Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

K-Means Algorithm for Clustering High-Achieving Student at Madrasah Tsanawiyah Yami Waled

Muhammad Hilman[1*], Martanto[2], Arif Rinaldi Dikananda[3], Ahmad Rifai[4]

[1,4]Informatics Enginneering, STMIK IKMI Cirebon, Indonesia
[2] Informatics Management, STMIK IKMI Cirebon, Indonesia
[3]Software Engineering, STMIK IKMI Cirebon, Indonesia
astraguppy@gmail.com[1], martantomusijo@gmail.com[2], rinaldi21crb@gmail.com[3], a.rifaaii1408@gmail.com[4]*

Abstract

This study aims to apply the K-Means algorithm to cluster students based on their mathematics grades at Madrasah Tsanawiyah Islamiyyah Yami Waled. By categorizing students into clusters of low, medium, and high academic achievement, the institution can develop more effective and targeted learning strategies. The data consisted of semester mathematics grades from 112 students, analyzed using the K-Means clustering algorithm. Clusters were evaluated using the Davies-Bouldin Index (DBI), with results showing three distinct clusters: Cluster 0 (low achievers, 54 students), Cluster 1 (medium achievers, 37 students), and Cluster 2 (high achievers, 21 students). The DBI score of 0.893 indicates good clustering quality, providing valuable insights for personalized learning approaches.

Keywords: *K-Means, Clustering, Education, Academic Achievement, Data Mining*

1. Introduction

The advancement of information technology has driven innovation in education, particularly in academic data management. The K-Means clustering algorithm has become a popular method for grouping data based on similar characteristics. Research shows that this algorithm is effective in understanding the distribution of students' academic abilities, especially in Mathematics [1]; & [2]. At Madrasah Tsanawiyah Islamiyyah Yami Waled, the implementation of the K-Means algorithm is expected to assist teachers in designing more effective learning strategies, particularly through grouping students based on their Mathematics grades. Previous studies, such as [3], have also demonstrated the success of this algorithm in clustering academic data in various contexts, while [4] highlight the importance of integrating information technology to improve data-driven education quality.

External The primary challenges in managing student data include limited technology utilization and educators' lack of data analysis capabilities [5]; [1]. Furthermore, earlier studies have focused more on algorithm accuracy without addressing practical implementation barriers [3]. This study aims to bridge that gap by utilizing the K-Means algorithm to analyze students' Mathematics grades comprehensively. By adopting this approach, the study is expected to provide technology-based solutions to support more effective and student-responsive educational decision-making, as suggested by previous research [6];[7].

Through the application of the K-Means algorithm, this research seeks to significantly contribute to education quality at Madrasah Tsanawiyah Islamiyyah Yami Waled. The clustering results will help identify students needing special attention and design targeted intervention programs. Moreover, this study aims to enrich the literature on applying data mining in education and support the development of technology-based educational policies [8]; [9]. With a data-driven approach, these findings are expected to enhance learning quality and serve as a foundation for developing more adaptive educational technology.

2. Litelatur Riview

Various studies demonstrate the effectiveness of the K-Means algorithm in clustering data for diverse purposes. [10] applied K-Means to group pharmaceutical data based on characteristics such as price, type, and form of medicine to assist in managing pharmacy inventories. [6] utilized K-Means to map provinces in Indonesia based on poverty levels, categorizing them into three clusters: high, medium, and low, providing insights for government policymaking. [1] employed K-Means for selecting high-achieving students based on their participation in learning activities, identifying clusters of students with high, medium, and low levels of engagement. Similarly, [8] applied K-Means to group students based on academic performance into three clusters—high, medium, and low achievers—helping teachers design targeted learning strategies. [11] focused on clustering Umrah pilgrims based on age, gender, and travel package preferences, aiding travel agencies in refining their marketing and services. Meanwhile, [7] implemented K-Means to predict recipients of Village Fund Cash Assistance (BLT) based on residents' economic conditions, ensuring more accurate and targeted aid distribution. Overall, these studies highlight that

K-Means is an effective method for data clustering to support decision-making in the fields of pharmaceuticals, education, economics, and public services.

Several studies have demonstrated the effectiveness of the K-Means clustering algorithm in various fields. According to [3], K-Means was applied to determine the Single Tuition Fee (UKT) for students by grouping them based on socioeconomic conditions, such as parental income, family dependents, and living conditions, to ensure a fair and targeted tuition policy. [5] focused on mapping high-achieving students based on academic performance, attendance, and extracurricular activities to identify top-performing students and optimize learning strategies tailored to each cluster's needs. Similarly, [9] applied K-Means to cluster Covid-19 spread data in West Java, categorizing regions into high, medium, and low transmission levels to assist policymakers in implementing effective prevention strategies. [4] utilized K-Means to analyze product sales at Yana Sport Store, identifying clusters of products based on sales performance to improve inventory management and marketing strategies. In health, [12] used K-Means to cluster Acute Respiratory Infection (ISPA) cases in Karawang, grouping regions based on disease incidence to develop more effective health interventions. Lastly, [2] applied K-Means to analyze student grades and group them into categories of high-achievers, potential students, and those needing academic support, enabling schools to design targeted educational programs. These studies collectively highlight the utility of the K-Means algorithm in clustering data for fair decision-making in education, healthcare, retail, and policymaking contexts.

The research conducted by [13], [14], [15], [16] and [17] explores the application of the K-Means clustering algorithm in various fields to solve data-driven problems. [13] focused on clustering patient diseases at Puskesmas Cigugur Tengah to identify common disease patterns and improve healthcare interventions. [14]. Implemented K-Means to classify high-achieving students based on academic performance and extracurricular involvement, aiding schools in determining optimal class placement. [15] analyzed sales levels of food and beverage menus to identify the most and least popular items, supporting restaurants in optimizing promotions and inventory management. [16] clustered lecturers based on attendance levels to evaluate and improve teaching performance and resource management at educational institutions. Meanwhile, [17] applied K-Means to assess the technical abilities of IT employees, enabling companies to design tailored training programs and optimize employee placement. Across these studies, the K-Means algorithm was shown to be an effective tool for analyzing diverse datasets, identifying meaningful patterns, and supporting strategic decision-making in healthcare, education, and business management.

3. Research Methods

This study uses the Knowledge Discovery in Databases (KDD) approach to build a model for clustering high-achieving students in Mathematics. The stages of KDD are illustrated in Figure 1 below.

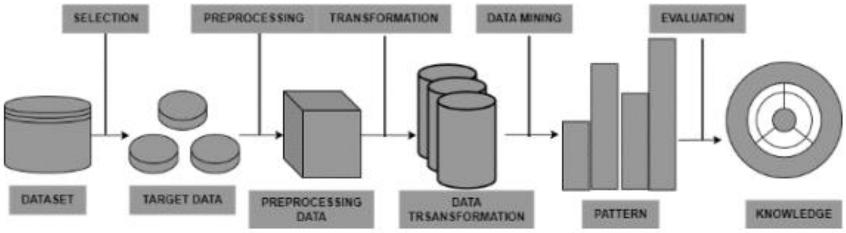


Fig. 1 : Stages of the KDD Method [18]

Description of Research Methods Using KDD:

1. Selection: Collecting student mathematics grade data at Madrasah Tsanawiyah Islamiyyah Yami Waled.
2. Preprocessing: Removing irrelevant values, addressing missing data, and normalizing grade ranges for K-Means.
3. Transformation: Converting grade data into a format suitable for the K-Means algorithm.
4. Data Mining: Using the K-Means algorithm to cluster students based on their mathematics grades.
5. Evaluation: Analyzing clustering results to assess whether the generated clusters align with the research objectives.

3.1. Data Sources

The data sources for the research on the K-Means Algorithm to Improve the Clustering Model of High-Achieving Students in Strategy Determination at Madrasah Tsanawiyah Islamiyyah Yami Waled will be obtained from primary data, specifically midterm and final semester mathematics grades of students directly collected from the school. Data collection will be conducted through the school's archival method to ensure valid and reliable grades. Access to the data will be obtained with official permission from the school authorities, and the data format will be adjusted to meet the analytical requirements for the K-Means algorithm.

3.2. Population and Sample

The identified population consists of all students at Madrasah Tsanawiyah Islamiyyah Yami Waled who have mathematics grades over a specific period. For sample selection, a purposive sampling method will be used, where samples are intentionally chosen from students with complete and relevant mathematics grade records suitable for analysis using the K-Means algorithm. The sample size will be determined based on the number of students meeting these criteria, totaling 112 students. Sampling is conducted while considering variations in students' mathematics proficiency levels as assessed in midterm and final semester evaluations for mathematics subjects.

3.3. Data Selection

Data collection was conducted using archival methods of mathematics grades from students at Madrasah Tsanawiyah Islamiyyah Yami Waled. The data was obtained directly from official school records, covering Mathematics exam results over a specific period. The data collection process was structured to ensure all data obtained was relevant and complete for clustering analysis. The data was then processed for implementing the K-Means algorithm, aiming to cluster students based on patterns in their mathematics grades.

3.4. Data Analysis

Data analysis was conducted using the K-Means algorithm. Below are descriptions of the Euclidean Distance and Centroid calculations used in the data analysis to evaluate clustering results for improving the clustering model of high-achieving students at Madrasah Tsanawiyah Islamiyyah Yami Waled:
The following is the calculation Euclidean Distance

$$d = \sqrt{\sum_{N}(X_i - Y_i)^2} \tag{1}$$

Where :

d = distance (2)
X_i = Attribut data point x (3)
Y_i = Attribut data point y (4)
i = *Index* attribut in data (5)
N = Result of addition $(X_i - Y_i)^2$ (6)

Here is the Centroid calculations

$$Centroid = \frac{the\ sum\ of\ all\ attribute\ values}{amount\ of\ data} \tag{7}$$

A centroid is the average point of a dataset, calculated by summing all attribute values and dividing by the number of data points.

4. Results and Discussion

At this stage, the researcher will describe the implementation of the K-Means algorithm in clustering high-achieving students at Madrasah Tsanawiyah Islamiyyah Yami Waled as the basis for determining more effective learning strategies. This process begins with the KDD (Knowledge Discovery in Databases) stages, which include selecting student data, preprocessing to prepare the data, and analyzing the clustering results. With this approach, students are grouped according to their achievement levels, allowing the madrasah to design more suitable strategies for each group. Below is the implementation of KDD in this research.

4.1. Data Selection

Selection is the initial stage of the Knowledge Discovery in Databases (KDD) process, aimed at identifying and collecting relevant data to achieve the research objectives. The selected data consists of the mathematics grades of students in both the odd and even semesters at Madrasah Tsanawiyah Islamiyyah Yami Waled, with a focus on grouping students based on their performance in mathematics. The mathematics grade data is shown in Table 1 below.

Table 1: Dataset

NO	NAMA	JENIS KELAMIN	TS-S1	AS-S1	TS-S2	AS-S2
1	ARYA PAERUL ADHA	L	75	77	80	82
2	ASEP SURYA	L	78	74	75	77
3	CANTIKA RAHAYU	P	76	79	77	79
4	DARUL ROHMAH	L	76	77	77	79
5	DWI SULISTIANI	P	78	77	81	83
6	FARHAN ABADAN	L	73	78	81	83
7	HAMDAN SIROJ MUQSITH	L	77	75	73	75
8	HOERUL ROZIKIN	L	79	76	74	76
9	JIHAN ALEXNIA RAMADHANI	P	76	78	80	82
10	M. LUTHFIHAN	L	74	74	76	78
11	M. YUSUF MAULANA	L	75	71	77	79
……….	……….	……….	……….	……….	……….	………
112	ZEVA FREDY PRATAMA	**L**	75	75	77	79

Table 1 above shows the mathematics grade dataset, which includes several important attributes that represent the identity and academic performance of students at Madrasah Tsanawiyah Islamiyyah Yami Waled. The first attribute is the Student Number (NO), which serves as a unique identifier for each student. The next attribute is the Student Name (NAMA), which records the identity of each student, followed by Gender (JENIS KELAMIN), represented by the code "L" for male and "P" for female. Next, there are four academic performance attributes: Mid-Term Grade 1 (TS-S1), Final Grade 1 (AS-S1), Mid-Term Grade 2 (TS-S2), and Final Grade 2 (AS-S2), which reflect the students' performance during two different semester periods. These grades will be used as the basis for the clustering process using the K-Means algorithm.

4.2. Data preprocessing

The preprocessing of data in this study aims to prepare the dataset optimally before being implemented in the clustering model. The initial step is to read the data from the CSV file to ensure that the data used aligns with the analysis needs. Next, attribute selection is performed to choose relevant and significant attributes for clustering purposes, such as students' academic scores and other supporting attributes. Then, roles are assigned to each attribute, with some attributes being focused on as the basis for the clustering process. With this preprocessing step, the data becomes more structured and ready to support accurate clustering. The implementation of this step can be seen in Fig 2 below.

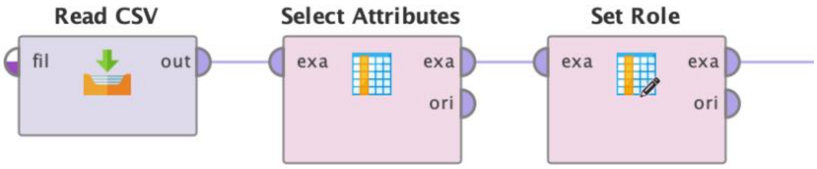
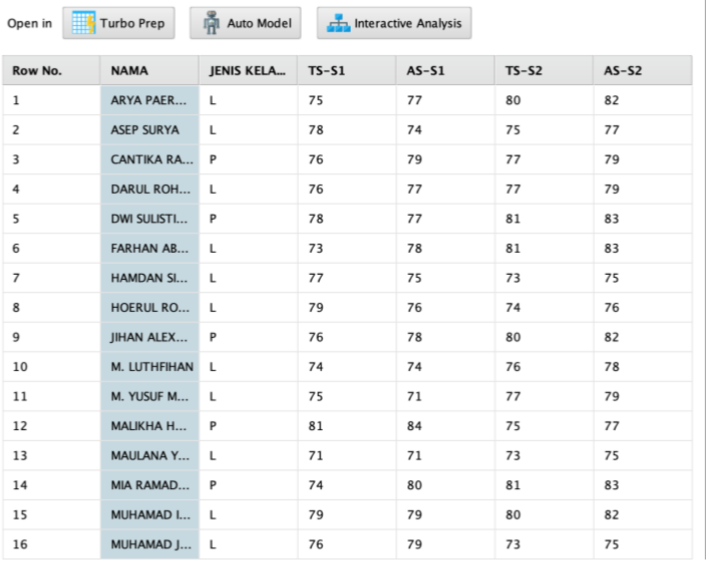


Fig. 2: Operator Read CSV, Select Attributes, and Set Role.

Fig. 2 shows the operators Read CSV, Select Attributes, and Set Role, with three operators used in the data preprocessing process. Below is an explanation of the function of each operator:

1. **Read CSV Operator**: This operator reads the CSV file containing student data. It imports the data from the external file into the processing environment, making it ready for the next preprocessing steps. Figure 4.2 shows the output of the Read CSV operator.



Row No.	NAMA	JENIS KELA...	TS-S1	AS-S1	TS-S2	AS-S2
1	ARYA PAER...	L	75	77	80	82
2	ASEP SURYA	L	78	74	75	77
3	CANTIKA RA...	P	76	79	77	79
4	DARUL ROH...	L	76	77	77	79
5	DWI SULISTI...	P	78	77	81	83
6	FARHAN AB...	L	73	78	81	83
7	HAMDAN SI...	L	77	75	73	75
8	HOERUL RO...	L	79	76	74	76
9	JIHAN ALEX...	P	76	78	80	82
10	M. LUTHFIHAN	L	74	74	76	78
11	M. YUSUF M...	L	75	71	77	79
12	MALIKHA H...	P	81	84	75	77
13	MAULANA Y...	L	71	71	73	75
14	MIA RAMAD...	P	74	80	81	83
15	MUHAMAD I...	L	79	79	80	82
16	MUHAMAD J...	L	76	79	73	75

Fig. 3: Output of the Read CSV Operator.

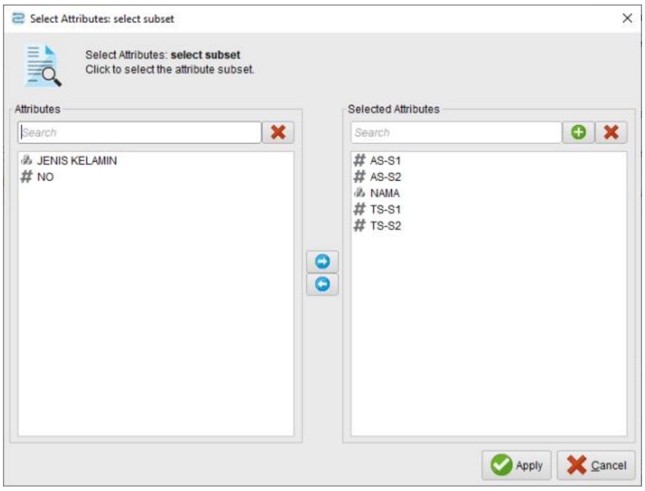
Fig. 3 displays the output from the Read CSV operator, which includes 6 attributes such as NAMA, JENIS KELAMIN, TS-S1, AS-S1, TS-S2, and AS-S2. The data types of these attributes are shown in Table 2.

Table 2: Data Types of the Mathematics Grade Dataset.

No	Nama Attribut	Tipe Data
1	Nama	*Polynominal*
2	Jenis Kelamin	*Polynominal*
3	TS-S1	*Integer*
4	AS-S1	*Integer*
5	TS-S2	*Integer*
6	AS-S2	*Integer*

Table 2 contains data with several attributes: "Nama," which is of the polynomial data type containing individual names; "Jenis Kelamin," also polynomial, indicating the gender of individuals; and four integer columns (TS-S1, AS-S1, TS-S2, and AS-S2) containing grades or scores that likely refer to certain measurements or categories in the analysis.

2. **Select Attributes Operator:** After the data is imported, the Select Attributes operator is used to choose the attributes relevant to the analysis goal. At this stage, attributes that are irrelevant or not directly related to the clustering process can be removed from the dataset.



Select Attributes: select subset

Select Attributes: **select subset**
Click to select the attribute subset.

Attributes

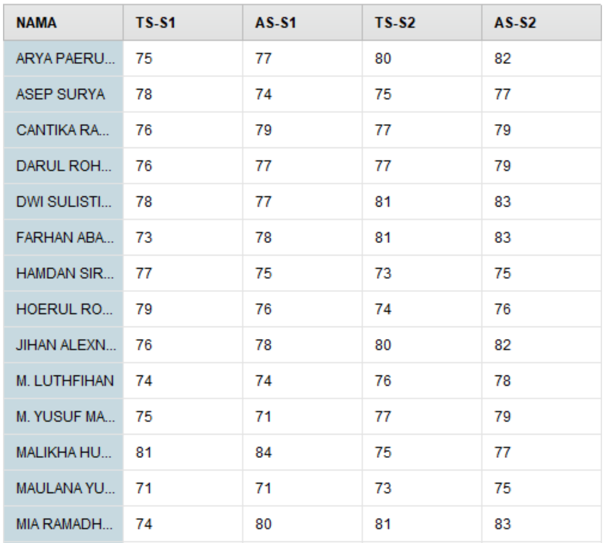
- JENIS KELAMIN
- NO

Selected Attributes

- AS-S1
- AS-S2
- NAMA
- TS-S1
- TS-S2

Fig. 4: Attribute Selection Process.

Fig. 4 shows the attributes that will be selected for further analysis. The selected attributes include NAMA, TS-S1, AS-S1, and TS-S2. The result of the attribute selection can be seen in Fig. 5 below.

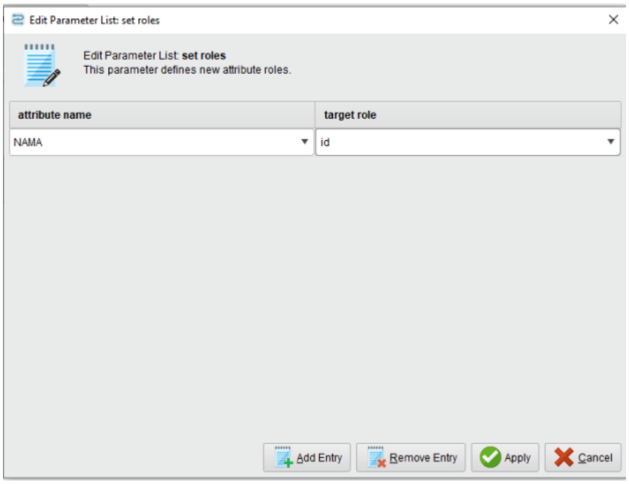


NAMA	TS-S1	AS-S1	TS-S2	AS-S2
ARYA PAERU...	75	77	80	82
ASEP SURYA	78	74	75	77
CANTIKA RA...	76	79	77	79
DARUL ROH...	76	77	77	79
DWI SULISTI...	78	77	81	83
FARHAN ABA...	73	78	81	83
HAMDAN SIR...	77	75	73	75
HOERUL RO...	79	76	74	76
JIHAN ALEXN...	76	78	80	82
M. LUTHFIHAN	74	74	76	78
M. YUSUF MA...	75	71	77	79
MALIKHA HU...	81	84	75	77
MAULANA YU...	71	71	73	75
MIA RAMADH...	74	80	81	83

Fig. 5: Output of the Select Attributes Operator.

Fig. 5 displays the output from the Select Attributes operator, which shows the attributes NAMA, TS-S1, AS-S1, and TS-S2 that will be grouped and analyzed further.

3. Set Role Operator: This operator assigns roles to each attribute in the dataset according to its function in the clustering process. For example, certain attributes may be set as labels or identifiers, while others may be designated as the basis for clustering. By defining the role of each attribute, the clustering process can proceed in a more targeted and effective manner.



Edit Parameter List: set roles

Edit Parameter List: **set roles**
This parameter defines new attribute roles.

attribute name	target role
NAMA	id

Fig. 6: Set Role Process.

Fig. 6 shows the attributes in the dataset assigned with roles. The attribute "NAMA" is assigned the role of ID, which serves as a unique identifier or identity for each data sample.

4.3. Transformation

Next, data transformation is performed to convert nominal attributes into numerical ones using the Nominal to Numerical operator. This conversion aims to optimize the analysis of categorical data. Figure 4.6 below shows the data transformation process.

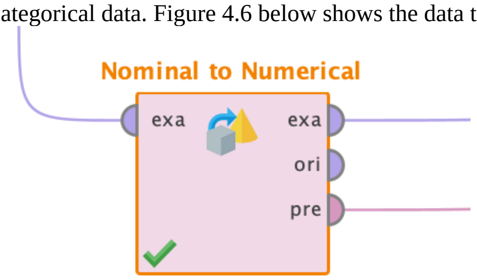
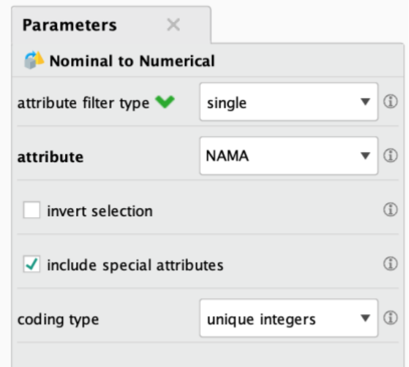


Fig. 7: Nominal to Numerical Operator.

Fig. 7 illustrates the use of the Nominal to Numerical operator with specific parameter settings to convert the nominal attribute "NAMA" into a numerical form. Fig. 8 below shows the parameters for the Nominal to Numerical operator.



Parameters

Nominal to Numerical

attribute filter type ✓ single

attribute NAMA

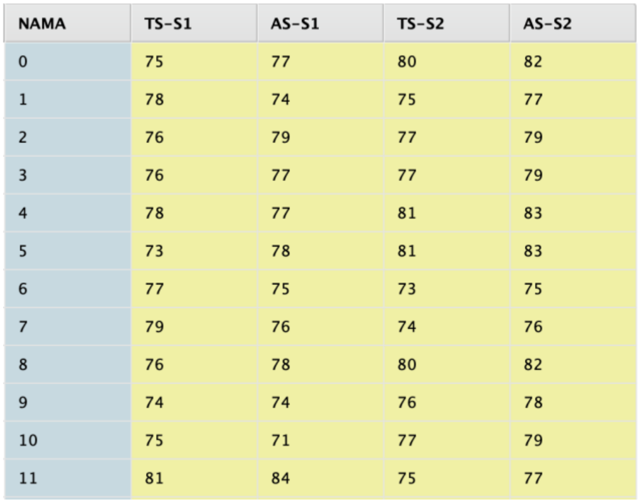
□ invert selection

☑ include special attributes

coding type unique integers

Fig. 8: Parameters of the Nominal to Numerical Operator.

Fig. 8 shows the configuration of the Nominal to Numerical operator. In this configuration, the "Attributes Filter Type" is set to "Single," meaning only one specific attribute will be processed in this conversion. The selected attribute is "NAMA," so the categorical values in the "NAMA" column will be converted into numerical form. The "Coding Type" is set to "Unique Integer," which will convert each unique category in the "NAMA" attribute into a different integer number. This transformation ensures that the data can be optimally processed by machine learning models, which require numerical data, allowing the "NAMA" attribute to be used in the clustering process. The result of the data transformation can be seen in Fig. 9 below.



NAMA	TS-S1	AS-S1	TS-S2	AS-S2
0	75	77	80	82
1	78	74	75	77
2	76	79	77	79
3	76	77	77	79
4	78	77	81	83
5	73	78	81	83
6	77	75	73	75
7	79	76	74	76
8	76	78	80	82
9	74	74	76	78
10	75	71	77	79
11	81	84	75	77

Fig. 9: Numeric Representation of the Name Column.

Fig. 9 shows the result of transforming the "NAMA" column into a numerical representation after the conversion process using the Nominal to Numerical operator. In this figure, the categorical values in the "NAMA" column have been changed to unique integer values for each distinct category. This process allows the data, which was previously in text or categorical form, to be used in machine learning models, such as clustering, which require numerical data. Each name in the "NAMA" column is now represented by a different number, according to the Unique Integer setting in the operator's configuration.

4.4. Data Minig

In this study, the K-Means algorithm is used to cluster students at Madrasah Tsanawiyah Islamiyyah Yami Waled based on their mathematics scores. After the data undergoes preprocessing and transformation, the K-Means algorithm is applied to group students based on the similarity of their average mathematics scores. The result of this process is the formation of student groups that can be used to design more targeted learning strategies.

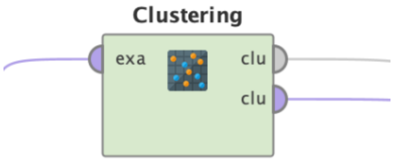


Fig. 10: K-Means Algorithm.

Fig. 10 shows the use of the K-Means algorithm operator to cluster the mathematics scores of students at Madrasah Tsanawiyah Islamiyyah Yami Waled. In this process, the Cluster Distance Performance operator is used to evaluate the quality of the clusters and determine the best number of clusters (K). In this case, the value of K is set to 6 to observe the possible variations in clustering that may provide the optimal result. Fig. 11 below shows the Cluster Distance Performance operator used:

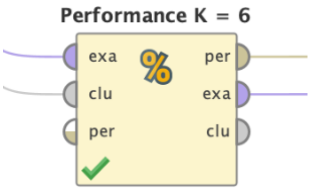


Fig. 11: Cluster Distance Performance Operator.

Fig. 11 shows the Cluster Distance Performance operator in the K-Means algorithm, which is used to evaluate the distance between data points in each cluster and how well each cluster represents the data. This operator works by measuring the average distance between data points within a cluster and the cluster's centroid. The smaller the distance between data points and the centroid, the better the quality of the formed cluster. Therefore, this operator helps determine the optimal number of clusters by identifying variations in the distance between clusters to find a more structured and organized division.

After evaluating the cluster quality using the Cluster Distance Performance operator, the next step is the use of the Sub Process operator. This operator is responsible for breaking down the K-Means algorithm into smaller, more specific processes, such as viewing each iteration of the cluster to find the Average Silhouette Width to determine the best cluster. By using the Sub Process operator, each phase of the clustering process can be carried out in a more modular and organized manner, ensuring that each sub-task can be executed separately before being combined to achieve the final result. Fig. 12 below shows the Sub Process operator.

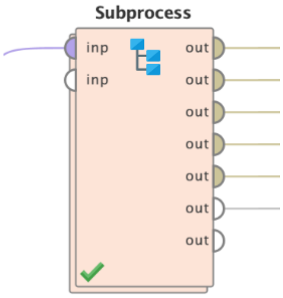


Fig. 12: Sub Process Operator.

Fig. 12 shows the Sub Process operator used to break down the clustering process into smaller, more specific tasks, allowing each phase of the clustering to be carried out more modularly and efficiently.

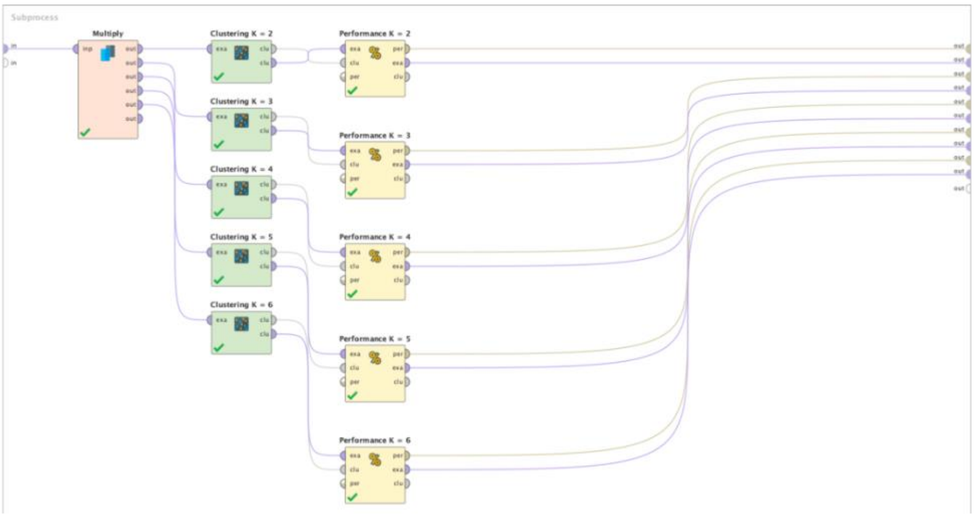


Fig. 13: Evaluation of Each Cluster.

Next, in the Sub Process operator, data input is duplicated (multiplication) and then passed through various clustering stages with different values of K, ranging from K=2 to K=6. Each clustering process generates an output in the form of cluster results (clu), which are then evaluated through the Performance stage for each value of K used. In the Performance evaluation process, several evaluation metrics are employed to assess the quality of the clustering results for each K value. Each K value will be compared with the others to determine the optimal result in selecting the most appropriate number of clusters for the data.

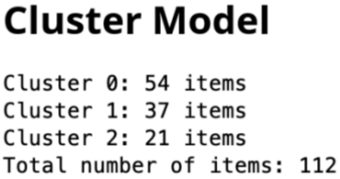
4.5. Evaluation and Interpretation

At this stage, an evaluation of the clustering results is conducted to determine the optimal number of clusters (K), using the Davies-Bouldin Index (DBI) as an indicator of cluster quality. The DBI results for each K value from the previous process are presented in Table 4.3 below to provide an overview of the clustering performance.

Table 3: Davies-Bouldin Index (DBI) for Each Cluster

K	*Davies Bouldin Index*
2	1.097
3	0.893
4	1.141
5	1.077
6	1.023

Table 3 shows the Davies-Bouldin Index (DBI) values for each number of clusters (K) tested. Based on the table, the best cluster is achieved at K=3 with a value of 0.893, indicating very good cluster separation.

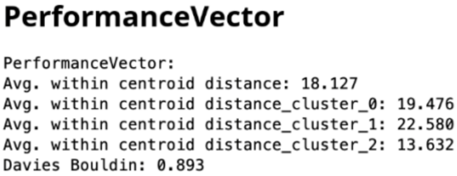


Cluster Model

```
Cluster 0: 54 items
Cluster 1: 37 items
Cluster 2: 21 items
Total number of items: 112
```

Fig. 14: Cluster Model K = 3.

Fig.14 shows the distribution of data points in each cluster formed from the clustering process with K=3, where a total of 112 items are divided into 3 clusters. The cluster with the most items is Cluster 0 with 54 items, while the cluster with the least items is Cluster 2 with only 21 items, and Cluster 1 contains 37 items. This distribution gives an insight into the variation in the number of data points in each cluster and helps assess the quality of the cluster separation in the clustering process.



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 18.127
Avg. within centroid distance_cluster_0: 19.476
Avg. within centroid distance_cluster_1: 22.580
Avg. within centroid distance_cluster_2: 13.632
Davies Bouldin: 0.893
```

Fig. 15: Performance Vector.

Fig. 15 shows the Performance Vector metric of the clustering results, which includes the average distance of each cluster to its respective centroid and the Davies-Bouldin Index (DBI). The average distance between points in each cluster shows how closely the data points in that cluster are grouped around the centroid. For example, Cluster 1 has the highest average distance (22.580), indicating that the data in this cluster is more spread out compared to other clusters. In contrast, Cluster 2 has an average distance of 22.580, which suggests that this cluster likely has data points that are closer to the centroid compared to Clusters 0 and 1.

A Davies-Bouldin Index (DBI) value of 0.893 indicates the overall quality of the clusters. The lower the DBI value, the better the clustering results, as DBI measures the ratio of the distance between clusters to the internal distance within the clusters. A DBI value below 1 indicates that the separation between clusters is good, although there is still room for improvement if the DBI value can be further reduced.

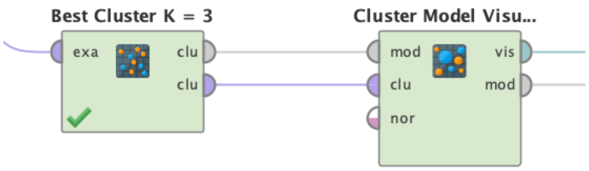
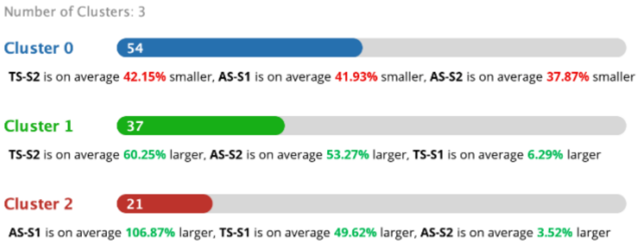


Fig. 16: Best Cluster & Cluster Model Visualizer Operator

Fig. 16 shows the result of the best cluster and the Cluster Model Visualizer operator. After the clustering process, the next step is to visualize the results of the best cluster.



Number of Clusters: 3

Cluster 0 54

TS-S2 is on average **42.15%** smaller, **AS-S1** is on average **41.93%** smaller, **AS-S2** is on average **37.87%** smaller

Cluster 1 37

TS-S2 is on average **60.25%** larger, **AS-S2** is on average **53.27%** larger, **TS-S1** is on average **6.29%** larger

Cluster 2 21

AS-S1 is on average **106.87%** larger, **TS-S1** is on average **49.62%** larger, **AS-S2** is on average **3.52%** larger

Fig. 17: Visualization of Cluster 1 - 19.

Fig. 17 shows the distribution of the average size differences for three main parameters (AS-S1, AS-S2, and TS-S2) across the 3 clusters formed. Each cluster has a different number of members, such as Cluster 0 with 54 members and Cluster 2 with 21 members, reflecting the variation in data patterns. The average pattern differences for each parameter provide an insight into the unique characteristics of each cluster. For instance, in Cluster 0, TS-S2 is 42.15% smaller, AS-S1 is 41.93% smaller, and AS-S2 is 37.87% smaller. In contrast, in Cluster 2, AS-S1 is 106.87% larger, TS-S1 is 49.62% larger, and AS-S2 is 3.52% larger. This analysis provides a deeper understanding of the characteristics of each cluster, supporting the understanding of data distribution patterns and the differences in parameter sizes among the groups.

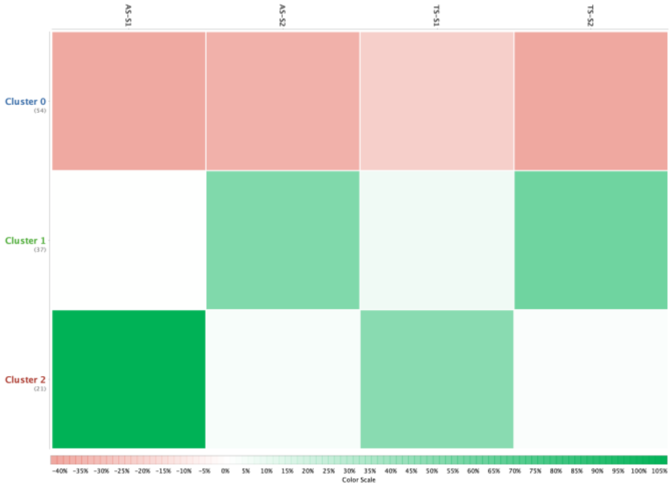
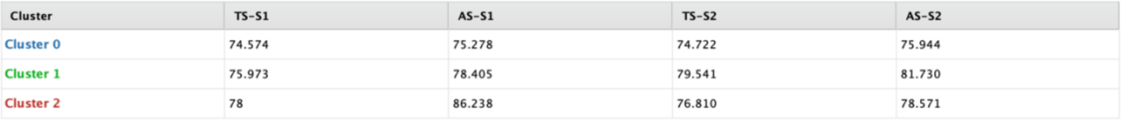


Fig. 18: Heatmap Visualization.

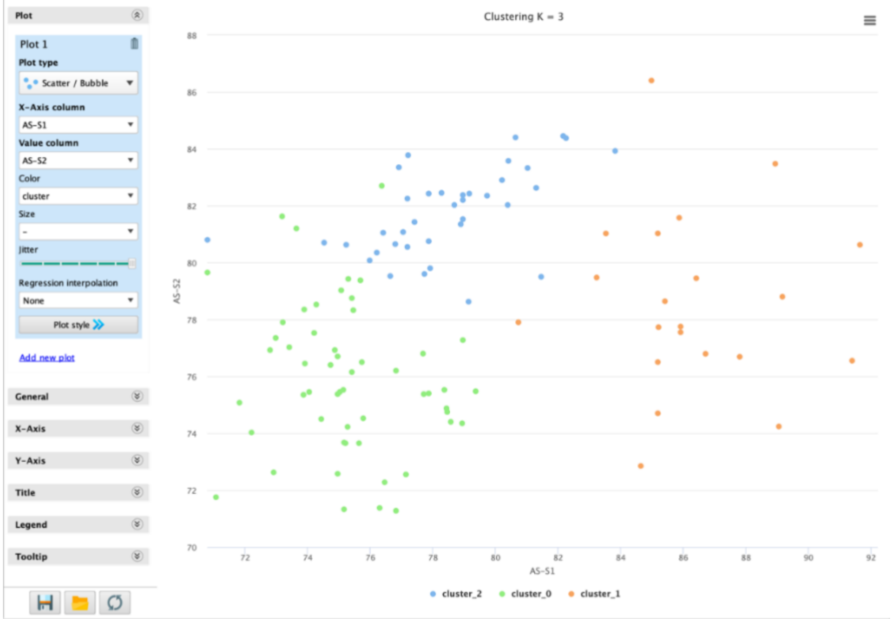
Fig.18 shows the visualization of the average size differences for three main parameters (AS-S1, AS-S2, TS-S1, and TS-S2) across the 3 clusters. The colors in this matrix represent the percentage of size differences, with green gradient indicating positive increases, and red gradient indicating decreases. In Cluster 0, red dominates, indicating that the parameters AS-S1, AS-S2, and TS-S2 are generally smaller compared to other clusters. On the other hand, Cluster 1 is dominated by green, showing an increase in these three parameters. Meanwhile, Cluster 2 shows significant differences with a stronger green dominance in AS-S1, indicating that this parameter has the highest increase compared to the other clusters.



Cluster	TS-S1	AS-S1	TS-S2	AS-S2
Cluster 0	74.574	75.278	74.722	75.944
Cluster 1	75.973	78.405	79.541	81.730
Cluster 2	78	86.238	76.810	78.571

Fig. 19: Centroid of Each Cluster.

Fig. 19 shows the average values of four main parameters (TS-S1, AS-S1, TS-S2, and AS-S2) for each cluster formed. In Cluster 0, the average values for TS-S1 and TS-S2 are relatively lower compared to the other clusters, with values of 74.574 and 74.722, respectively. The average values for AS-S1 and AS-S2 in this cluster are also quite low, at 75.278 and 75.944, respectively. Cluster 1 has higher average values for most parameters, especially AS-S2, which reaches 81.730, the highest value among all clusters. Additionally, the values for TS-S1 and TS-S2 in Cluster 1 are also relatively high, at 75.973 and 79.541, respectively. In Cluster 2, AS-S1 shows the highest average value, 86.238, indicating that this parameter dominates the cluster. Additionally, the average values for AS-S2 and TS-S2 in Cluster 2 are also relatively high, at 78.571 and 76.810, respectively. This data provides insights into how each parameter contributes differently to the characteristics of each cluster, helping to understand the average value distribution patterns within the data groups.



Plot 1
Plot type: Scatter / Bubble
X-Axis column: AS-S1
Value column: AS-S2
Color: cluster
Size: -
Jitter
Regression interpolation: None

Fig. 20: Scatter Plot Visualization.

Fig. 20 shows the scatter plot visualizing the data distribution for the parameters TS-S1 (horizontal axis) and AS-S1 (vertical axis), with a focus on a specific cluster, Cluster 2. Each point on the plot represents an individual data point that is grouped based on the clustering results, with different colors indicating cluster membership. From the plot, the data distribution pattern shows the relationship between TS-S1 and AS-S1 values. Most of the points fall within a close range of values, reflecting the uniformity of distribution in a specific cluster.

In this study, the K-Means algorithm provides a clear illustration of the academic achievement clustering of students at Madrasah Tsanawiyah Islamiyyah YAMI WALED. Three main clusters were formed based on differing academic scores. Cluster 0, with 54 members, represents a group with relatively low academic performance, indicated by lower average scores across various parameters compared to the other clusters (e.g., TS-S2 is on average 42.15% lower). Cluster 1, consisting of 37 students, shows better performance with higher average scores, especially on TS-S2, which is 60.25% higher than the overall average. Cluster 2, containing 21 students, has the highest academic achievement with AS-S1 reaching 106.87% above average, making it the group with the best academic performance. Additionally, the Davies-Bouldin Index (DBI) value of 0.893 for this clustering result indicates good cluster quality. A DBI value lower than 1 suggests that the separation between clusters is fairly clear, although there is still room for improvement. Based on these findings, it can be concluded that the K-Means algorithm successfully grouped students into three categories of academic performance: high achievers (Cluster 2), moderate achievers (Cluster 1), and students needing improvement (Cluster 0).

This study demonstrates that the K-Means algorithm is effective in detecting patterns among students. The findings are consistent with those of [10] and [8], which also concluded that K-Means is effective in clustering numerical data, particularly when applied to educational data. [10] emphasized how the algorithm can provide significant results in selecting students with higher potential for personalized educational interventions. [8] supports this conclusion, highlighting that K-Means can deliver clear clustering, separating students based on performance indicators such as academic scores and participation in school activities.

Furthermore, this study shows that K-Means delivers more optimal clustering results compared to some other algorithms, a finding not always evident in previous research. In studies by [7] and [4], for example, it was found that K-Means was less optimal in datasets with complex category variations or unevenly distributed data. However, in this study, optimal results were achieved because the variables used, such as AS and TS scores, as well as extracurricular participation, had a relatively consistent data distribution across semesters. This indicates that K-Means is more suitable when applied to datasets that are uniform and structured, which allows the algorithm to form clusters with clearer boundaries.

The application of K-Means in this study also supports the approach of [5] and [13], who recommend using this algorithm in educational environments to understand the relationship between academic performance and student involvement in extracurricular activities. In this context, the study demonstrates that clustering based on academic scores plays a vital role in identifying students who need additional support, as well as those with potential for specialized enrichment programs. The study also shows that K-Means has potential as a strategic analysis tool for clustering high-achieving students, particularly for educational institutions with structured data, such as semester grades and student attendance data. By comparing this study with [6] and [9], we can conclude that K-Means is a relevant tool for institutions with numerical data to design student development programs according to their performance clusters.

Overall, this study confirms the effectiveness of K-Means in mapping high-achieving student profiles and providing valuable clustering results for educational strategy planning. Thus, the algorithm plays an important role in offering a better understanding of students' academic conditions, providing a strong foundation for decision-making in supporting student development at Madrasah Tsanawiyah Islamiyyah Yami Waled.

5. Conclusion

As a result of the research conducted, the author summarizes the main findings based on the research objectives and the answers to the problem formulation. The following conclusions are drawn to provide a clear overview of the implementation and analysis of the K-Means algorithm in improving the clustering model for high-achieving students at Madrasah Tsanawiyah Islamiyyah Yami Waled. Firstly, the study successfully implemented the K-Means algorithm to cluster students at Madrasah Tsanawiyah Islamiyyah Yami Waled based on their odd and even semester mathematics scores. The clustering results revealed three main clusters: low achievers (Cluster 0, 54 students), moderate achievers (Cluster 1, 37 students), and high achievers (Cluster 2, 21 students). This process was carried out through the Knowledge Discovery in Databases (KDD) stages, which included data selection, preprocessing, transformation, and result analysis.

Secondly, the clustering results provide a clear picture of the distribution of student abilities, with Cluster 0 representing students with low average scores, while Cluster 2 includes those with high scores. This information helps the institution in developing more personalized and targeted teaching strategies, such as giving special attention to low-performing students and providing greater challenges to high-performing students. Thus, the K-Means algorithm proves to be effective in supporting data-driven strategic decision-making to enhance the quality of education at Madrasah Tsanawiyah Islamiyyah Yami Waled.

References

[1] F. P. Dewi, P. S. Aryni, and Y. Umaidah, “Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 2, pp. 111–121, 2022, doi: 10.14421/jiska.2022.7.2.111-121.

[2] E. A. Saputra and Y. Nataliani, “Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means,” *J. Inf. Syst. Informatics*, vol. 3, no. 3, pp. 424–439, 2021, doi: 10.51519/journalisi.v3i3.164.

[3] Haris Kurniawan, Sarjon Defit, and Sumijan, “Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal,” *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 2, pp. 80–89, 2020, doi: 10.52158/jacost.v1i2.102.

[4] A. Nugraha, O. Nurdiawan, and G. Dwilestari, “PENERAPAN DATA MINING METODE K-MEANS CLUSTERING UNTUK ANALISA PENJUALAN PADA TOKO YANA SPORT,” 2022.

[5] J. Hutagalung, “Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 1, pp. 606–620, 2022, doi: 10.35957/jatisi.v9i1.1516.

[6] A. Bahauddin, A. Fatmawati, and F. Permata Sari, “Analisis Clustering Provinsi Di Indonesia Berdasarkan Tingkat Kemiskinan Menggunakan Algoritma K-Means,” *J. Manaj. Inform. dan Sist. Inf.*, vol. 4, no. 1, pp. 1–8, 2021, doi: 10.36595/misi.v4i1.216.

[7] Y. Filki, “Algoritma K-Means Clustering dalam Memprediksi Penerima Bantuan Langsung Tunai (BLT) Dana Desa,” *J. Inform. Ekon. Bisnis*, vol. 4, pp. 166–171, 2022, doi: 10.37034/infeb.v4i4.166.

[8] S. Dewi, S. Defit, and Y. Yuhandri, “Akurasi Pemetaan Kelompok Belajar Siswa Menuju Prestasi Menggunakan Metode K-Means,” *J. Sistim Inf. dan Teknol.*, vol. 3, pp. 28–33, 2021, doi: 10.37034/jsisfotek.v3i1.40.

[9] N. Mirantika, “Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat,” *Nuansa Inform.*, vol. 15, no. 2, pp. 92–98, 2021, doi: 10.25134/nuansa.v15i2.4321.

[10] T. Asy Aria, M. Julkarnain, and F. Hamdani, “KLIK: Kajian Ilmiah Informatika dan Komputer Penerapan Algoritma K-Means Clustering Untuk Data Obat,” *Media Online*, vol. 4, no. 1, pp. 649–657, 2023, doi: 10.30865/klik.v4i1.1117.

[11] M. Djaka Permana, A. Lia Hananto, E. Novalia, B. Huda, and T. Paryono, “Klasterisasi Data Jamaah Umrah pada Tanurmutmainah Tour Menggunakan Algoritma K-Means,” *J. KomtekInfo*, vol. 10, pp. 15–20, 2023, doi: 10.35134/komtekinfo.v10i1.332.

[12] V. Ramadhan and Apriade Voutama, “Clustering Menggunakan Algoritma K-Means Pada Penyakit ISPA di Puskesmas Kabupaten Karawang,” *J. Pendidik. dan Konseling*, vol. 4, no. 5, pp. 462–473, 2022.

[13] C. A. Sugianto, A. H. Rahayu, and A. Gusman, “Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Cigugur Tengah,” *J. Inf. Technol.*, vol. 2, no. 2, pp. 39–44, 2020, doi: 10.47292/joint.v2i2.30.

[14] A. Sulistiyawati and E. Supriyanto, “Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan,” vol. 15, no. 2.

[15] H. Syahputra, “Clustering Tingkat Penjualan Menu (Food and Beverage) Menggunakan Algoritma K-Means,” *J. KomtekInfo*, vol. 9, pp. 29–33, 2022, doi: 10.35134/komtekinfo.v9i1.274.

[16] I. Virgo, S. Defit, and Y. Yuhandri, “Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering,” *J. Sistim Inf. dan Teknol.*, vol. 2, pp. 23–28, 2020, doi: 10.37034/jsisfotek.v2i1.17.

[17] D. Zakiyah, N. Merlina, and N. A. Mayangky, “Penerapan Algoritma K-Means Clustering Untuk Mengetahui Kemampuan Karyawan IT,” *Comput. Sci.*, vol. 2, no. 1, pp. 59–67, 2022, doi: 10.31294/coscience.v2i1.623.

[18] K. Gustipartsani, N. Rahaningsih, R. Danar Dana, and I. Yulia Mustafa, “Data Mining Clustering Menggunakan Algoritma K-Means Pada Data Kunjungan Wisatawan Di Kabupaten Karawang,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3595–3601, 2024, doi: 10.36040/jati.v7i6.8282.