieving the validity of clustering solutions. This index is a index of relative clustering validity, which means comparing the clustering results with a hypothetical ideal clustering. Lower DBI values indicate better clustering solutions.

- Higher DB index values indicate worse clustering solutions. This is because a higher DBI value indicates that the cluster is not well separated and the cluster is not compact.
- However, lower DB index values are more desirable. This value indicates that the clusters are well separated and compact, which is often a good indication of a successful clustering solution.

The Davies Bouldin index (DBI) formula is used to assess the quality of clustering results. This index measures the extent to which clusters are separated from each other and the extent to which they are dense. The lower the DBI value, the better the clustering results (clusters are denser and more clearly separated). The following is the formula for calculating the Davies-Bouldin Index in a clustering.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i\neq j}\left(\frac{d(ci,cj)}{ai+aj}\right) \qquad (3)$$

where :

- K is the number of clusters.
- ai is the size (density) of cluster i, usually calculated as the average distance between each point in the cluster and the cluster center (centroid).
- d (ci,cj) is the distance between cluster centers ci and cj.
- Max i=j that we look for pairs of clusters that have the maximum ratio between density and distance between centers.

This formula calculates the DBI value by adding up the maximum value of the ratio of the distance between cluster centers to the number of standard deviations for each different pair of clusters. Lower DBI values indicate better clustering.

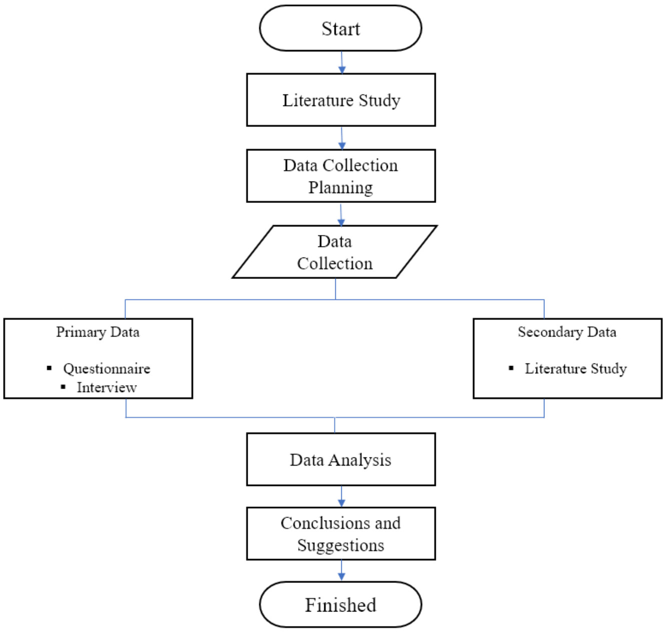


Figure 2. Qualitative flow diagram

2.8 Research Flow Diagram

The following is our qualitative research flow diagram. Qualitative research is a research method that uses descriptive data, in the form of written or spoken language, obtained from people or actors who can be observed. This approach aims to explain and analyze phenomena involving individuals or groups, such as events, social dynamics, attitudes, beliefs and perceptions. There is an example of a qualitative research flow diagram, as presented in Figure 2.

3. RESULTS AND DISCUSSIONS

3.1 Data Cleaning and Integration

Data cleaning and integration refers to the process of improving and combining data from various sources so that it is ready for analysis. Data cleaning involves identifying and correcting errors or discrepancies in data, such as duplication or missing values. Meanwhile, data integration is combining data from different sources to form a consistent and complete data set.

a. Checking empty data

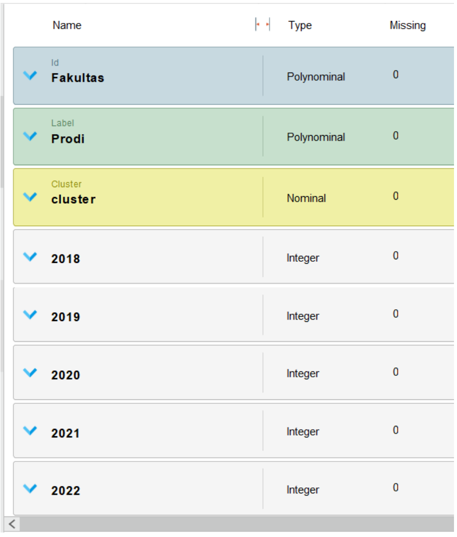
Figure 3 shows that the dataset has gone through a checking and validation process for missing values, and no empty data was found in all columns. This condition is excellent because it allows data analysis to be performed without requiring additional cleaning. This dataset is suitable for application in various analysis methods or direct modeling.

b. Checking for data that deviate significantly (outliers)

Figure 4 is a dataset showing potential outliers in the numerical column, especially in 2020, 2021, and 2022, because there is a large difference between the minimum and maximum values. Meanwhile, in categorical data, some categories with a very small number of entries (such as FPPsi or cluster_0) may need to be analyzed further to understand whether they are normal or an anomaly. Additional analyzes, such as distribution visualization and statistical calculations, are needed to confirm whether the data are indeed outliers.

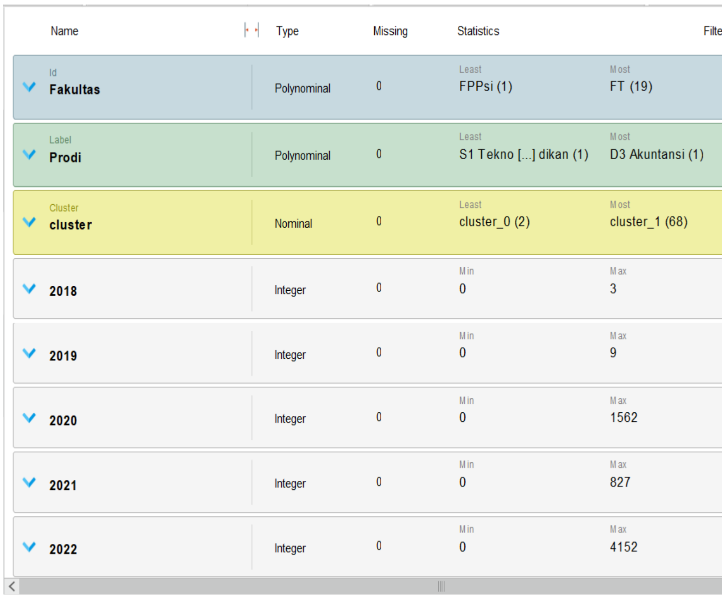
3.2 Data Mining

Figure 5 shows the data analysis flow for clustering experiments with various methods. The dataset is processed through several cleaning and attribute selection steps before further processing. This process allows exploring several clustering algorithms in parallel and comparing their performance. The final result is the selection of the best clustering method based on the performance evaluation carried out. Trials were carried out to obtain optimal clusters, starting from cluster 2 to cluster 6. Each clustering result was then evaluated for its performance.



Name	Type	Missing
Id Fakultas	Polynominal	0
Label Prodi	Polynominal	0
Cluster cluster	Nominal	0
2018	Integer	0
2019	Integer	0
2020	Integer	0
2021	Integer	0
2022	Integer	0

Figure 3. Blank data checking



Name	Type	Missing	Statistics		Filte
Id Fakultas	Polynominal	0	Least FPPsi (1)	Most FT (19)	
Label Prodi	Polynominal	0	Least S1 Tekno [...] dikan (1)	Most D3 Akuntansi (1)	
Cluster cluster	Nominal	0	Least cluster_0 (2)	Most cluster_1 (68)	
2018	Integer	0	Min 0	Max 3	
2019	Integer	0	Min 0	Max 9	
2020	Integer	0	Min 0	Max 1562	
2021	Integer	0	Min 0	Max 827	
2022	Integer	0	Min 0	Max 4152	

Figure 4. Checking for data with high deviation

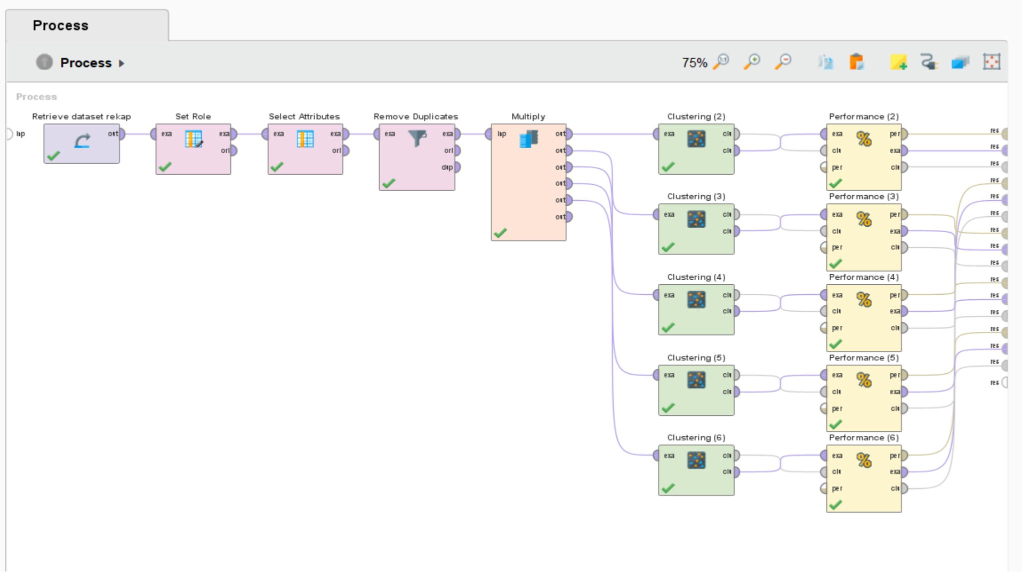


Figure 5. Creating a K-Means clustering algorithm model

The image above shows a workflow diagram of a data analytics platform, such as RapidMiner. The following is a detailed explanation of each step in the process:

a. Retrieve Dataset
This component is used to retrieve or load a dataset (named recap in the image). This dataset is likely the starting point of the analysis process.

b. Set Role
This step sets the role of the attribute in the dataset. For example: determining which is the label (variable target) and marking the attribute as an ID, custom attribute, or non-processing attribute.

c. Select Attributes
At this stage, certain attributes are selected for analysis. Possibly irrelevant attributes are discarded so that only important columns are used for further processing.

d. Remove Duplicates
This component is used to delete duplicate data rows (based on certain criteria). The goal is to ensure the dataset has no redundancy.

e. Multiply

This component duplicates the dataset into several copies (replicas). This allows the same dataset to be processed in paraller in the next step, namely applying the clustering method with different parameters or algorithms.

f. Clustering
The duplicated dataset is previously processed through various clustering methods (or the same algorithm with different parameters). In the picture there are 6 clustering components: Clustering 2 to Clustering 6. Each works in parallel to produce groups (clusters) from the dataset based on their characteristics.

g. Performance Evaluation
After each clustering process, a performance evaluation is carried out. This evaluation measures the quality of clustering results using certain metrics, such as silhouette score, inertia, or other cluster validity.

3.3 Model Evaluation

Model evaluation is an important step to determine the effectiveness of a clustering method. In the figure, evaluation is carried out on several clustering models to compare their performance in parallel. The results of this evaluation will help select the clustering model with the best results based on appropriate metrics.

Table 2 is the Davies Bouldin index, a clustering evaluation metric that assesses the quality of clusters based on the density ratio within clusters and the separation between clusters. A smaller DBI value indicates better clustering results. This metric is very useful in unsupervised analysis, but it is important to note its sensitivity to the number of clusters and data structure. The smallest Davies Bouldin value, namely 0.079, was obtained in the clustering model with the number of clusters k = 3. Therefore, the clustering model chosen was with k = 3 which shows the best grouping quality compared to other k values in the range k = 2 to k = 6.

The Davies Bouldin index measures how good the resulting clusters are, where the value is more low indicates better cluster. This result is encouraging that the K-Means model has been evaluated and the user needs to check this Davies Bouldin value for determine the quality of the resulting clustering.

Table 3 is data grouping clustering, which is a technique in data analysis which aims to group data into groups or clusters that have similarities based on certain characteristics or features. In data grouping, objects that have higher similarity to each other are placed in one group cluster, while objects that are different are grouped separately. Grouping the dataset with k = 3 produces three clusters, namely cluster 0 which consists of 68 data, cluster 1 which consists of 1 data, and cluster 2 which also consists of 1 data. Shows the amount of data in each cluster for each k value.

Table 2. Davies Bouldin value

Number of Clusters	Davies Bouldin Value
2	0.488
3	0.079
4	0.511
5	0.419
6	0.498

Table 3. Number of clusters for each K value

Number of Clusters	Data Grouping
2	Group 0 : 2 data Group 1 : 68 data
3	Group 0 : 68 data Group 1 : 1 data Group 2 : 1 data
4	Group 0 : 63 data Group 1 : 1 data Group 2 : 1 data Group 3 : 5 data
5	Group 0 : 59 data Group 1 : 1 data Group 2 : 6 data Group 3 : 1 data Group 4 : 3 data
6	Group 0 : 36 data Group 1 : 1 data Group 2 : 3 data Group 3 : 1 data Group 4 : 26 data Group 5 : 3 data

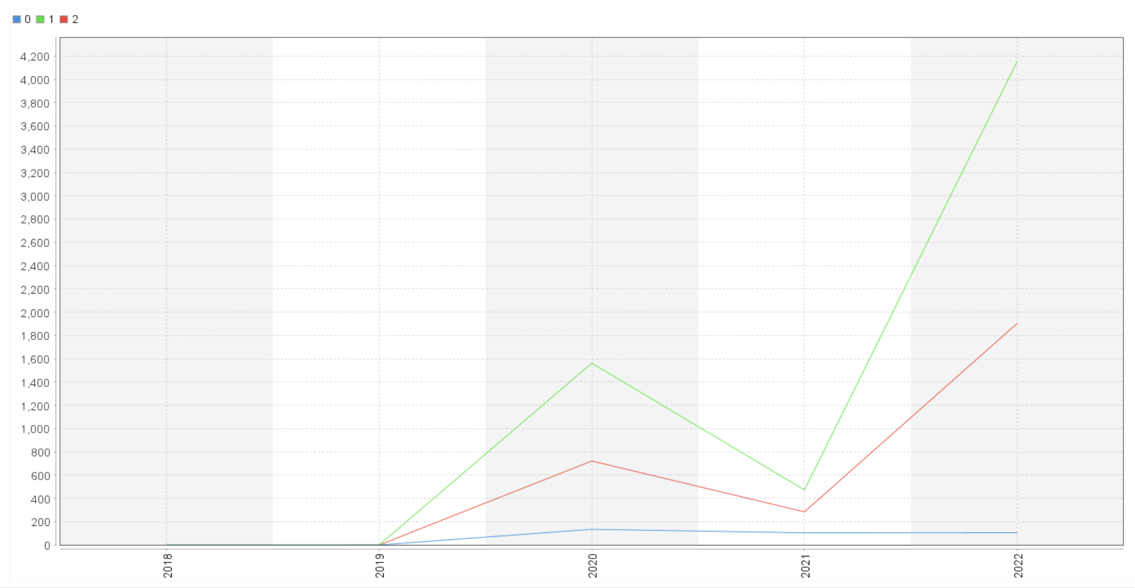


Figure 6. Plot the results from Cluster 0, Cluster 1, and Cluster 2

Figure 6 is a graph that depicts changes in Malang University student achievement data from various study programs, which are divided into three clusters, namely cluster 0, cluster 1, and cluster 2, showing changes in scores for each cluster from the period 2018 to 2022. The blue cluster (0) has the lowest number of achievements throughout that period. The green cluster (1) has the highest number of achievements and the best achievements compared to other clusters. Meanwhile, the red cluster (2) has a number of achievements with consistent improvement, this study program has quite good performance but is still below cluster 1.

Table 4 is data that shows that most of the activities or numbers in study programs have started to encrease drastically since 2020, especially in cluster_1 and cluster_2. This indicates a significant change in the distribution of data or activity in that year, which may be related to external factors such as educational policies or development programs. Cluster_0 tends to be more stable, with a more moderate increase compared to other clusters. The results of this data grouping are based on attributes such as faculty and study program, with the clustering results divided into three groups: cluster_0, cluster_1, and cluster_2. The data covers the period 2018 to 2022 for each row, providing an idea of how data from various study programs and faculties are grouped according to certain patterns based on the attributes analyzed.

3.4 Results Visualization

Data visualization in K-Means clustering helps in understanding how the algorithm groups data based on similarities. This visualization makes it easy to see how data is distributed within clusters and how well they are separated. With proper visualization we can see what K-Means is clustering produces well-separated clusters.

Table 4. Cluster results for each data

Row Now.	Fakultas	Prodi	Cluster	2018	2019	2020	2021	2022
1	FEB	S1 Pendidikan …	Cluster_0	0	0	396	36	120
2	FEB	S1 Akuntansi	Cluster_1	1	1	1562	473	4152
3	FEB	S1 Pendidikan …	Cluster_2	0	0	722	285	1903
4	FEB	S1 Manajemen	Cluster_0	0	0	707	61	339
5	FEB	S1 Pendidikan …	Cluster_0	0	0	586	86	549
6	FEB	S1 Pendidikan …	Cluster_0	0	0	266	18	78
7	FEB	S1 Ekonomi P …	Cluster_0	0	0	421	24	251
8	FEB	D3 Akuntansi	Cluster_0	0	0	69	18	0
9	FEB	D3 Manajemen …	Cluster_0	0	0	35	7	7
10	FIK	S1 Pendidikan …	Cluster_0	1	0	291	159	37
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
70	FT	D3 Tata Busana	Cluster_0	0	0	0	1	0

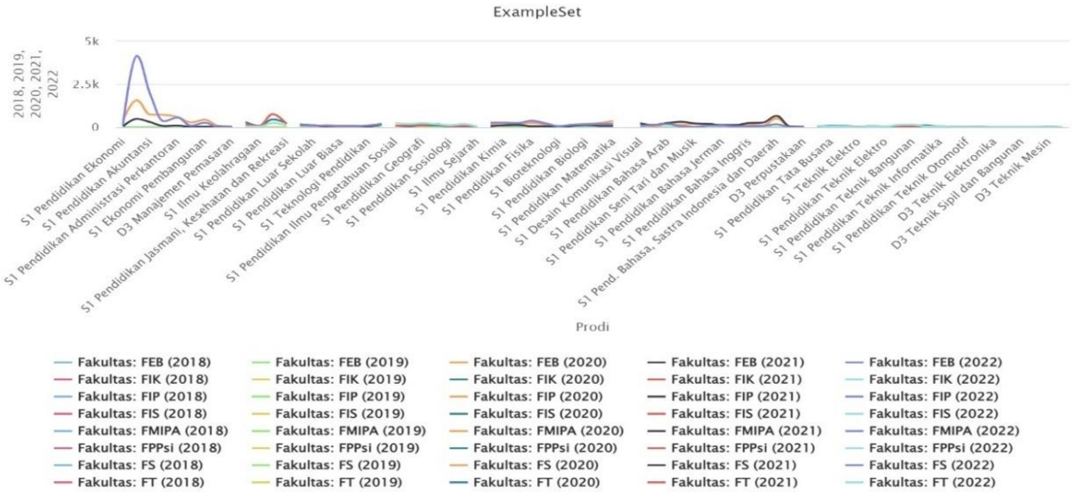


Figure 7. Visualization of each study program

Figure 7 is a line graph that depicts the amount of data for various study programs from 2018 to 2022. The graph presents the amount or identity of data per year for each study program and highlights data variations between faculties which are represented by different line colors. This figure makes possible to compare changes in the number or value of data between study programs in different years as well as examine differences between faculties. The peaks and troughs in this graph reflects changes that occur over time, with some courses having high scores in a given year, while others may be stable or experience little change.

The following is Figure 8 to Figure 12, a graph that depicts data from various faculties for the period 2018 to the period 2022. Each line represents data from a different faculty and there are several different faculties and there are several faculties recorded for each year such as FEB, FIK, FIP, FMIPA, FPPsi, FS, and FT. Overall this graph appears to show a trend in the data, by faculty and year. For a more in-depth interpretation, it would be useful to know more about what this graph measuring (for example, student achievement in each study program and graduation) to provide more specific context.

Figure 8 shows a graph of data for the 2018 period, with results mostly in the range 0 to 1000 for the majority of faculties. Several important points for 2018 can be seen as follows, the 2018 period shows scores in the low range, most faculties have results close to 0 to 1000. The faculty that has the highest score for 2018 is probably FEB 2018 with a score that looks greater than other faculties. Other faculties, such as FIK, FIP, FMIPA, and others, have much lower results, most of which are below 500. In general, 2018 shows significant differences between FEB and other faculties in this graph.

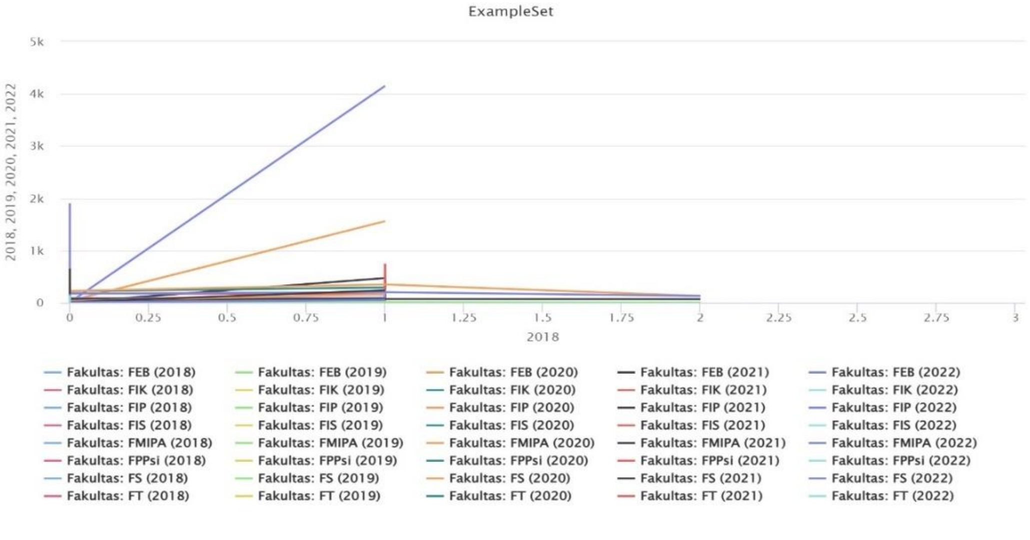


Figure 8. Visualization of study programs for the 2018

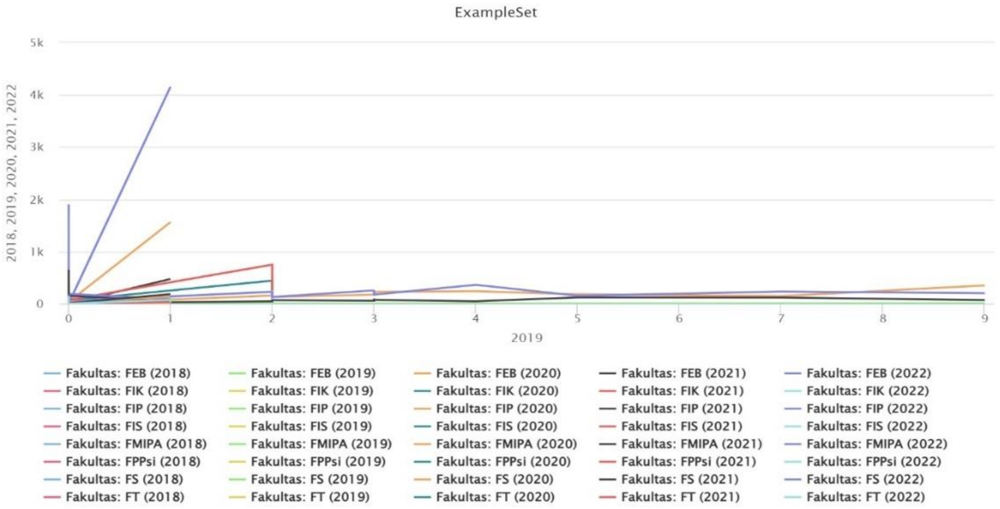


Figure 9. Visualization of study program for the 2019

Figure 9 shows a graph of data for the 2019 period, which has relatively small fluctuations compared to 2018, and subsequent years. In 2019, most faculties had stable scores and were below 1000, without significant changes compared to other periods. Only a few faculties showed small improvements, while most remained consistent. Faculties that appear to stand out from other faculties may have a small spike, but it is not very significant when compared to other years (for example 2020 or 2021).

Figure 10 shows a data graph for the 2020 period, which is the beginning of major changes for several faculties. Where several faculties, such as the faculty of Engineering FT and the faculty of Economics and Business FEB, experienced significant improvements, their scores, rose sharply compared to previous years. Meanwhile, other faculties such as FIK, FIP, and FIS seem to remain stable with relatively low scores, below 1000. Overall, the 2020 period shows the beginning of major changes for several faculties, especially with data spikes not seen in previous years, previous year.

Figure 11 shows a data graph for the 2021 period, where the faculty of Engineering FT and the faculty of Economics and Business FEB experienced a significant increase, with FT reaching more than 3000 and FEB around 2000. Several faculties, such as FMIPA and FIP, experienced growth stable with scores still below 1000. Meanwhile, other faculties such as FIK and FIS remain consistent with low scores, close to 0 to slightly above 100. Year 2021 marks major development for FT and FEB compared to other faculties 2021 shows a big increase in several faculties, especially FT and FEB which are leaders in the graph. Other faculties tend to grow slowly or remain stable with smaller values.

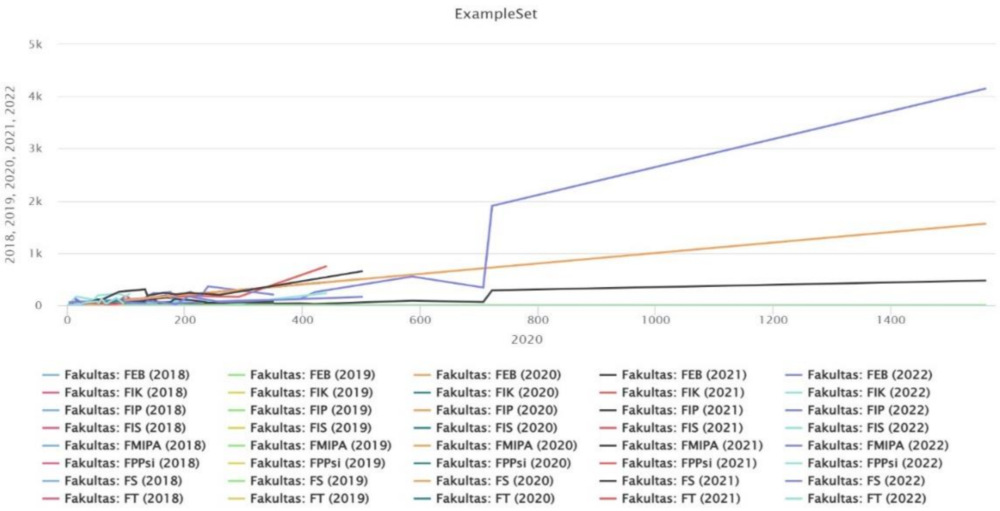


Figure 10. Visualization of study program for the 2020

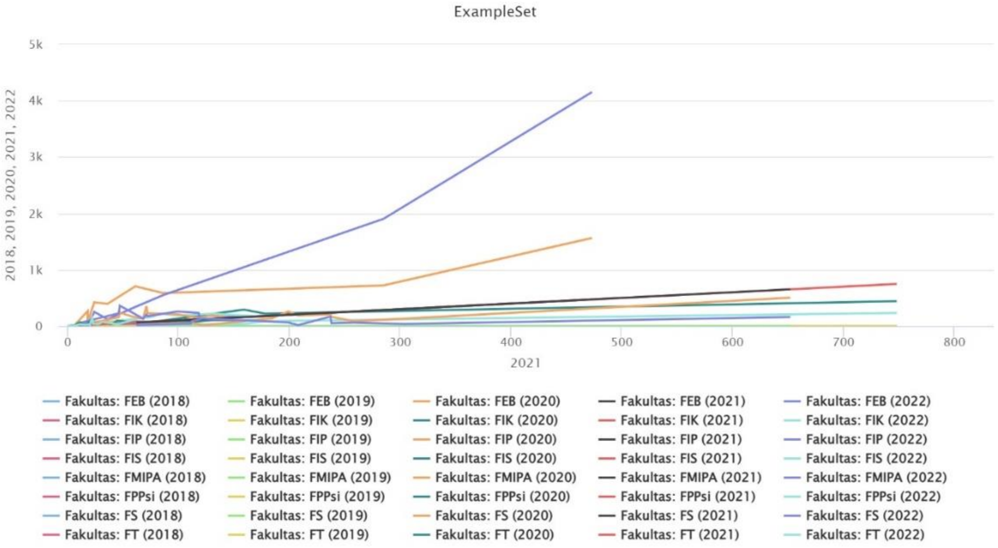


Figure 11. Visualization of study program for the 2021

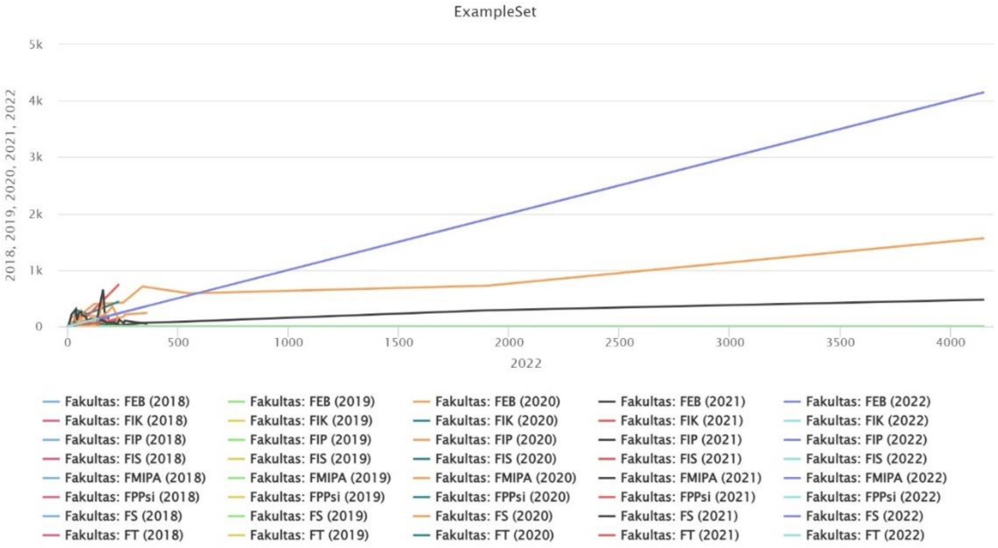


Figure 12. Visualization of study programs for the 2022

Figure 12 shows graphic data for the 2022 period, where the faculty of Engineering FT continues to show significant improvement, reaching a score close to 5000, making it the faculty with the highest data on the graph, followed by the faculty of FEB which also continues its growth to close to 3000, occupying second place highest. Other faculties such as FMIPA and FIP grew steadily with scores below 1000, while FIK and FIS remained consistent with low scores, close to 0 to slightly above 100, FT and FEB will be the most prominent faculties in 2022.

4. CONCLUSION

The clustering model can be used to group data in a dataset. This grouping assists in decision-making, especially in determining groups that require policies or specific strategies to improve performance, so the initial data in the unfavorable group can be enhanced to enter a better group. Using an achievement dataset, students from 2018 to 2022 resulted in 3 groups. Cluster 0 has the lowest performance throughout the year, with a slight increase but relatively stable with a value of 100. Cluster 1 has the highest number of achievements, especially in 2022, which shows a significant spike; this indicates that the study program has better achievement performance than the cluster other with a value of 4.100. Meanwhile, Cluster 2 shows a consistent increase, with a spike in 2022 after experiencing a decline in 2021; study programs in this cluster have pretty good performance but are still below Cluster 1 with a value of 1.900.

For further research, add other factors, such as economic conditions or campus facilities, for a more in-depth analysis. Try other clustering methods to compare results. Analyze data over a more extended period to see trends more clearly. Evaluate the effectiveness of achievement improvement programs based on groups. Use interactive data visualization to facilitate understanding and decision-making. It is hoped that these suggestions can help research develop further and provide more significant benefits for the university.

REFERENCES

[1] D. R. Retnowati, A. Fatchan, and K. Astina, (2016), "Prestasi Akademik dan Motivasi Berprestasi Mahasiswa Universitas Negeri Malang," *Jurnal Pendidikan*, vol. 1, no. 3, pp. 521–525.

[2] A. A. Saputro and R. Helilintar, (2020), "Perancangan Prediksi Prestasi Nilai Akademik Mahasiswa Menggunakan Metode K-Means Clustering," in *Seminar Nasional Inovasi Teknologi*, pp. 49–55.

[3] Z. Zaifullah and T. Yulianto, (2022), "Analisis Cluster Untuk Pengelompokkan Prestasi Mahasiswa Angkatan 2013 Fakultas MIPA Universitas Islam Madura," *Zeta - Math Journal*, vol. 7, no. 1, pp. 1–10, doi:10.31102/zeta.2022.7.1.1-10.

[4] J. Ramadhani, M. D. Nawar, and N. M. P. Aritonang, (2024), "Penilaian Pengelompokan Data Prestasi Siswa Menggunakan Metode K-Means Untuk Mengenali Siswa Berprestasi," *J-Com (Journal of Computer)*, vol. 4, no. 1, pp. 15–22, doi:10.33330/j-com.v4i1.2977.

[5] S. N. B. Sembiring, H. Winata, and S. Kusnasari, (2022), "Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means," *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, vol. 1, no. 1, pp. 31–40, doi:10.53513/jursi.v1i1.4784.

[6] I. Vhallah, S. Sumijan, and J. Santony, (2018), "Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Metode Clustering K-Means," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 2, no. 2, pp. 572–577, doi:10.29207/resti.v2i2.308.

[7] B. Basmalia and Z. Fatah, (2024), "Prediksi Kelulusan Siswa Menggunakan Algoritma K- Nearest Neighborgs (K-NN) Di SMK Al Hasyimy Ibrahimy," *Gudang Jurnal Multidisiplin Ilmu*, vol. 2, no. 10, pp. 110–115, doi:ttps://doi.org/10.59435/gjmi.v2i11.1050.

[8] A. Wasik, Z. Fatah, and A. Munazilin, (2024), "Implementasi data mining untuk memprediksi penjualan accessoris handphone dan handphone terlaris menggunakan metode k-nearest neighbor (k-nn) 1," in *Seminar Nasional Sains dan Teknologi "SainTek,"* vol. 1, no. 2, pp. 469–479.

[9] S. Widaningsih, (2019), "Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, KNN dan SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, doi:10.36787/jti.v13i1.78.

[10] Narwati, (2010), "Pengelompokan Mahasiswa Menggunakan Algoritma K-Means," *Jurnal Dinamika Informatika*, vol. 2, no. 2, pp. 1–7.

[11] A. Yudhistira and R. Andika, (2023), "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *Journal of Artificial Intelligence and Technology Information (JAITI)*, vol. 1, no. 1, pp. 20–28, doi:10.58602/jaiti.v1i1.22.

[12] S. Suliman, (2021), "Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan dan Sosial Ekonomi Dengan Algoritma K-Means Clustering," *Jurnal Sistem Informasi dan Sistem Komputer*, vol. 6, no. 1, pp. 1–11, doi:10.51717/simkom.v6i1.48.

[13] R. N. Sukmana, A. Mita, and A. Abdurrahman, (2019), "Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma K-Means Clustering," *Jurnal teknologi informasi dan komunikasi*, vol. 8, no. 1, pp. 11–15, doi:10.58761/jurtikstmikbandung.v8i1.125.

[14] E. A. Saputra and Y. Nataliani, (2021), "Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means," *Journal of Information Systems and Informatics*, vol. 3, no. 3, pp. 424–439, doi:10.33557/journalisi.v3i3.164.

[15] S. Suraya, M. Sholeh, and D. Andayati, (2023), "Penerapan Metode Clustering Dengan Algoritma K-Means Pada Pengelompokan Indeks Prestasi Akademik Mahasiswa," *Skanika*, vol. 6, no. 1, pp. 51–60, doi:10.36080/skanika.v6i1.2982.

[16] I. A. Kurniawan, R. M. H. Bhakti, and B. Irawan, (2024), "Implementasi Data Mining Untuk Mengukur Prestasi Siswa SD Menggunakan Metode K-Means Clustering," *JCRD: Journal of Citizen Research and Development*, vol. 1, no. 2, pp. 262–268, doi:10.57235/jcrd.v1i2.3324.