Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Improving Student Achievement Clustering Model Using K-Means Algorithm in Pasundan Majalaya Vocational School

Sopian Abdul Mukhsyi[1*], Ade Irma Purnamaari[2], Agus Bahtiar[3], Kaslani[4]

[1,2]*Informatics Engineering, STMIK IKMI Cirebon, Indonesia*
[3]*Information System, STMIK IKMI Cirebon, Indonesia*
[4]*Computerized Accounting, STMIK IKMI Cirebon, Indonesia*
sopianabdulm47@gmail.com[1*], irma2974@yahool.com[2], agusbahtir038@gmail.com[3], kaslani@ikmi.ac.id[4]

Abstract

This study analyzes and enhances the student achievement clustering model at SMK Pasundan Majalaya using the K-Means algorithm. The Knowledge Discovery in Databases (KDD) method and RapidMiner AI Studio 2024.1.0 were used to process data from 125 students based on 15 metrics, including academic scores and attendance rates. For group evaluation, the Elbow method and Davies-Bouldin Index (DBI) were employed. The results showed optimal clustering with 2 groups and a DBI value of 0.893. Analysis results revealed significant differences in characteristics between the two groups. Cluster_1 consists of 38 students and has lower score patterns (60-80), with attendance rates of 94-100%, and a positive correlation between attendance and academic achievement. On the other hand, Cluster_0 consists of 86 students and shows higher score patterns (67.5-87.5), with attendance rates of 80-100%, and demonstrates a positive correlation between attendance and academic achievement. Schools can use this clustering model to create learning approaches that are better suited to each student group.

Keywords: K-Means Algorithm, Clustering, Davies-Bouldin Index, Student Achievement, Knowledge Discovery in Databases (KDD)

1. Introduction

Many aspects of human life have changed due to rapid advances in informatics. Today, innovations in many areas, such as education, health, economics and governance, depend on information and communication technology (ICT). The use of technology in education, in particular, has opened up new opportunities to improve education management and learning quality. Educational data analytics, one of the most important applications of informatics, has become a very useful tool to improve our understanding of the learning process and the factors that influence it. These developments have not only changed the way we view and manage education, but have also opened up avenues to improve the quality of learning.

Educational data analysis has enormous potential, but there are still many major problems when using it in primary schools. Processing and interpreting student attendance data effectively to produce a relevant picture of academic achievement is one of the main challenges. It is very difficult for primary education to find attendance patterns that correlate with academic achievement, which in turn hinders their ability to make targeted interventions. There are no analytical models tailor-made for the primary education context, which exacerbates this problem. This is because most previous research has focused on higher education. The need for methods that can uncover the complex relationship between attendance and academic achievement that are easy to use and understand by school employees who may not be familiar with advanced data analysis is growing.

Previous research has explored various aspects of educational data analysis, but there are still significant gaps in the primary school context. The study conducted by [1] demonstrated the effectiveness of data mining techniques in analyzing student attendance patterns and their relationship with academic performance. However, this study focused on higher education and did not specifically address the unique dynamics that exist at the primary school level. Meanwhile, [2] conducted a comprehensive review of machine learning techniques for predicting student academic performance, but again, the majority of the studies analyzed were from secondary and higher education contexts. Applying an improved K-means algorithm for student achievement analysis, shows the potential of clustering techniques in educational contexts. However, its application is limited to relatively homogeneous datasets and does not consider the complexity of long-

term attendance data common in primary schools. These studies, while providing a valuable foundation, highlight the need for a more specific and contextually relevant approach for analyzing student achievement data at the primary school level.

The main objective of this research is to develop and evaluate a data analysis model based on the K-means clustering algorithm that can improve the understanding of the relationship between students' attendance and their academic achievement at the primary school level. Specifically, this research aims to: identify student attendance patterns that correlate with certain levels of academic achievement, develop a predictive model that can help identify students who are at risk of declining academic achievement based on their attendance patterns, provide an analytical tool that can be used by educators and school administrators to take preventive actions and targeted interventions. The significance of this research lies in its potential to fill the gap in the literature regarding educational data analysis at the primary school level, as well as providing practical contributions in the form of analytical models that can be implemented to improve the quality of education and support data-driven decision-making in primary schools [3].

The methodological approach used in this study combines data mining techniques, specifically the K-means clustering algorithm, with statistical analysis and data visualization. Attendance and academic achievement data at Pasundan Majalaya Vocational High School will be collected, processed, and analyzed using the K-means algorithm to identify clusters of students with similar attendance and achievement characteristics. The process will involve several key stages, including data preprocessing to handle missing values and outliers, data normalization to ensure comparability between variables, and clustering parameter optimization to improve the accuracy of the results. The clustering results will be validated using cluster evaluation metrics such as Silhouette Score and Calinski-Harabasz Index to ensure the quality and stability of the resulting clusters. Post-hoc analysis will be conducted to interpret the characteristics of each cluster and identify significant patterns that emerge. This approach will not only yield meaningful insights into the relationship between attendance and academic performance, but will also generate a model that can be used for future prediction and decision-making.

The implications of the results of this study are expected to make significant contributions to educational practice and primary school management. By understanding the relationship between attendance patterns and academic achievement, schools can develop more effective intervention strategies to improve students' attendance and, in turn, their academic achievement. The data analysis model developed in this study can serve as an early warning system to identify students who are at risk of declining academic achievement, enabling timely preventive action. Furthermore, the results of this study can serve as a basis for the development of evidence-based education policies, refining the student attendance management system at the primary school level. From a theoretical perspective, this research contributes to the development of analytical models tailored to the specific needs of primary education, expanding the application of data mining techniques in this context. This may pave the way for further research on the integration of advanced data analytics in basic education management. In practical terms, the findings of this research could encourage the adoption of data-driven approaches in educational decision-making, improving the efficiency of resource allocation and ultimately contributing to improving the overall quality of basic education. The long-term potential of this research includes the development of an analytics platform that can be adopted by a wide range of stakeholders.

2. Literature Review

2.1. Data Mining

Data Mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [2].Data mining is the process of identifying data that has potential usefulness so that in the end it is easy to interpret. The task of data mining is divided into two, namely supervised learning and unsupervised learning. Supervised Learning requires modeling to perform analysis. Examples of Supervised Learning are Classification, Statistical Regression, and Association Rule, while Unsupervised Learning is learning that is not supervised by variables and does not make hypotheses before analysis. and does not make a hypothesis before analyzing analysis. The model will be created based on the results. Example of Unsupervised Learning is Clustering [3].

2.2. Clustering

One of the techniques known in Data Mining is clustering. The definition of scientific clustering in Data Mining is the grouping of a number of data or objects into clusters (groups) so that each cluster will contain data that is as similar as possible and different from objects in other clusters [4].The most widely used clustering method is the K-Means clustering method. The main drawback of this method is that the results are sensitive to the selection of the initial cluster center and the calculation of local solutions to achieve optimal conditions. Cluster analysis is a multivariate technique that has the main goal of grouping objects based on their characteristics. Cluster analysis classifies objects so that every object object that is closest in similarity to other objects is in the same cluster [5].

2.3. K-Means

K-means is one of the clustering algorithms that uses the partitioning method. K-means is a clustering algorithm that divides each data item into one cluster. The following are the steps in the K-means algorithm [6], [7], [8].
The research steps are as follows.

1. Determine the number of clusters (k) in the dataset.
2. Determine the centroids. ** The determination of the centroid value is done randomly at the initial stage, but at the iterative stage the following formula is used:

$$Vk=Nk\sum i/1NkXi$$

Where:
Vk = centroid of the k-th cluster
Xi = i-th data
Nk = number of objects that are members of the k-th cluster

3. Calculate the shortest distance to the centroid for each dataset. The centroid distance used is the Euclidean distance with the formula:

$$D_E = \sqrt{(x-s)^2 + (y-t)^2}$$

Where:
DE= Euclidean distance
i= number of objects
(x,y) = object coordinates
(s,t) = centroid coordinates

4. Group objects based on distance to nearest centroid.
5. Repeat step 2 until the centroid reaches the optimal value.

K-means has been widely used to analyze data distribution in various studies. For example, Kurniawan et al. (2023) used K-means to cluster the pharmaceutical data of Puskesmas, and the clustering results helped optimize pharmaceutical inventory management. Furthermore,

3. Research Methods

This research uses a quantitative method with a data mining approach, using the Knowledge Discovery in Databases (KDD) process to find patterns in student achievement data at Pasundan Majalaya. Data selection, preprocessing, transformation, data mining [9], and evaluation are important stages in the KDD process used. The Davies-Bouldin Index (DBI) method is used to evaluate the accuracy and clarity of the separation made between clusters. In the data mining stage, the K-Means algorithm was used to cluster students based on their performance metrics, such as academic grade point average, attendance, and participation rate. The result yields several clusters that represent groups of students with comparable patterns or achievement metrics. The following figure shows the research process conducted:

Figure 1: Research Methods

Table 1: Activity Description Research Methods

Stages	Activity	Activity Description
1. Selection	Data Collection	Collecting daily attendance data and academic grades of Pasundan Majalaya Vocational High School students. Ensure that the data collected covers all students within a certain period of time and has complete and consistent information.
2. Pre-Procession	Pre-processing	Performed data cleaning to ensure quality, including removing duplicate data and normalizing attendance and academic grades to have a uniform scale to facilitate further analysis.
3. Data Transformation	Preparing Data for the analysis process	Transforming the cleaned data into a format suitable for the K-Means algorithm, including data normalization and selection of important variables such as number of attendance, absenteeism, and grade point average. into a format suitable for the K-Means algorithm, including data normalization and selection of important variables such as number of attendance, absenteeism, and grade point average.
4. Data Mining	K-Means Clustering Method.	Implement the K-Means Clustering algorithm to group student achievement based on attendance and academic achievement data. Run the algorithm with a predetermined number of clusters to generate different groups of students.
5. Evaluation	Evaluation of results.	Evaluate the clustering results using metrics such as Silhouette Score to assess the quality of the clustering. Examining the patterns formed in each cluster to understand

			the relationship between attendance and student achievement.
6.	Knowledge	Results	Data Mining Process decisions and actions in research from rapidminer results and DBI value results

3.1. Data Source

This study was conducted at Pasundan Majalaya Vocational High School (SMK), using data from the school's official report card. This data was taken from even semester records in the 2024/2025 school year, The dataset consists of 125 student data which includes academic grades such as Mathematics, Indonesian Language, Science, and Social Studies. The data used in this study comes from primary data collected directly from the school without conducting field data collection such as interviews or surveys. In addition, secondary data comes from references from related literature and journals on this issue, such as the application of the K-Means algorithm and student achievement clustering methods.

3.2. Population

Population is a generational field consisting of subjects or objects with certain numbers and characteristics that have been grouped determined by researchers to study and draw conclusions, this study comes from Tuberculosis (TB) cases in the Cirebon City Health Office work area. The population includes all TB patients recorded in this area during the time period specified in the study, namely Tuberculosis cases in 2024 which were recapitulated from January 01 to November 01, 2024.

3.3. Data Collation Techniques

Data collection will be conducted thoroughly and in accordance with a predetermined structure to ensure accurate data, by seeking information on student achievement at SMK Pasundan Majalaya. This includes academic grades, attendance rates. Before further processing, the data is validated to ensure its accuracy and completeness. Using the K-Means algorithm, the data was then categorized based on the range of academic grades and attendance rates. Through this process, students are grouped into high and low achievement categories, which facilitates analysis and helps make better decisions.

3.4. Data Analysis Techniques

In this study, the K-Means algorithm was used to group students based on student achievement. The first step involves collecting student achievement data. Then, the data is processed and normalized to ensure accuracy. Next, the K-Means algorithm is used to divide students into groups based on the similarity of their achievements. With the help of RapidMiner application, Knowledge Discovery in Database (KDD) approach was used to select the ideal number of clusters. The results of this analysis show that there are groups of students that can be used to implement more focused learning strategies and improve the quality of education in Pasundan Majalaya

4. Results and Discussion

4.1. Research Result

This study uses the k-means clustering algorithm to improve the student achievement clustering model at Pasundan Majalaya Vocational High School. Data processing was conducted using the Knowledge Discovery in Databases (KDD) method, assisted by RapidMiner AI Studio 2024.1.0 software. The main stages of the KDD method include data selection, data cleaning, data transformation, and analysis to produce appropriate clusters. With the k-means algorithm, student achievement data is grouped based on academic grades, attendance. The clusters created are divided into groups of students with high, medium, and low achievement. The clustering results can help schools make a more targeted learning approach and improve student achievement in the future.

4.1.1 Data Selection

The data used in this study was obtained directly from Pasundan Majalaya Vocational High School. The dataset totaled 125 students with 15 main attributes, namely student name, gender, number of academic grades, percentage of attendance, total days. This data was then processed using the k-means algorithm to improve the student achievement clustering model.

Table 2: Data Selection

Name Student	L/P	PABP	Pp	B.Indonesia	MTK	B.INGGRIS	Mapel_KK	Score_Number	S	I	A	Number_Attendance	Day Total_Attendance	Attendance Percentage
Cita Liana	P	80	70	52	88	58	71	74.83	0	1	0	44	45	98%
Claresya Syarah	P	78	75	34	87	60	70	76.17	0	0	0	45	45	100%
Maudy Zalsya Febiyanti	P	76	70	48	87	50	72	72.83	11	0	0	34	45	76%
Giza Ramadani	P	80	72	30	88	55	71	75.67	0	0	0	45	45	100%
Jihan Mardiyah	P	82	70	44	88	66	78	78	4	0	0	41	45	91%
Leni Marlina	P	78	70	40	87	62	72	76.5	0	0	0	45	45	100%

Lisna Meilani	P	78	72	56	88	55	76	75	2	1	0	42	45	93%
Luvita Salsabila	P	78	70	40	88	52	69	73.17	2	0	0	43	45	96%

4.1.2. Preprocessing/ Cleaning

At this stage, the collected data is processed to remove data that is incomplete, inconsistent, or contains noise. Only clean and ready-to-use data is then processed further. The first stage in data preprocessing is inserting the aim to replace missing values. An image of the Replace Missing Value Operator insertion process can be seen in Figure 2 below.

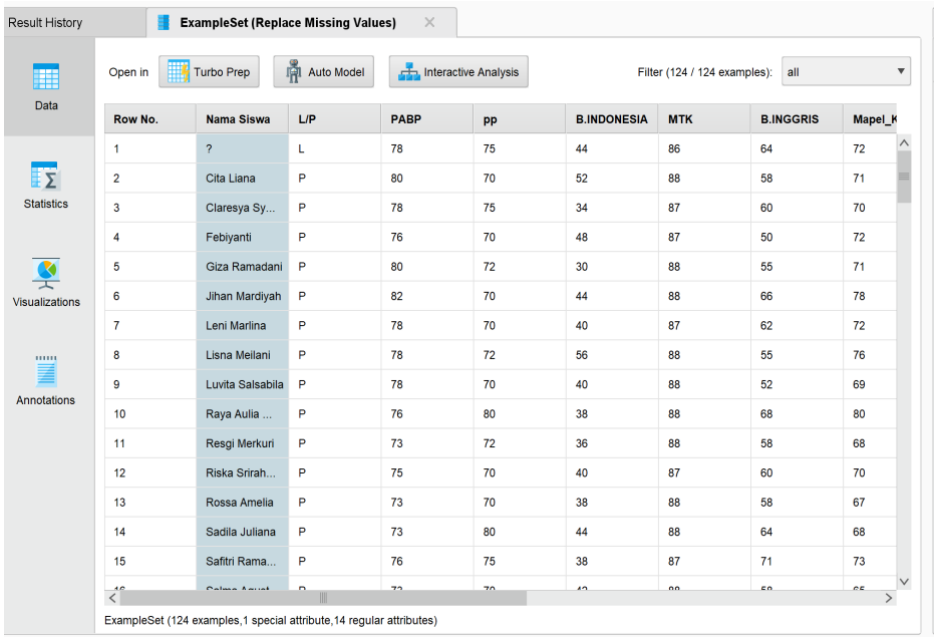

Figure 2: Data processing results

4.1.3. Transfomation

The data transformation process consists of two stages. First, the nominal to numerical operator is used to convert attributes to numerical form. By using this operator, data that was originally in nominal format will be converted into numerical format. This is necessary because the K-Means algorithm only has the ability to process data in numerical format. Before starting the process, make sure the filter type attribute is selected as "subset" because the student name and l/p columns are converted into numeric form with the coding type used using unique integers. Figure 4 below shows an illustration of the nominal to numeric operator entry process.

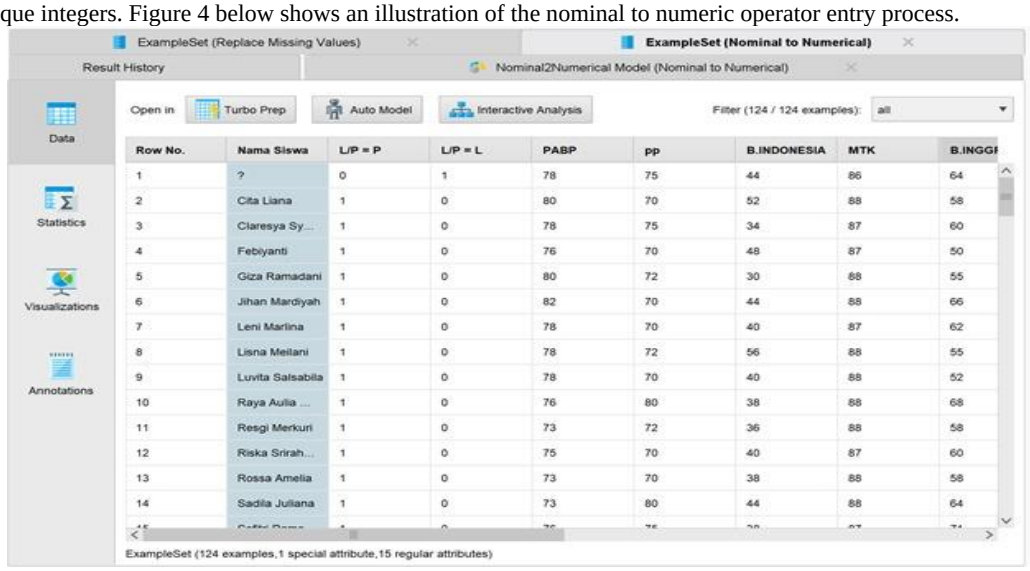

Figure 3: Transformation Result

4.1.4. Data Mining

In the data mining process, the first step is to use the K-Means algorithm to classify student achievement using Rapidminer. The algorithm is implemented in the process, with the determination of parameters that can be seen in the figure, namely k = 2 which is done a maximum

of 10 times. The image of the k-means clustering process can be seen in Figure 5 below.namely K = 2 which is done a maximum of 10 times the clustering. An image of the K-Means clustering process can be seen in Figure 5 below.

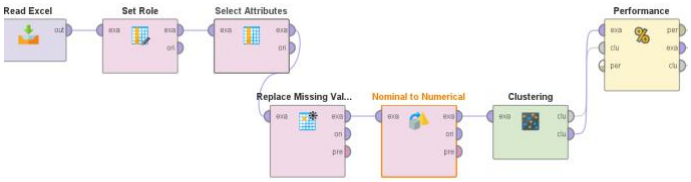

Figure 4: Clustering K-Means

Next is the stage of adding a performance operator involving the process of finding the distance value and DBI value to measure the performance of a communication system or device effectively. The picture of the performance stage can be seen in Figure 5 below.

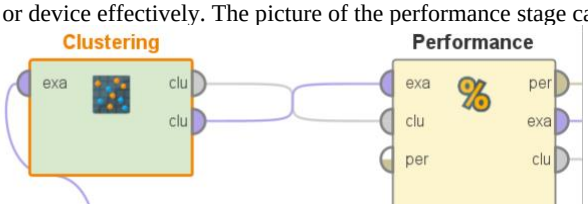

Figure 5: Performation

4.1.2 Interpretation / Evaluation

In the process of clustering data, model evaluation is performed to determine the best clusters. In this method, Davies-Bouldin Index (DBI) values are evaluated with lower DBI values indicating better cluster quality. In the applied K-means model, varying values of k were tested ranging from k=2 to k=10, and each value was evaluated to determine which DBI was the smallest and which cluster was formed. Below is Table 4.2 Experimental results of k=2 to k=10 using DBI evaluation.

Table 3: Evaluation DBI Results

Cluster	Number Of Cluster Members	DBI Value Result	Nilai Avg. within centroid distance
2	Cluster 0: 86 items Cluster 1: 38 items	0.893	483.963
3	Cluster 0: 26 Cluster 1: 45 Cluster 2: 53	0.981	320.986
4	Cluster 0: 41 Cluster 1: 18 Cluster2: 48 Cluster3: 17	1.022	262.469
5	Cluster 0 : 15 Cluster 1:17 Cluster 2: 14 Cluster 3: 16 cluster 4: 35	1.038	224.599
6	Cluster 0 : 39 Cluster 1 : 14 cluster 2 : 8 Cluster 3 : 34 cluster 4 : 15 Cluster 5 : 14	1.084	198.800
7	Cluster 0 : 16 Cluster 1 : 10 Cluster 2 : 13 cluster 3 : 13 Cluster 4 : 8 Cluster 5 : 31 Cluster 6 : 33	1.098	179.683
8	Cluster 0 : 24 Cluster 1 : 10 Cluster 2 : 13 Cluster 3 : 28 Cluster 4 : 9 Cluster 5 : 21 Cluster 6 : 7 Cluster 7 : 12	1.174	163.670
9	Cluster 0 : 36 Cluster 1 : 5 Cluster 2 : 2 Cluster 3 : 4	1.029	162.224

	Cluster 4 : 13 Cluster 5 : 14 Cluster 6 : 11 Cluster 7 : 7 Cluster 8 : 32		
10	Cluster 0 : 32 Cluster 1 : 5 Cluster 2 : 20 Cluster 3 : 7 Cluster 4 : 1 Cluster 5 : 27 Cluster 6 : 8 Cluster 7 : 13 Cluster 8 : 9 Cluster 9 : 2	0.984	151.110

4.1.3 Knowledge

The results of research using the K-Means clustering algorithm using Rapid miner tools and the results of evaluating the Davies-Bouildin Index and using the elbow method can be seen as follows.

a. Davies-Bouldin Index (DBI) value

Clustering using the k-means clustering algorithm with Rapidminer tools and DBI evaluation obtained the optimal and best cluster value at K = 2 with the smallest DBI value result of 0.893 with cluster_0 members: 86 items and Cluster 1: 38 items. This shows that clustering with two clusters produces the optimal cluster and has the best degree of separation. Can be seen in the figure 6 below.

Attribute	cluster_0	cluster_1
L/P = P	0.453	0.263
L/P = L	0.547	0.737
PABP	77.302	79.526
pp	72.186	82.842
B.INDONESIA	40.407	53.342
MTK	86.442	84.553
B.INGGRIS	58.988	74.579
Mapel_KK	71.535	73.553
JUMLAH_NILAI	406.872	448.395
S	0.791	0.342
I	0.640	0.316
A	0.070	0.158
Jumlah_kehadiran	43.523	44.184
Total_hari	45	45
Presentase Kehadiran	0.968	0.982

Figure 6: Model Cluster Results

b. Elbow Method

The Elbow method is used to see changes in the average distance within clusters or called the Avg. within centeroid distance value which measures when the number of clusters increases. The elbow graph is generated by entering the average within-cluster distance of each k into the graph/plot, then looking at the point where the graph starts to form an elbow. In this study, the elbow occurs at K=2 which indicates that cluster 2 is the optimal number of clusters. With the similarity of the DBI evaluation results and the Elbow method reinforces that two clusters are the best choice in this modeling. The Elbow method is used to see changes in the average distance within clusters or called the Avg. within centeroid distance value which measures when the number of clusters increases. The elbow graph is generated by entering the average distance within the cluster of each k into the graph/plot, then looking at the point where the graph starts to form an elbow. In this study, the elbow occurs at K=2 which indicates that cluster 2 is the optimal number of clusters. With the similarity of the DBI evaluation results and the Elbow method reinforces that two clusters are the best choice in this modeling. Can be seen in Figure 7 below.

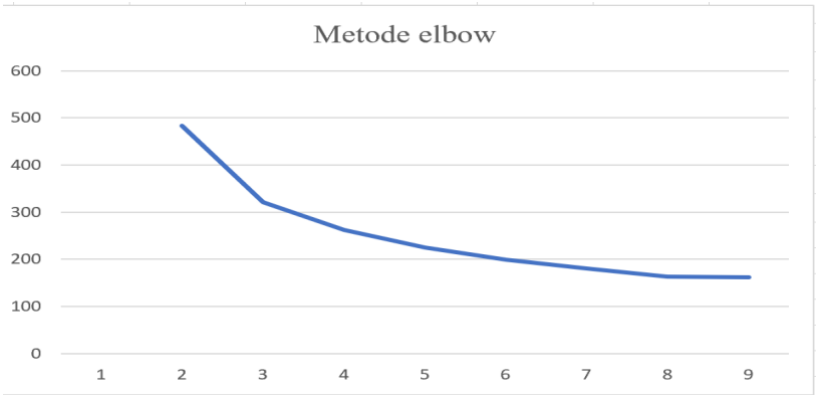

Figure 7: Method Elbow

c. Visualization

Can be seen in Figure 8 is a data graph generated from Rapidminer visualization which serves to identify student groups that can be used for further analysis such as decision making and health treatment. Can be seen in the image below.

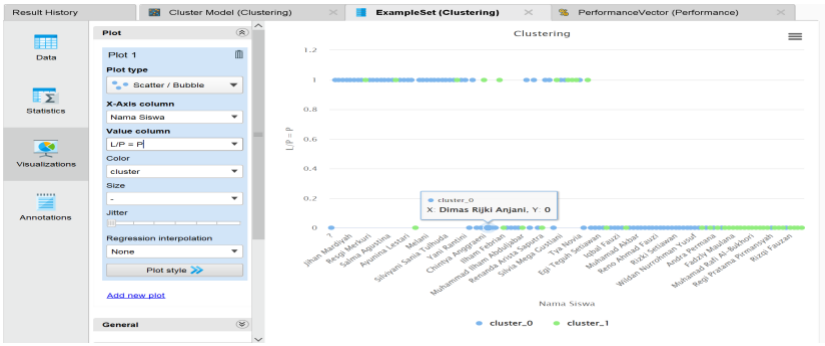

Figure 8: visualization result

From the visualization plot image above, it can be explained that the graph shows two groups that are given different colors. Cluster_0 with blue color shows the group with the majority of L/P = P values are at 0 in cluster_0, which indicates the dominance of male students. On the other hand, most of the L/P = P values are at 1, indicating that female students dominate this cluster. In addition, this distribution corresponds to the mean L/P = P values in the centroid table; the value for cluster_0 is 0.453 and for cluster_1 is 0.737, indicating that there is a difference in gender dominance in each cluster.

4.2 Discussion

In the discussion section, researchers outline the objectives that have been presented in Chapter 1, namely how to apply the k-means algorithm in clustering student achievement results and evaluate the effectiveness of the k-means algorithm in clustering student data using the Davies-Bouldin index (DBI) evaluation to determine the best cluster. The final result of this research is the analysis of clustering data using the k-means algorithm.

a) Grouping student achievement using the K-Means clustering algorithm.

The application of the k-means algorithm in clustering student achievement using Rapidminer with DBI evaluation, the smaller the DBI value, the better and more optimal the cluster formed. The optimal clustering is found at k = 2 with a DBI value of 0.893, formed into 2 clusters, cluster_0 and cluster_1 with a total of 125 people. Cluster_0: 86 people and Cluster_1: 38 people. The following table shows the results of the DBI value.

Table 4: DBI Results

K	DBI
2	0.893
3	0.981
4	1.022
5	1.038
6	1.084
7	1.098
8	1.174
9	1.029
10	0.984

b) Analysis results of student achievement grouping model using k-means algorithm

Based on the test results with the value of k = 2 to k = 10 as shown in table 4.8, it can be seen that the smallest Davie-Bouldin Index (BDI) value is obtained at k = 2. The smaller DBI value indicates that the clusters formed are more optimal, therefore, k = 2 is chosen as the optimal value for clustering student achievement data.

NAMA SISWA	L/P	PABP	pp	B.INDONESIA	MTK	B.INGGRIS	Mapel_KK	JUMLAH_NILAI	S	I	A	Jumlah_kehadiran	Total_hari	Presentase Kehadiran	CLASTER
Tania Juliani	P	73	73	88	87	58	85	77.33	0	1	0	44	45	97.78%	Claster_0
Yani Rantini	P	73	75	80	87	67	82	77.33	1	0	0	44	45	97.78%	Claster_0
Alfariz Eka Putra	L	73	75	84	88	58	75	75.5	2	3	0	40	45	88.89%	Claster_0
Alifia Shofia Lifa	P	73	71	84	87	55	71	73.5	0	1	0	44	45	97.78%	Claster_0
Andri Saputra Pratama	L	74	61	83	87	50	61	69.33	0	1	0	44	45	97.78%	Claster_0
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	
Anisa Nur Oktapiani	L	73	65	86	87	66	65	73.67	0	0	0	45	45	100.00%	Claster_0
Chintya Anggraeni	P	74	79	81	88	68	79	78.17	1	0	0	44	45	97.78%	Claster_0
Dimas Rijki Anjani	L	74	75	89	87	54	75	75.67	1	0	0	44	45	97.78%	Claster_0
Esal Pirmansyah	L	76	81	90	87	60	81	79.17	0	0	0	45	45	100.00%	Claster_0
Feisal Nur Agung	L	74	75	80	87	55	75	74.33	0	0	0	45	45	100.00%	Claster_0

Figure 9: culster_0 members

NAMA SISWA	L/P	PABP	pp	B.INDONESIA	MTK	B.INGGRIS	Mapel_KK	JUMLAH_NILAI	S	I	A	Jumlah_kehadiran	Total_hari	Presentase Kehadiran	CLASTER
Ikşan Sulaeman	L	78	60	40	89	57	80	67.33	0	0	0	45	45	97.78%	Claster_1
M. Fikri Hidayatulloh	L	75	78	58	89	62	68	71.67	0	0	1	44	45	97.78%	Claster_1
Moch Rifki Fauzi	L	78	75	36	86	53	70	66.33	1	0	0	44	45	93.33%	Claster_1
Much Rizki Nurahman	L	84	75	28	86	51	67	65.17	0	3	0	42	45	97.78%	Claster_1
Muhamad Akbar	L	75	80	48	86	64	68	70.17	0	0	1	44	45	97.78%	Claster_1
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	...
Muhamad Nizar	L	77	80	38	86	57	73	68.5	1	0	0	44	45	100.00%	Claster_1
Muhamad Riski	L	82	60	36	86	58	65	64.5	0	0	0	45	45	97.78%	Claster_1
Nasbiandra Algifari	L	78	78	44	86	53	68	67.83	0	0	0	45	45	100.00%	Claster_1
Raihan Andika Rachman	L	81	78	46	86	62	53	67.67	0	1	0	44	45	95.56%	Claster_1
Reno Ahmad Fauzi	L	85	78	44	86	64	65	70.33	0	0	0	45	45	97.78%	Claster_1

Figure 10: clusters_1 members

From the two images above showing K=2 data, the two clusters—Cluster_0 and Cluster_1—display differences in academic performance and student attendance rates. Students in this cluster have higher average scores, with the majority scoring above 75 in almost all subjects, especially Mathematics and English, and have excellent attendance rates, with percentages above 97%. In contrast, Cluster_1 consists of more diverse students with lower average scores; some students show significant weaknesses in subjects such as Mapel KK and Indonesian Language, with scores even below 50. Additionally, this cluster also has students with slightly lower attendance rates, such as 93.33%, which can impact their learning outcomes.

5. Conclusion

Based on the results of this research, the discussion that has been carried out, it can be concluded as follows:

1. By applying the K-Means algorithm it shows the ability to group students effectively based on their attendance data. Cluster_0 consists of 86 students with a lower attendance rate, and Cluster_1 consists of 38 students with a higher attendance rate. These two clusters have the smallest Davies-Bouldin Index (DBI) value, namely 0.893.
2. Based on data visualization analysis, it can be concluded that there are significant differences in the relationship patterns between attendance and academic scores in both student groups, where Cluster_0 shows a strong positive correlation with higher academic scores (67.5-87.5) and more varied attendance patterns (80-100%), while Cluster_1 displays a different pattern with lower academic scores (60-80) and a narrower attendance range (94-100%) without clear correlation, this indicates that attendance level is not always a primary determining factor in students' academic achievement, and there are other factors that may influence academic performance, especially in the Cluster_1 group.

6. Suggestions

Based on the research that has been done, several suggestions can be given for further research:

1. Based on the research results, schools must create learning strategies that are more appropriate to each group of students. For example, for students in Cluster_0 who have low attendance rates, a special mentoring program can be held which includes providing motivation, additional study guidance, or a more personalized approach to increase student attendance and achievement.
2. Cluster_0, where the majority of students are male, requires greater attention in terms of their motivation to learn and their involvement in academic activities. One way to improve their attendance and overall achievement is through intervention programs involving parents, teachers, and counselors.
3. Schools can consider collecting additional data, such as behavioral data, family background, and participation in extracurricular activities, to maximize the effectiveness of the K-Means algorithm and help discover other relevant patterns to improve the quality of education.

References

[1] M. Hedayetul, I. Shovon, and M. Haque, "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree," 2012. [Online]. Available: www.ijacsa.thesai.org

[2] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," 2023.

[3] M. Rafi, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," vol. 12, no. 2, pp. 121–129, 2020.

[4] Z. Sitorus and U. Asahan, "PENERAPAN DATA MINING UNTUK CLUSTERING PENDUDUK MISKIN DI KOTA TANJUNGBALAI MENGGUNAKAN METODE ALGORITMA K-MEANS," vol. 4307, no. 1, pp. 212–218, 2024.

[5] M. Miranda, N. Rahaningsih, and R. D. Dana, "Analisis Clustering Data Anak Balita di Posyandu Kampung Sukarame Menggunakan Algoritma K-Means," vol. 6, no. 1, pp. 136–141, 2024.

[6] M. Veronica, H. Effendi, and O. Saleh, "Clustering Tingkat Kedisiplinan Pegawai Pada Pengadilan Tinggi Palembang Menggunakan Algoritma K-Means," pp. 261–266, 2023.

[7] Ramadani, S., Ambarita, I., & Pardede, A. M. H. (2019). Metode K-Means untuk pengelompokan masyarakat miskin dengan menggunakan jarak kedekatan Manhattan City Dan Euclidean (Studi kasus kota binjai). *Inf. Syst. Dev*, *4*(2), 15-29.

[8] Aditya Putra Prananda, Pardede, A. M. H., & Rahmadani. (2024). Segmentation Algorithm K – Means Based On The Maturity Level Of Blueberries. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *3*(2), 584–589. https://doi.org/10.59934/jaiea.v3i2.433

[9] Pardede A. M. H. *et al* 2019 Implementation of Data Mining to Classify the Consumer's Complaints of Electricity Usage Based on Consumer's Locations Using Clustering Method *Journal of Physics: Conference Series* **1363** 12079