Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i1.1029 Published By DRPM-UBD

Clustering of High School Students Academic Scores Using the K-Means Algorithm

Chairunisa Azzahra[1], Sriani[2]

[1,2]Computer Science Study Program, State Islamic University of North Sumatra, Medan, Indonesia
Email: [1]nisaazzahra438@gmail.com, [2]sriani@uinsu.ac.id

Abstract

The clustering of student subject scores in senior high school is conducted using the K-Means Clustering algorithm. The issue addressed in this study is how to optimally group students based on their academic scores to help schools understand the distribution of student abilities. This clustering is essential as a foundation for evaluating and improving the learning system. The research methodology includes data collection and preprocessing, determining the optimal number of clusters using the Davies-Bouldin Index (DBI), and applying the K-Means Clustering algorithm. The analysis results indicate that the optimal number of clusters is three, with an average DBI value of 1.226. Cluster 0 is categorized as "very good" (46 students), Cluster 1 as "good" (70 students), and Cluster 2 as "less good" (51 students).The clustering results can be utilized for more targeted learning interventions and curriculum adjustments. Schools can implement remedial programs or additional classes for students in the "less good" cluster to improve their academic performance. Meanwhile, students in the "very good" cluster can be provided with advanced learning materials or opportunities to participate in academic competitions. Additionally, clustering outcomes provide valuable insights for refining teaching strategies, allocating resources more effectively, and personalizing learning approaches to suit each student's needs. Furthermore, these clustering results support academic decision-making by enabling educators and administrators to identify student performance trends and address potential learning gaps. This data-driven approach helps schools enhance overall educational quality by adapting teaching methods and policies based on empirical findings.

Keywords: student performance clustering, educational data mining, K-Means, clustering evaluation, Davies-Bouldin Index

1. INTRODUCTION

Education Foundation Medan is an institution that operates in the field of education. Education plays an essential role in developing children's potential to achieve safety and happiness. Furthermore, education is also a process of humanism, known as "humanizing humans," which requires respect for the fundamental rights of every individual. Students, in other words, are not machines that can be controlled at will; they are a generation that needs guidance and attention in their maturation process. The goal of education is to shape individuals

who are independent, critical thinkers, and possess good morals, not only to fulfil basic life needs but also to create more meaningful human beings [1], [2], [3].

The goal of this clustering is to group students based on their abilities to help the school improve student quality and identify elite classes. As the amount of data increases, clustering becomes less efficient, making a computer-based system necessary [4]. Fast and accurate clustering can enhance the quality of learning. Elite classes consist of above-average students, and determining their selection is crucial for the school's improvement. The process involves assessing students' performance against the minimum passing grade. Report card data from the 2013 curriculum shows variations in evaluation criteria [5]. The number of students in elite classes can vary yearly based on school development and student admissions. Clustering large student data efficiently requires a system to identify students for elite classes, which will improve learning outcomes. At , with its two majors (Science and Social Sciences), this study aims to determine the elite class composition, motivating high-achieving students to maximize their potential while encouraging lower-performing students to improve and join the elite class [6], [7].

The K-Means algorithm is part of Clustering Data Mining, where it is used to create new groups through the formation of clusters [8]. In the K-Means algorithm, the cluster formation process is based on the characteristics of each object present in the data. These characteristics are determined by the closest distance value between data points. The process in the K-Means algorithm is iterative, where iterations are performed to calculate the proximity between each data object. During the iterative process, the center point value is first determined as the basis for group assignment. Based on this value, data processing will be carried out for the entire dataset [9], [10].

A study by [7] demonstrated that applying K-Means clustering in a class clustering system effectively classified data. The iterative process of centroid distance calculation and cluster point determination helped save time in grouping students for elite classes. The web-based clustering information system produced flexible, accessible results for authorized users. The study At Junior High School involved 192 students, with 96 placed in elite classes and 96 not. User acceptance of the system was 97.56%, indicating its effectiveness [7].

Data mining is the process of gathering and analysing data to uncover valuable insights [11]. Its primary objective is to extract meaningful information from datasets. This field, which falls under computer science, emphasizes machine learning techniques. By leveraging data and information, data mining enables predictions and informed decision-making.[12]

The K-Means algorithm is a non-hierarchical clustering technique that begins by creating initial cluster partitions and iteratively refines them until minimal changes

occur. It groups similar data points together while separating those with distinct characteristics. The algorithm's objective is to minimize variation within clusters while maximizing differences between them. Known for its simplicity and ease of implementation, K-Means is widely used, particularly in Senior High Schools to medium-scale applications. It organizes data into k clusters based on centroids, though its effectiveness heavily depends on the chosen value of k and the initial centroid selection, which is often randomized [13], [14].

Several studies have applied the K-Means clustering algorithm in various fields. [15].categorized students' academic performance into four clusters, with Cluster 02 showing excellent performance (90 students, average score 87.01), Cluster 0 showing good performance (204 students, average score 81.51), Cluster 03 showing fairly good performance (123 students, average score 80.96), and Cluster 01 showing poor performance (326 students, average score 77.30) [7]. implemented K-Means clustering in determining elite student classes, achieving efficient classification of subgroups, saving time, and improving system performance. Study by [16] applied K-Means clustering to classify population density in Deli Serdang Regency, resulting in three clusters: very dense, dense, and moderate. Study by [17] used K-Means clustering to categorize smokers over the age of 15, identifying three distinct clusters. Study by [18] applied K-Means clustering to analyze discounts on Honda motorcycles, classifying vehicles into three clusters based on eligibility for discounts. These studies demonstrate the effectiveness of K-Means clustering in various applications, from education to population density and product pricing.

2. METHODS

In this study, data analysis was conducted using the K-means clustering method to group elite classes based on students' subject grades. The process began by selecting and collecting relevant grade data from school academic records. After preprocessing the data to ensure scale consistency and data integrity, K-means clustering was applied to group students based on the proximity of their grades. The analysis provided valuable insights into the distribution of students' academic performance, which can be used to develop more focused educational strategies. Figure 1 is the flowchart designed by the researchers.

The research begins with modelling using the K-Means Clustering algorithm, which is applied to student grade data to group students based on their performance patterns. In testing the K-Means method for clustering student subject grades to determine elite classes, the researcher starts by determining the optimal number of clusters. After randomly initializing the centroids, the K-Means iterative process groups students into clusters based on their proximity to the centroids of their grades. This paper illustrates how the K-Means clustering method is tested specifically for determining elite classes.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

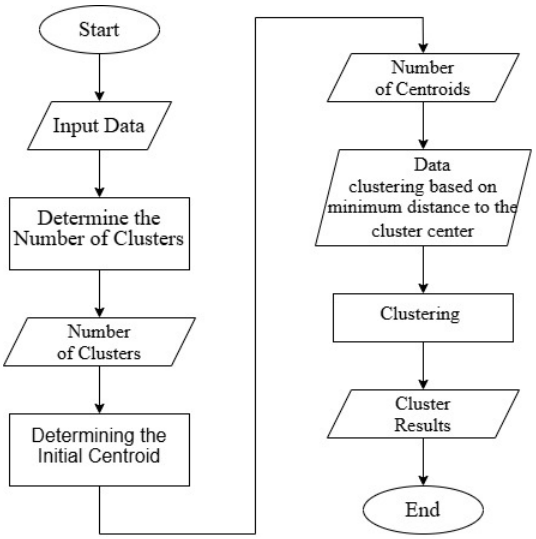


Figure 1 Research Methodology

2.1 Dataset

The selection of the dataset sample (representative data) is the initial stage in applying the data clustering technique. The sample dataset to be used consists of 167 student data entries and 16 attributes (Fiqh, Tauhid, Tafsir, Hadis, Nahwu, Shorof, English, Indonesian, Mathematics, Physics, Biology, Chemistry, Memorization of Surah, Extracurricular Activities, Behaviour, Attendance). The initial central centroids must be randomly selected after determining the sample dataset. The first centroid is from the 2nd data point, the second centroid is from the 11th data point, and the third centroid is from the 18th data point, which were initially chosen as central centroids in this study, as shown in Table 1.

Table 1. Sample of Student Data

No	Fiqh	Tauhid	Tafsir	⋯	Extracurricular	Behaviour	Attendance
1	9	9	9	⋯	0	8	9
2	7	7	8	⋯	0	8	9
3	9	9	9	⋯	8	8	10
4	7	8	5	⋯	0	8	9
…	…	…	…	…	…	…	…
164	6	9	6	⋯	0	8	8
165	6	8	5	⋯	8	8	9
166	8	9	8	⋯	8	8	10
167	8	9	9	⋯	0	8	4

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.2 Initial Centroid

The initial centroid is the starting point selected either randomly or based on a specific method in the K-Means algorithm to initiate the data clustering process. The selection of the initial centroid is crucial as it can affect the final clustering result, as shown in Table 2. This initial centroid serves as a temporary cluster centre and will be updated in each iteration until convergence is reached or no significant changes occur in the clustering process [19].

Table 2. Initial Centroid

Centroid	Fiqh	Tauhid	Tafsir	…	Extracurricular	Behaviour	Attendance
D2	7	7	8	…	0	8	9
D11	9	9	8	…	0	8	9
D18	9	9	9	…	8	8	9

3. RESULTS AND DISCUSSION

3.1. Calculating Centroid Distance

After determining the initial centroids, calculations are performed using the Euclidean formula as shown in Equation 1.

$$(ij) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

Iteration 1:

1. Calculating the closest centroid distance to Centroid 1 (D2) with values (7, 7, 8, 8, 9, 8, 7, 9, 6, 8, 6, 7, 7, 0, 8, 9).

$$C1 = \sqrt{(7-7)^2 + (7-7)^2 + (8-8)^2 + (8-8)^2 + (9-9)^2 + (8-8)^2 + (7-7)^2 + (9-9)^2 + (6-6)^2 + (8-8)^2 + (6-6)^2 + (7-7)^2 + (7-7)^2 + (0-0)^2 + (8-8)^2 + (9-9)^2} = 0$$

$$C2 = \sqrt{(7-9)^2 + (7-9)^2 + (8-8)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (7-8)^2 + (9-9)^2 + (6-9)^2 + (8-9)^2 + (6-9)^2 + (7-9)^2 + (7-8)^2 + (0-0)^2 + (8-8)^2 + (9-9)^2} = 5{,}916079783$$

$$C3 = \sqrt{(7-9)^2 + (7-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (7-8)^2 + (9-9)^2 + (6-9)^2 + (8-9)^2 + (6-8)^2 + (7-8)^2 + (7-7)^2 + (0-8)^2 + (8-8)^2 + (9-9)^2} = 9{,}53939201$$

2. Calculating the closest centroid distance to Centroid 2 (D11) with values (9, 9, 8, 9, 9, 9, 8, 9, 9, 9, 9, 9, 0, 8, 9).

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

$$C1 = \sqrt{(9-7)^2 + (9-7)^2 + (8-8)^2 + (9-8)^2 + (9-9)^2 + (9-8)^2 + (8-7)^2 + (9-9)^2 + (9-6)^2 + (9-8)^2 + (9-6)^2 + (9-7)^2 + (8-7)^2 + (0-0)^2 + (8-8)^2 + (9-9)^2} = 5{,}916079783$$

$$C2 = \sqrt{(9-9)^2 + (9-9)^2 + (8-8)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (8-8)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (8-8)^2 + (0-0)^2 + (8-8)^2 + (9-9)^2} = 0$$

$$C3 = \sqrt{(9-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (8-8)^2 + (9-9)^2 + (9-9)^2 + (9-9)^2 + (9-8)^2 + (9-8)^2 + (8-7)^2 + (0-8)^2 + (8-8)^2 + (9-9)^2} = 8{,}246211251$$

From the calculation of the nearest centroids above, the results of Iteration 1 are obtained and presented in Table 3.

Table 3 Iteration 1 Results

Data	C1	C2	C3	Minimum	Cluster
1	4,242641	3,605551	8,544004	3,605551	2
2	0	5,91608	9,539392	0	1
3	10,19804	8,185353	2,645751	2,645751	3
4	4	6,082763	10,14889	4	1
5	5,744563	8,485281	11,6619	5,744563	1
6	10,04988	9,486833	3,464102	3,464102	3
7	9,746794	11,6619	8,246211	8,246211	3
8	9,219544	9,486833	5,291503	5,291503	3
9	8,831761	8,544004	2,236068	2,236068	3
10	4,582576	5,830952	10	4,582576	1
11	5,91608	0	8,246211	0	2
12	8,660254	8,831761	3,162278	3,162278	3
13	10,0995	11,44552	8,185353	8,185353	3
14	10,14889	11,48913	8,246211	8,246211	3
15	8,888194	9,69536	5,09902	5,09902	3
16	9,949874	8	2	2	3
17	9,110434	8,485281	2,828427	2,828427	3
18	9,539392	8,246211	0	0	3
19	6	5	9,848858	5	2
20	6,480741	10,04988	12,76715	6,480741	1

Based on the calculation using the clustering formula, the grouping is determined based on the Senior High School lest distance to the nearest centroid, as follows: Old Cluster (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

New Cluster (2 1 3 1 1 3 3 3 3 1 2 3 3 3 3 3 3 3 2 1).

The data grouping has changed, and the calculation for the next iteration will continue. Table 4 shows the results of Iteration 2 calculations, obtained from the distance to the nearest centroid.

Table 4 Iteration 2 Results

Data	C1	C2	C3	Min	Cluster	Description
1	5,734108	2,848001	8,710227	2,848001	2	Safe
2	3,475629	4,630815	8,833726	3,475629	1	Safe
3	10,93984	8,530989	4,285797	4,285797	3	Safe
4	1,442221	4,333333	8,710227	1,442221	1	Safe
5	2,946184	6,666667	9,654087	2,946184	1	Safe
6	10,21176	9,261629	2,456567	2,456567	3	Safe
7	8,641759	10,41367	5,134984	5,134984	3	Safe
8	8,915156	8,931841	2,490259	2,490259	3	Safe
9	9,595832	8,412953	2,387758	2,387758	3	Safe
10	3,078961	4,136558	8,767443	3,078961	1	Safe
11	6,743886	2,108185	8,833726	2,108185	2	Safe
12	9,092854	8,393119	1,742045	1,742045	3	Safe
13	8,790904	10,30102	5,069654	5,069654	3	Safe
14	9,004443	10,38161	5,086065	5,086065	3	Safe
15	9,070832	9,006171	3,141346	3,141346	3	Safe
16	10,46327	8,273116	3,563901	3,563901	3	Safe
17	9,395744	8,192137	2,588704	2,588704	3	Safe
18	10,40577	8,43274	3,46911	3,46911	3	Safe
19	4,481071	3,480102	8,861982	3,480102	2	Safe
20	4,3909	8,530989	10,94081	4,3909	1	Safe

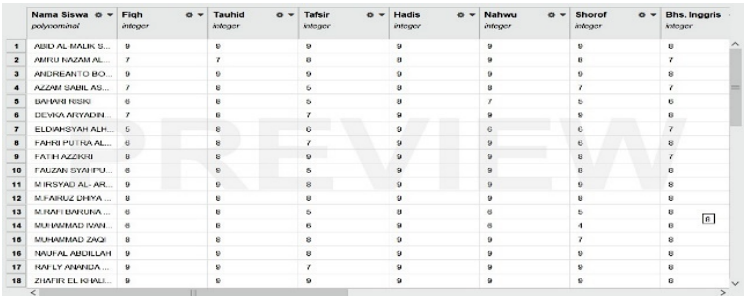
Based on the calculations using the Clustering formula, the grouping is determined based on the Senior High Schoolllest distance to the nearest centroid, as follows:
Old Cluster (2 1 3 1 1 3 3 3 3 1 2 3 3 3 3 3 3 3 2 1)
New Cluster (2 1 3 1 1 3 3 3 3 1 2 3 3 3 3 3 3 3 2 1).

If there are no further changes in the clusters after performing the calculations using the clustering algorithm in iterations 1 and 2, the iteration process will be stopped.

Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.2. Data Selection

Finding the file location is the first step. It was previously created with the "read excel" operator, which functions for the xlsx format. The button on the "read excel" operator no longer has a yellow exclamation mark, as it displays the imported data. This indicates that the data has been captured by the operator and is ready to be processed, as shown in Figure 2.



	Nama Siswa	Fiqh	Tauhid	Tafsir	Hadis	Nahwu	Shorof	Bhs. Inggris
	polynominal	integer	integer	integer	integer	integer	integer	integer
1	ABID AL MALIK S...	9	9	9	9	9	9	8
2	AMRU NAZAM AL...	7	7	8	8	9	8	7
3	ANDREANTO BO...	9	9	9	9	9	9	8
4	AZZAM SABIL AS...	7	8	5	8	8	7	7
5	BAHARI RISKI	6	8	5	8	7	5	6
6	DEVKA ARYADIN...	7	8	7	9	9	9	8
7	ELDIANSYAH ALH...	5	8	6	9	6	6	7
8	FAHRI PUTRA AL...	6	8	7	9	9	6	8
9	FATIH AZZIKRI	8	8	9	9	9	8	7
10	FAUZAN SYAHPU...	6	9	5	9	9	8	6
11	M IRSYAD AL- AR...	9	9	8	9	9	9	8
12	M.FAIRUZ DHIYA ...	8	8	8	9	9	8	8
13	M.RAFI BARUNA ...	6	8	5	8	6	5	8
14	MUHAMMAD IVAN...	6	8	6	9	6	4	8
15	MUHAMMAD ZAQI	8	8	8	9	9	7	8
16	NAUFAL ABDILLAH	9	9	8	9	9	9	8
17	RAFLY ANANDA ...	9	9	7	9	9	9	8
18	ZHAFIR EL KHALI...	9	9	9	9	9	9	8

Figure 2 Dataset

Next, differentiate the sequence of attribute names, coordinates, and the anticipated positions to be input using the "set role" operator into the "label" category. This is done to classify the "label" data and calculate and update the results in real time, as shown in Figure 3.

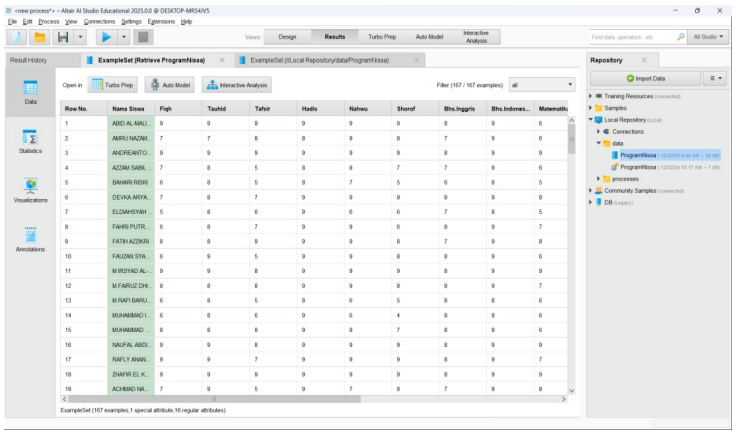


Figure 3 Set Role Result

3.3. Preprocessing

The purpose of this preprocessing step is to eliminate missing values from the data to be used, ensuring that the processing does not result in errors or issues. The initial processing for the previous table is performed as shown in Figure 4

Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Nama Siswa	Nominal	0	ZIVANA CINDERASUCI ...	ABID AL [...] IRING (1)	ABID AL-MALIK SEMBIRING (1), A
Fiqh	Integer	0	4	9	7.999
Tauhid	Integer	0	5	9	7.940
Tafsir	Integer	0	4	9	7.072
Hadis	Integer	0	5	9	8.449
Nahwu	Integer	0	4	9	7.928
Shorof	Integer	0	4	9	7.569
Bhs.Inggris	Integer	0	5	9	7.132
Bhs.Indonesia	Integer	0	5	9	7.892
Matematika	Integer	0	3	9	6.425

Figure 4 Preprocessing Data

3.4. K-Means Clustering

One of the stages that needs to be completed is determining how many clusters exist. The value of k, or the number of clusters, is used based on (the number of clusters, the optimal number, and the ideal participants in each cluster). The result of this clustering will be used as a reference to determine the elite student classes at Senior High School. Figure 5 is the process design for creating K-Means clusters of student data using the RapidMiner tool.

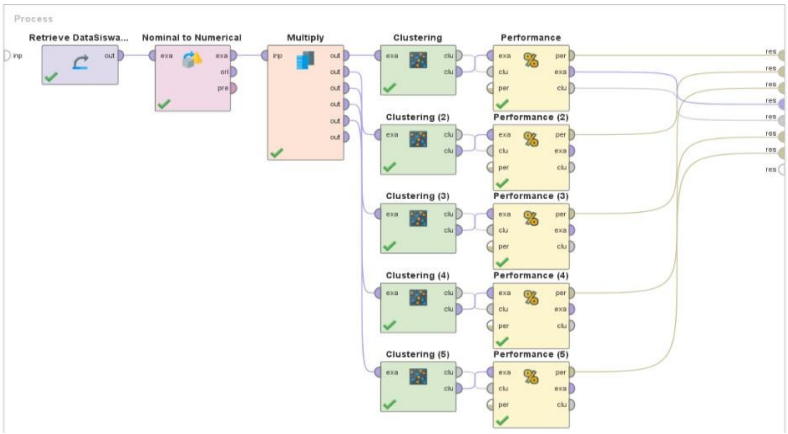


Figure 5 K-Means Design

Here are the clustering results obtained from RapidMiner, where the data has been grouped into clusters based on certain criteria. The clustering process helps in identifying patterns or similarities among data points, making it easier to analyse and make informed decisions. The results allow us to categorize students into different groups based on their attributes, which can then be used to determine the placement of students in the appropriate classes, such as identifying top-performing students or those in need of additional support.

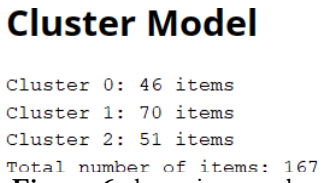


Figure 6 clustering results

3.5. Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering results by measuring the compactness and separation between clusters. A lower DBI value indicates better clustering performance, where clusters are both compact and well-separated [20]. In this study, the average DBI value of 1.226 reflects the overall quality of the clustering process using the K-Means algorithm. The result suggests that the algorithm effectively groups students' academic scores into clusters with a moderate level of compactness and separation. However, this value also indicates that there is potential for improvement in cluster quality through parameter optimization or the use of alternative clustering methods. These findings can serve as a foundation for personalized educational interventions, where each student cluster can receive tailored learning strategies to improve academic performance. Further research is recommended to explore different clustering techniques or hybrid methods to achieve better cluster quality and educational outcomes.

3.6. Discussion

The clustering analysis performed in this study aims to efficiently group student data to identify patterns that can be used for academic performance evaluation, class placement, and personalized learning strategies. By utilizing the K-Means clustering algorithm, the study demonstrates how students' attributes can be categorized into meaningful groups based on their proximity to centroids using Euclidean distance calculations. The results provide valuable insights into how clustering techniques can enhance decision-making in education, particularly in determining student classifications based on performance metrics.

The Euclidean distance formula was employed to determine the closeness of data points to the centroids, ensuring accurate student grouping. The calculations in Iteration 1 (Table 3) produced an initial classification, with students being assigned to clusters based on their nearest centroids. The iteration process continued until the cluster assignments stabilized, as seen in Iteration 2 (Table 4), where the new cluster assignments remained the same as the previous iteration. This stability indicates that the algorithm successfully converged to an optimal grouping, eliminating the need for additional iterations. The unchanged clustering

assignments confirm that the algorithm has effectively structured the dataset into well-defined groups, reducing computational complexity and ensuring efficiency. The process highlights the importance of choosing an appropriate number of clusters (k), as incorrect selections may lead to overfitting (too many clusters) or underfitting (too few clusters that fail to capture variation in the data). The clustering results suggest that students with similar academic attributes tend to fall within the same clusters, reinforcing the effectiveness of this method for identifying student performance trends.

The data selection and preprocessing steps are crucial to ensuring that the clustering results are reliable. Before analysis, missing values were eliminated to prevent distortions in cluster formation. The "Set Role" function was applied to appropriately label and differentiate attributes such as coordinate values, student scores, and predicted classifications (Figure 3). This ensured that the input variables were properly structured, allowing for accurate real-time updates during clustering.

The preprocessing step (Figure 4) further reinforced data integrity by removing inconsistencies, which could otherwise lead to erroneous clustering outputs. A clean dataset ensures that the clustering algorithm operates effectively, minimizing potential biases or misclassifications. The removal of incomplete data is essential, as missing values can significantly impact distance calculations, leading to inaccurate centroid placements and poor cluster assignment.

The K-Means clustering process (Figure 5) was applied to identify student groups based on predefined criteria, aiming to categorize students into different performance-based clusters. The results (Figure 6) show how student data was successfully partitioned into clusters, enabling educators to better understand academic trends and develop targeted interventions for students based on their grouping. One of the significant advantages of K-Means clustering in this study is its ability to handle large datasets efficiently, making it ideal for school performance analysis. By grouping students based on similarities in their attributes, educational institutions can: (1) Identify high-performing students and provide them with advanced learning opportunities. (2) Recognize students needing additional support and tailor intervention programs accordingly. (3) Optimize resource allocation by ensuring that students receive personalized assistance based on their cluster characteristics. However, one of the key challenges in K-Means clustering is its sensitivity to the initial selection of centroids. Poor initial centroid placement can lead to suboptimal clustering, requiring multiple iterations to achieve a stable solution. In this study, the convergence in Iteration 2 indicates that the initial centroid selection was effective, leading to meaningful group differentiation.

To validate the clustering results, the Davies-Bouldin Index (DBI) was employed as a performance evaluation metric. A DBI score of 1.226 was obtained, indicating

a moderate level of compactness and separation between clusters. While a lower DBI value would signify stronger clustering performance, the current score suggests that the K-Means algorithm was successful in grouping students based on their academic attributes but may still benefit from optimization. A DBI score greater than 1 suggests that there may be overlapping clusters or inconsistencies in how the data is partitioned. Future refinements to the clustering model may include:

1) Parameter tuning: Adjusting the number of clusters (k) to identify the most effective grouping.
2) Hybrid clustering approaches: Incorporating hierarchical or density-based clustering methods to improve classification accuracy.
3) Feature selection: Including additional academic performance indicators such as attendance, participation in extracurricular activities, and past academic trends to enhance cluster differentiation.

Despite the potential for improvement, the DBI value confirms that the K-Means model provides a reasonable structure for analysing student data. The clustering process successfully captured meaningful patterns and similarities among students, serving as a foundation for educational decision-making and performance analysis.

The findings from this study underscore the practical applications of clustering in educational settings. By classifying students based on their performance metrics, schools and educators can implement targeted learning strategies that address the unique needs of each student group. The results highlight the importance of data-driven decision-making in education, particularly in: (1) Developing customized learning plans based on student abilities. (2) Improving resource distribution by identifying at-risk students. (3) Enhancing institutional strategies for academic performance assessment. Future studies should explore alternative clustering techniques, such as fuzzy clustering or deep learning-based classification, to further refine student categorization. Additionally, incorporating real-time student performance tracking into the clustering model could enhance adaptability and ensure that learning strategies remain relevant over time.

The clustering analysis successfully demonstrated the effectiveness of the K-Means algorithm in categorizing students based on their academic attributes. By employing Euclidean distance calculations, the study identified clear student groupings, providing a structured framework for performance evaluation and academic interventions. Although the DBI score suggests room for improvement, the findings indicate that the clustering model provides valuable insights for personalized education strategies. The results emphasize the need for continuous refinement in clustering methodologies, ensuring higher precision in student classification and improved decision-making processes in education. In summary, this study highlights how machine learning-based clustering techniques can be leveraged to enhance academic planning, optimize student placement, and improve learning outcomes. Future research should focus on exploring hybrid clustering

approaches, integrating additional educational factors, and evaluating the long-term impact of data-driven student classification.

4. CONCLUSION

The K-Means clustering analysis successfully categorized Senior High School students into three distinct clusters based on their academic performance. These clusters provide valuable insights into student achievement levels, enabling educators to implement tailored learning strategies to support different performance groups. Cluster 0 (Excellent) consists of 46 students who consistently achieve high scores across all subjects, with centroid values ranging between 7.5 and 9. However, Mathematics recorded the lowest score within this group at 7, suggesting a potential area for improvement. Cluster 1 (Good) includes 70 students with moderate academic performance, where centroid values range between 6 and 9. Mathematics and Physics have the lowest recorded scores at 6, and no scores were recorded for Extracurricular Activities, indicating a possible gap in student participation outside of core subjects. Cluster 2 (Very Good) comprises 51 students whose performance is slightly lower than Cluster 0 but still strong. Centroid values range from 5 to 8, with the lowest scores recorded in Mathematics (5) and Physics (6), highlighting subjects where additional support may be beneficial.

To validate the effectiveness of the clustering results, the Davies-Bouldin Index (DBI) was employed, yielding an average validity index score of 1.226. This value indicates that the clustering structure is well-formed and effective in distinguishing student performance levels. While the clustering model successfully grouped students into meaningful categories, further optimization of cluster parameters or the exploration of alternative clustering techniques could enhance classification accuracy and refine student support strategies. K-Means clustering provides a data-driven approach to identifying student performance trends, allowing for personalized interventions, optimized resource allocation, and improved educational outcomes. Future research should focus on enhancing cluster quality, integrating additional academic indicators, and exploring hybrid machine learning techniques to further refine student classification and support decision-making in education.

REFERENCES

[1] S. Ujud, T. D. Nur, Y. Yusuf, N. Saibi, and M. R. Ramli, "Penerapan Model Pembelajaran Discovery Learning Untuk Meningkatkan Hasil Belajar Siswa Sma Negeri 10 Kota Ternate Kelas X Pada Materi Pencemaran Lingkungan," *J. Bioedukasi*, vol. 6, no. 2, pp. 337–347, 2023, doi: 10.33387/bioedu.v6i2.7305.

[2] R. Kurniawan, M. M. M. Mukarrobin, and M. Mahradianur, "Klasterisasi Tingkat Pendidikan Di Dki Jakarta Pada Tingkat Kecamatan Menggunakan Algoritma K-Means," *Technol. J. Ilm.*, vol. 12, no. 4, p. 234, 2021, doi: 10.31602/tji.v12i4.5633.

[3] B. Harahap and A. Rambe, "Implementasi K-Means Clustering Terhadap Mahasiswa yang Menerima Beasiswa Yayasan Pendidikan Battuta di Universitas Battuta Tahun 2020/2021 Studi Kasus Prodi Informatika," *Informatika*, vol. 9, no. 3, pp. 90–97, 2021, doi: 10.36987/informatika.v9i3.2185.

[4] D. Leman and E. Syahrin, "Sistem Cerdas Rekomendasi Klinik Pratama di Kota Medan Berbasis Data Mining Dengan Metode K-Means Untuk Pasien BPJS dan Umum," no. September, pp. 204–214, 2024.

[5] F. Asril, *Pengembangan Kecerdasan Majemuk pada Pembelajaran Tematik Kelas V di MI Modern Al Azhary Ajibarang*, Doctoral dissertation, UIN Prof. KH Saefuddin Zuhri, 2023.

[6] 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.

[7] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk.v15i2.1162.

[8] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, pp. 40–47, 2023.

[9] I. D. Setiawan and A. Triayudi, "Penerapan Data Mining Dengan Menggunakan Algoritma Clustering K-Means Untuk Pembagian Jurusan Pada Sekolah Menengah Atas," *J. Comput. Syst. Informatics*, vol. 5, no. 2, pp. 380–392, 2024, doi: 10.47065/josyc.v5i2.4970.

[10] Y. Andini, J. T. Hardinata, and Y. P. Purba, "Penerapan Data Mining Terhadap Tata Letak Buku Di Perpustakaan Sintong Bingei Pematangsiantar Menggunakan Metode Apriori," *J. TIMES*, vol. 11, no. 1, pp. 9–15, 2022, doi: 10.51351/jtm.11.1.2022661.

[11] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," *J. Ilm. Inform. Dan Ilmu Komput.*, vol. 3, no. 1, pp. 46–56, 2024, doi: 10.58602/jima-ilkom.v3i1.26

[12] N. Nursobah, S. Lailiyah, B. Harpad, and M. Fahmi, "Penerapan Data Mining Untuk Prediksi Perkiraan Hujan dengan Menggunakan Algoritma K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, 2022, doi: 10.47065/bits.v4i3.2564.

[13] S. N. Br Sembiring, H. Winata, and S. Kusnasari, "Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 1, p. 31, 2022, doi: 10.53513/jursi.v1i1.4784.

[14] A. Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 20–28, 2023, doi: 10.58602/jaiti.v1i1.22.

[15] S. Anwar, T. Suprapti, G. Dwilestari, and I. Ali, "Pengelompokkan Hasil Belajar Siswa dengan Metode Clustering K-Means," *JURSISTEKNI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 60–72, 2022.

[16] … Preddy, P. Marpaung, I. Pebrian, and W. Putri, "Penerapan Data Mining Untuk Pengelompokan Kepadatan Penduduk Kabupaten Deli Serdang Menggunakan Algoritma K-Means," *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 2, pp. 64–70, 2023.

[17] Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Clustering Perokok Usia Lebih dari 15 Tahun," *Bull. Inf. Technol.*, vol. 4, no. 4, pp. 501–507, 2023, doi: 10.47065/bit.v4i4.1067.

[18] R. Mauliadi, "Data Mining Menggunakan Algoritma K-Means Clustering dalam Analisis Tingkat Potongan Harga Terhadap Harga Jual Sepeda Motor Honda," *J. Inform. Ekon. Bisnis*, vol. 4, pp. 7–9, 2022, doi: 10.37034/infeb.v4i4.156.

[19] A. Supriyadi, A. Triayudi, and I. D. Sholihati, "Perbandingan Algoritma K-Means Dengan K-Medoids Pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 6, no. 2, pp. 229–240, 2021, doi: 10.29100/jipi.v6i2.2008.

[20] N. R. Saputra, G. Z. Muflih, and T. Informatika, "Pengelompokan Wilayah Indonesia Berdasarkan Komponen Indeks Pembangunan Manusia dengan Pendekatan Algoritma K-Means Clustering," vol. 8, pp. 156–167, 2025.