Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Clustering of High-Achieving Students Based on Scores at Junior High School Level Using K-Means Algorithm

Silva Audia[1]*, Relita Buaton[2], Zira Fatmaira[3]

[1,2,3]Information System, STMIK Kaputama, Indonesia
Silvaaudiaa17@icloud.com [1], bbcbuaton@gmail.com [2]*

Abstract

Education plays a crucial role in shaping quality human resources, and student achievement evaluation at the junior high school level is essential for supporting academic guidance, learning programs, and recognition of outstanding students. However, the increasing number of students often makes the process of identifying and categorizing achievement more complex. This study aims to develop a student clustering model at SMP Budi Utomo Binjai using the K-Means algorithm as part of a data mining approach. The input data consisted of 638 student records covering three main variables: average score, counseling score, and extracurricular score. Data were preprocessed and transformed before being processed using MATLAB R2014a, which provides a kmeans() function to automatically group the data into clusters. Several clustering trials were conducted with three to six clusters to evaluate the grouping performance. The results showed that students could be grouped into categories of high, medium, and low achievement, with each cluster having different characteristics of average, counseling, and extracurricular scores. Variance analysis indicated that clusters with smaller variance values represented more compact and homogeneous groupings, while clusters with higher variance values were more heterogeneous. The findings demonstrate that the K-Means algorithm is effective in grouping student performance data objectively, providing useful insights for teachers and school administrators to design more targeted learning strategies, academic interventions, and recognition systems. This research highlights the potential of data mining techniques to support decision-making processes in the education sector.

***Keywords**: Clustering; Data Mining; Junior High School; K-Means Algorithm; Student Achievement*

1. Introduction

Education is one of the main factors in developing quality human resources. At the junior high school level, evaluating student achievement is essential as a basis for decision-making, whether in determining learning programs, providing academic guidance, or giving awards to outstanding students. However, the large number of students often becomes an obstacle in the process of identifying and grouping achievements.

SMP Budi Utomo Binjai has a relatively large number of students in each grade, thus requiring an efficient method to group students based on their scores. This grouping is useful for understanding student characteristics, adjusting learning strategies, and giving awards or special attention more objectively. One method that can be applied is K-Means Clustering. K-Means is a non-hierarchical data mining algorithm used to cluster data into several groups based on similarity of characteristics. Its principle is to first determine the number of clusters (k), then the algorithm randomly selects initial centroids, calculates the distance of each data point to the centroids, and assigns the data to the nearest cluster. This process repeats until the centroids stabilize and the final optimal clusters are formed. Through this approach, students can be grouped into high, medium, and low achievement categories more objectively and systematically.

This study aims to develop a student grouping model at SMP Budi Utomo Binjai using the K-Means algorithm as a reference for evaluation, guidance, and learning strategy planning. Data mining with the clustering technique was chosen because it can extract important information from educational data, as demonstrated by previous studies (Nanda et al., 2023; Ani et al., 2021; Saputra & Pakereng, 2023) that proved the effectiveness of K-Means in identifying outstanding students.

Based on this background, the research problem can be formulated as follows: how the process of grouping student achievement data based on scores can be effectively carried out, what results are obtained when the K-Means algorithm is applied in such grouping, and how the implementation of K-Means can help schools identify high, medium, and low-achieving students more objectively..

2. Main Body

The steps carried out for processing student data at SMP Budi Utomo Binjai using the Clustering method with the K-Means algorithm produce new information and knowledge regarding high-achieving students based on their scores. The grouping is conducted by considering the variables of Average Score, Counseling Score, and Extracurricular Score. The results of this process provide insights that can be used as a basis for further strategies in delivering instruction tailored to the characteristics and needs of students.

In this study, data processing was carried out using MATLAB, a numerical computing software that offers strong support for data analysis, statistics, and the implementation of machine learning algorithms. MATLAB provides the kmeans() function, which enables researchers to perform data clustering automatically.

2.1. Discussion of Input Data

The system input data consists of data obtained from SMP Budi Utomo Binjai. The system input data was stored in Microsoft Office Excel as a data container, then transformed based on the transformation values of each variable used. The details of the input data, variables, and transformation values applied in the system are as follows:

1. Input Data

   File name : analisis.xlsx

   Number of records : 638 data entries

   Variables : - X = Average Score

   - Y = Counseling Score

   - Z = Extracurricular Score

2. Clustering Groups : 3 clusters
3. Data Transformation Values for Variables

Table 1: Transformed Data Values of Variables

No	Variabel	Transformasi	Nilai Transformasi
1	Nilai Rata-rata	91-100	1
		81-90	2
		71-80	3
		61-70	4
		<60	5
2	Nilai Bimbingan Konseling	91-100	1
		81-90	2
		71-80	3
		61-70	4
		<60	5
3	Nilai Ekskul	91-100	1
		81-90	2
		71-80	3
		<70	4

2.2. Implementation

In this chapter, the results of testing the data mining analysis software for student data at SMP Budi Utomo using MATLAB R2014a will be explained to obtain new information and knowledge regarding the students of SMP Budi Utomo Binjai, based on the variables of Average Score, Counseling Score, and Extracurricular Score. Implementation in this context refers to the process in which the transformed data is applied into the programming environment used and processed according to the clustering method with the K-Means algorithm. This allows us to determine the extent to which the system performs in processing the data and whether the results meet the users' needs.

2.2.1. Clustering Trial Using MATLAB

After implementing the method into the system design for data clustering, it is necessary to conduct testing to evaluate the results of the implementation. The testing is carried out by processing the system's input data. The display below shows the results of the clustering

process, including the data grouping visualization in the form of a graph with coordinate points representing data clusters and their centroids. The results of the clustering process are presented as follows:

1. Results of 3 Clusters

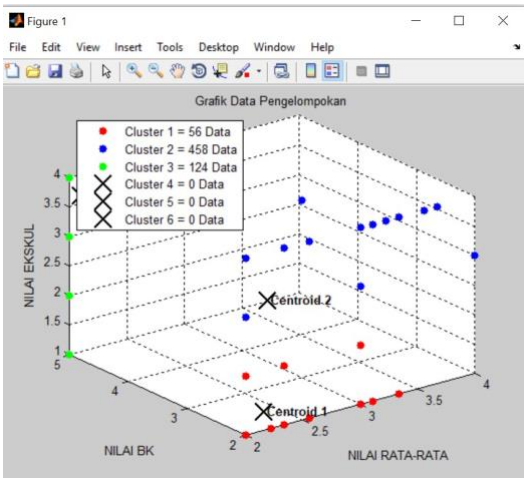

Fig. 2 : Results of 3 Clusters

In the figure above, it can be seen that Cluster 1 contains 56 data points, Cluster 2 contains 458 data points, and Cluster 3 contains 124 data points, with a total of 638 data points. The system trial results represent the output generated from the processing of the input data tested using MATLAB 2014a. The description of the clustering results using the K-Means algorithm is as follows:

Table 2 : Cluster Centroids of the System Results

No	Pusat Cluster	Variabel			Jumlah Data
		X	Y	Z	
1	Centroid 1	2.1567	2	1.3214	58
2	Centroid 2	2.1863	2.0021	3.1986	458
3	Centroid 3	2.0967	5	3.6451	124

Based on the table of clustering results obtained from the testing process, it can be observed that:

a) Group 1, with 56 data points, can be classified based on an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 91–100.
b) Group 2, with 458 data points, can be classified based on an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 71–80.
c) Group 3, with 124 data points, can be classified based on an Average Score of 81–90, Counseling Score of <60, and Extracurricular Score of <70.
d) Cluster variance.

```
Variance Cluster 1 = 0.3287
Variance Cluster 2 = 0.3226
Variance Cluster 3 = 0.5260
```

Based on the variance results, it can be concluded that Cluster 2 has the tightest data distribution (variance = 0.3226), followed by Cluster 1 (variance = 0.3287), while Cluster 3 shows the largest spread (variance = 0.5260), indicating that the data within this cluster is more heterogeneous. This means that Cluster 2 and Cluster 1 can be considered as better groupings since their members are relatively homogeneous, whereas Cluster 3, although still grouped, has greater data variation, which affects the overall quality of the clustering.

2. Results of 4 Clusters

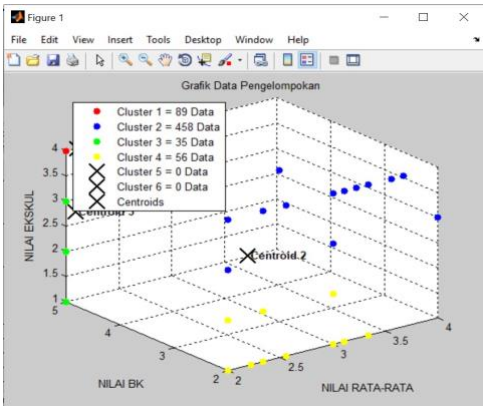

Fig. 2 : Results of 4 Clusters

In the figure above, it can be seen that Cluster 1 contains 89 data points, Cluster 2 contains 458 data points, Cluster 3 contains 35 data points, and Cluster 4 contains 56 data points, with a total of 638 data points. The system trial results represent the output generated from processing the input data tested using MATLAB 2014a. The description of the clustering results of community data using the K-Means clustering algorithm is as follows:

Table 3 : Cluster Centroids of the System Results

No	Pusat Cluster	Variabel			Jumlah Data
		X	Y	Z	
1	Centroid 1	2.1011	5	4	89
2	Centroid 2	2.1863	2.0021	3.1986	458
3	Centroid 3	2.0857	5	2.7428	35
4	Centroid 4	2.1567	2	1.3214	56

Based on the clustering results table obtained from the testing process, the following observations can be made:

a) Group 1, with 89 data points, is classified with an Average Score of 81–90, Counseling Score <60, and Extracurricular Score 71–80.
b) Group 2, with 458 data points, is classified with an Average Score of 81–90, Counseling Score 81–90, and Extracurricular Score 71–80.
c) Group 3, with 35 data points, is classified with an Average Score of 81–90, Counseling Score <60, and Extracurricular Score 81–90.
d) Group 4, with 56 data points, is classified with an Average Score of 81–90, Counseling Score 81–90, and Extracurricular Score 91–100.
e) Cluster Variance

```
Variance Cluster 1 = 0.3226
Variance Cluster 2 = 0.4980
Variance Cluster 3 = 0.3287
Variance Cluster 4 = 0.0909
```

Based on the variance values of each cluster, it is shown that Cluster 4 has the smallest variance of 0.0909, indicating the tightest and most homogeneous data distribution among all clusters. Meanwhile, Cluster 2 has the largest variance of 0.4980, meaning the data in this cluster is more dispersed and heterogeneous compared to the others. Cluster 1 (0.3226) and Cluster 3 (0.3287) fall within the medium variance range, suggesting that both clusters group the data fairly well, although some variability remains. Therefore, it can be concluded that the most optimal clustering result is represented by Cluster 4, while Cluster 2 requires further attention due to its higher data variation.

3. Results of 5 Clusters

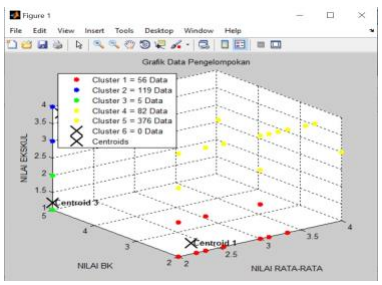

Fig. 3 : Results of 5 Clusters

In the figure above, it can be seen that Cluster 1 contains 56 data points, Cluster 2 contains 119 data points, Cluster 3 contains 5 data points, Cluster 4 contains 82 data points, and Cluster 5 contains 376 data points, with a total of 638 data points. The system trial results represent the output generated from the processing of the input data tested using MATLAB 2014a. The description of the community data clustering results using the K-Means clustering algorithm is as follows:

Table 4 : Cluster Centroids of the System Results

No	Pusat Cluster	Variabel			Jumlah Data
		X	Y	Z	
1	Centroid 1	2.1567	2	1.3214	56
2	Centroid 2	2.1008	5	3.7478	119
3	Centroid 3	2	5	1.2000	5
4	Centroid 4	3.0325	2.0121	3.2926	82
5	Centroid 5	2.0017	2	3.1781	376

Based on the clustering results table obtained from the testing process, the following can be observed:

a) Group 1, with 56 data points, is classified by an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 91–100.
b) Group 2, with 119 data points, is classified by an Average Score of 81–90, Counseling Score <60, and Extracurricular Score <70.
c) Group 3, with 5 data points, is classified by an Average Score of 81–90, Counseling Score <60, and Extracurricular Score of 91–100.
d) Group 4, with 82 data points, is classified by an Average Score of 71–80, Counseling Score of 81–90, and Extracurricular Score of 71–80.
e) Group 5, with 376 data points, is classified by an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 71–80.
f) Cluster variance.

```
Variance Cluster 1 = 0.1470
Variance Cluster 2 = 0.2415
Variance Cluster 3 = 0.2445
Variance Cluster 4 = 0.2568
Variance Cluster 5 = 0.5260
```

Based on the variance calculation results, Cluster 1 has the smallest variance value of 0.1470, indicating the tightest and most homogeneous data distribution. Cluster 2 (0.2415), Cluster 3 (0.2445), and Cluster 4 (0.2568) show relatively moderate variance values, meaning that although some diversity exists, the data within these clusters remain fairly compact. Meanwhile, Cluster 5 has the highest variance value of 0.5260, which indicates that its data is more dispersed and heterogeneous compared to the other clusters. Thus, it can be concluded that Cluster 1 represents the most optimal clustering result, while Cluster 5 is considered the least optimal due to its high data variation.

4. Results of 6 Clusters

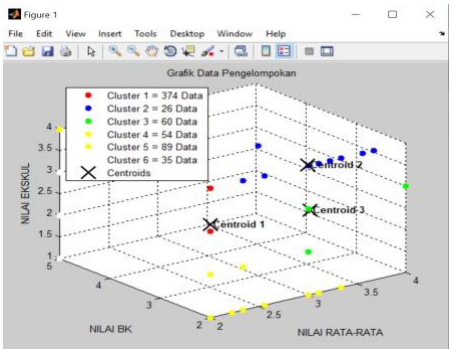

Fig. 4 : Results of 6 Clusters

In the figure above, it can be seen that Cluster 1 consists of 374 data points, Cluster 2 consists of 26 data points, Cluster 3 consists of 60 data points, Cluster 4 consists of 54 data points, Cluster 5 consists of 89 data points, and Cluster 6 consists of 35 data points, with a total of 638 data points. The system trial results represent the output generated from the processing of input data tested using MATLAB 2014a. The description of the community data clustering results using the K-Means clustering algorithm is as follows :

Table 5 : Cluster Centroids of the System Results

No	Pusat Cluster	Variabel			Jumlah Data
		X	Y	Z	
1	Centroid 1	2	2	3.1737	374

2	Centroid 2	3.0128	2.0384	4	26
3	Centroid 3	3.0166	2	2.9666	60
4	Centroid 4	2.1255	2	1.2962	54
5	Centroid 5	2.1011	5	4	89
6	Centroid 6	2.0857	5	2.7428	35

Based on the clustering results table obtained from the testing process, the following can be observed:

a) Group 1, with 374 data points, is classified by an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 71–80.
b) Group 2, with 26 data points, is classified by an Average Score of 71–80, Counseling Score of 81–90, and Extracurricular Score <70.
c) Group 3, with 60 data points, is classified by an Average Score of 71–80, Counseling Score of 81–90, and Extracurricular Score of 81–90.
d) Group 4, with 54 data points, is classified by an Average Score of 81–90, Counseling Score of 81–90, and Extracurricular Score of 91–100.
e) Group 5, with 89 data points, is classified by an Average Score of 81–90, Counseling Score <60, and Extracurricular Score <70.
f) Group 6, with 35 data points, is classified by an Average Score of 81–90, Counseling Score <60, and Extracurricular Score of 71–80.
g) Cluster variance.

```
Variance Cluster 1 = 0.1470
Variance Cluster 2 = 0.0909
Variance Cluster 3 = 0.4980
Variance Cluster 4 = 0.2445
Variance Cluster 5 = 0.2308
Variance Cluster 6 = 0.2524
```

Based on the variance calculation results, it is shown that Cluster 2 has the smallest variance value of 0.0909, indicating that the data within this cluster is very compact and homogeneous, making its clustering quality highly optimal. On the other hand, Cluster 3 has the highest variance of 0.4980, showing that the data is more dispersed and heterogeneous compared to the other clusters. Meanwhile, Cluster 4 (0.2445), Cluster 5 (0.2308), and Cluster 6 (0.2524) fall within the medium variance range, which means they still maintain a reasonable level of compactness despite some diversity among members. Thus, it can be concluded that the best clustering result is represented by Cluster 2, whereas Cluster 3 represents the lowest clustering quality.

3. Conclusion

Berdasarkan hasil penelitian menggunakan algoritma K-Means dengan bantuan MATLAB R2014a terhadap 638 data siswa SMP Budi Utomo Binjai, diperoleh beberapa temuan penting :

a) Pada uji coba 3 cluster, jumlah siswa terbagi menjadi 56 siswa pada Cluster 1, 458 siswa pada Cluster 2, dan 124 siswa pada Cluster 3. Variance yang dihasilkan menunjukkan bahwa Cluster 2 (0,3226) memiliki sebaran data paling rapat, diikuti Cluster 1 (0,3287), sementara Cluster 3 memiliki variance terbesar (0,5260).
b) Pada uji coba 4 cluster, siswa terbagi menjadi 89 siswa pada Cluster 1, 458 siswa pada Cluster 2, 35 siswa pada Cluster 3, dan 56 siswa pada Cluster 4. Variance terkecil terdapat pada Cluster 4 (0,0909) yang menandakan pengelompokan paling homogen, sedangkan variance terbesar terdapat pada Cluster 2 (0,4980).
c) Pada uji coba 5 cluster, jumlah siswa terbagi menjadi 56 siswa pada Cluster 1, 119 siswa pada Cluster 2, 5 siswa pada Cluster 3, 82 siswa pada Cluster 4, dan 376 siswa pada Cluster 5. Variance terkecil ada pada Cluster 1 (0,1470), sementara Cluster 5 memiliki variance terbesar (0,5260).
d) Pada uji coba 6 cluster, jumlah siswa terbagi menjadi 374 siswa pada Cluster 1, 26 siswa pada Cluster 2, 60 siswa pada Cluster 3, 54 siswa pada Cluster 4, 89 siswa pada Cluster 5, dan 35 siswa pada Cluster 6. Hasil variance menunjukkan bahwa Cluster 2 adalah yang paling optimal dengan nilai 0,0909, sedangkan Cluster 3 paling lemah dengan variance 0,4980.

Secara keseluruhan, dapat disimpulkan bahwa pengelompokan terbaik diperoleh pada uji coba 4 cluster dan 6 cluster, di mana variance terendah berada pada Cluster 4 (0,0909) dan Cluster 2 (0,0909). Hal ini menunjukkan bahwa algoritma K-Means mampu mengelompokkan siswa secara objektif menjadi kelompok berprestasi tinggi, menengah, dan rendah, sehingga hasilnya dapat dijadikan dasar strategi pembelajaran yang lebih tepat sasaran.

References

[1] Abduloh, S. P., Suntoko, M. P., Tedi Purbangkara, S. P., & Ade Abikusna, M. P. (2022). Peningkatan dan pengembangan prestasi belajar peserta didik. uwais inspirasi indonesia.

[2] Ani, H., Nofriansyah, D., & Mariami, I. (2021). Implementasi Data Mining Untuk Pengelempokan Buku Di Perpustakaan Yayasan Nurul Islam Indonesia Baru Dengan Metode K-Means Clustering. Jurnal Cyber Tech, 1(1).
[3] Dewi, F. P., Aryni, P. S., & Umaidah, Y. (2022). Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran. JISKA (Jurnal Informatika Sunan Kalijaga), 7(2), 111–121.
[4] Fatah, V. F., Susanti, S., Ariyanti, M., & Nursyamsiyah, N. (2021). Penyesuaian Diri Siswa Tahun Pertama SMP Dimasa Pandemi Covid 19. Jurnal Keperawatan, 6(2), 232–239.
[5] Fatimah, S., Suriati, S., & Usman, A. (2022). Pengelompokkan Tingkat Pemahaman Guru PAUD Terhadap Pembelajaran Berbasis STEAM Menggunakan Metode X-Means Clustering. Explorer, 2(1), 24–31.
[6] Haryanto, S., Mawaddah, N., Rahman, R., Fatmawati, F., & Octafiona, E. (2024). Analysis of Islamic Counselling and Learning Motivation: Keys to Successful Student Academic Achievement. Journal of Education Research, 5(2).
[7] Juariah, S. (2023). Paradigma Pendidikan Islam Dan Pengembangan Sumber Daya Insani Dalam Membentuk Etika Dan Karakter Dalam Masyarakat Islam. Kaipi: Kumpulan Artikel Ilmiah Pendidikan Islam, 1(2), 65–71.
[8] Khesya, N. (2021). Mengenal Flowchart Dan Pseudocode Dalam Algoritma Dan Pemrograman. Muchamad, M. K. (2022). Pengantar Simulasi Sistem Komunikasi Digital Menggunakan Matlab. Syiah
[9] Kuala University Press.
[10] Nanda, W. S., Pardede, A. M. H., & Simanjuntak, M. (2023). Analisis Data Mining Untuk Klasterisasi Data Rekam Medis Menggunakan Algoritma K-Means Pada Rumah Sakit Sylvani Binjai. Indonesian Journal of Education And Computer Science, 1(3), 82–88.
[11] Parinsi, M. T., Mewengkang, A., & Rantung, T. (2021). Perancangan Sistem Informasi Sekolah Di Sekolah Menengah Kejuruan. Edutik: Jurnal Pendidikan Teknologi Informasi Dan Komunikasi, 1(3), 227–240.
[12] Permadi, A., & Wiyaja, Y. A. (2023). Pengelompokan Terbaik Menggunakan Algoritma K-Means Pada Dataset Bus Biskita Bogor. INTERNAL (Information System Journal), 6(1), 88–100.
[13] Rahayu, N. D., Anshor, A. H., & Afriantoro, I. (2024). Penerapan Data Mining untuk Pemetaan Siswa Berprestasi menggunakan Metode Clustering K-Means. JUKI: Jurnal Komputer Dan Informatika, 6(1), 71–83.
[14] Rosaly, R., & Prasetyo, A. (2019). Pengertian Flowchart Beserta Fungsi dan Simbol-simbol Flowchart yang Paling Umum Digunakan.
[15] Saputra, D. A., & Pakereng, M. A. I. (2023). Analisis Data Nilai Siswa Kelas 8 Berbasis Nilai Pengetahuan Untuk Menentukan Siswa Berprestasi dengan K-Means Clustering (Kasus SMP Negeri 4 Salatiga). J-SAKTI (Jurnal Sains Komputer Dan Informatika), 7(2), 630–638.
[16] Smrti, N. N. E., Sukenada, A. I. P. G., Kadek, D. T. R. N., Adnan, A., & Ode, J. P. P. (2023). Flowgorithm Sebagai Penunjang Pembelajaran Algoritma dan Pemrograman. Jurnal Bangkit Indonesia, 12(1), 56–64.
[17] Surbakti, N. K. (2021). Data Mining Pengelompokan Pasien Rawat Inap Peserta BPJS Menggunakan Metode Clustering (Studi Kasus: RSU. Bangkatan). Journal of Information and Technology, 1(2), 47–53.
[18] Syamsiah, S. (2019). Perancangan Flowchart dan Pseudocode Pembelajaran Mengenal Angka dengan Animasi untuk Anak PAUD Rambutan. STRING (Satuan Tulisan Riset Dan Inovasi Teknologi), 4(1), 86–93.
[19] Ulfah, U., & Arifudin, O. (2021). Pengaruh aspek kognitif, afektif, dan psikomotor terhadap hasil belajar peserta didik. Jurnal Al-Amar: Ekonomi Syariah, Perbankan Syariah, Agama Islam, Manajemen Dan Pendidikan, 2(1), 1–9.