Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

Clustering of the Best Senior High Schools in Serdang Bedagai Regency Using the K-Means Method

Tania Annisa Siagian[1*], Nurdin[2], Munirul Ula[2]

[1]Departement of informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia
[2]Departement of informatics Technology, Universitas Malikussaleh, Lhokseumawe, Indonesia
Email: [1]tania.210170248@mhs.unimal.ac.id, [2]nurdin@unimal.ac.id, [3]munirulula@unimal.ac.id
Email Penulis Korespondensi: tania.210170248@mhs.unimal.ac.id*
Submitted: **27/05/2025**; Accepted: **13/06/2025**; Published: **30/06/2025**

Abstract− This study aims to cluster the best Senior High Schools (SMA) in Serdang Bedagai Regency using the K-Means method. Five evaluation indicators were used in the clustering process: accreditation, school status, number of teachers, achievements, and facilities. A total of 41 schools were analyzed using a non-hierarchical approach, with the optimal number of clusters determined through the Elbow Method, resulting in three groups: excellent, good, and fair. Data normalization was performed using the Min-Max method to ensure equal scaling among variables. The clustering results using the K-Means algorithm formed three clusters that represent the quality of schools based on transformed numerical data. The K-Means method proved capable of providing a general overview of school quality grouping, which can serve as a basis for policy-making to improve the quality of education in the region.

Keywords: K-Means, Clustering, Senior High School, Elbow Method, Serdang Bedagai

1. INTRODUCTION

Education is a fundamental pillar in the development of a nation. One of the major challenges still faced in Indonesia's education system is the inequality in quality among educational institutions. This issue is particularly evident in certain regions, such as Serdang Bedagai Regency, which has a large number of Senior High Schools (SMA) with varying standards. These differences include school accreditation, status (public or private), number of teaching staff, student achievements, and availability of educational facilities. In such conditions, the public often finds it difficult to choose the best school that fits their children's needs. On the other hand, local governments require accurate and structured data to develop targeted educational improvement policies. Without an objective classification system, prioritizing development and allocating educational resources becomes suboptimal.

To address these issues, data mining techniques particularly clustering can be applied to group schools objectively based on selected indicators [1]. Clustering is a method of grouping data into several clusters based on the similarity of characteristics [2]. One of the most widely used clustering algorithms is K-Means. This method is favored for its speed, ease of implementation, and efficiency in handling large-scale data [3]. K-Means works by first determining the number of clusters (K), then randomly initializing cluster centroids. Each data point is then measured against each centroid using the Euclidean Distance formula, and grouped based on the shortest distance. This process is repeated iteratively until the centroid positions stabilize and no longer change significantly [4].

To improve the clustering results, the data must undergo transformation and normalization. This study employed the Min-Max Normalization method to standardize the scale of each variable, ensuring that no single attribute dominates the distance calculation [5]. The optimal number of clusters was determined using the Elbow Method, while the quality of the clustering results was evaluated using the Davies-Bouldin Index (DBI). A smaller DBI value indicates better clustering, with compact and well-separated clusters [6].

Previous studies have used K-Means in various educational contexts. For example, [7] Clustered students based on academic performance, resulting in informative performance segmentation. [8] utilized K-Means for school classification in Seruyan Regency based on facilities and teaching staff, which helped the local government prioritize development efforts. [9] applied K-Means to map regions based on the Human Development Index, providing spatial insights into development disparities. [10] compared K-Means and K-Medoids for grouping academic programs and found that K-Means produced more optimal clustering results [11]. Meanwhile, [12] used K-Means to automatically generate exam question packages, demonstrating the algorithm's effectiveness in maintaining consistent difficulty levels across questions.

However, these studies have not specifically addressed school clustering based on the integration of five quality indicators simultaneously, nor have they focused on particular regional contexts with unique local educational policy needs. Most of the previous research relied on limited variables, such as academic scores or infrastructure alone, and did not incorporate real combined datasets (both primary and secondary data) that reflect the full complexity of educational quality. In contrast, this study considers not only quantitative indicators from government records but also qualitative inputs such as facilities data obtained through school-level questionnaires, resulting in a more comprehensive and context-specific analysis.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

This is the research gap addressed by the present study. By utilizing the most recent data from 41 senior high schools in Serdang Bedagai Regency in 2024 and integrating accreditation, school status, number of teachers, student achievements, and facilities, this study presents a more contextual clustering model that can be directly used by local policymakers. The objective of this study is to classify high schools in Serdang Bedagai Regency into three quality categories using the K-Means algorithm based on the five aforementioned indicators. The clustering results are expected to provide an objective overview of the educational landscape in the region, which can serve as a basis for decision-making in educational planning and policy evaluation. Furthermore, this approach is expected to support the development of a transparent, accountable, and data-driven education system in the region.

2. RESEARCH METHODOLOGY

This research was conducted through several systematic stages. The stages of research conducted are as follows:

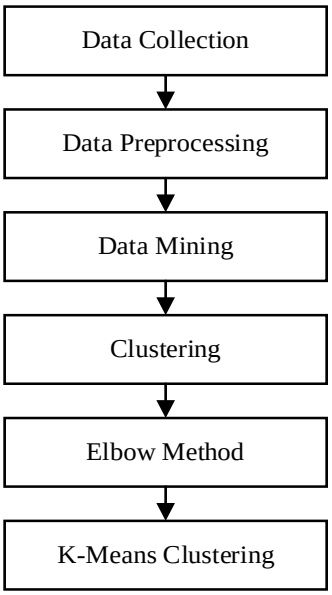

Figure 1. Research Stage

1. Data Collection

This study utilized data from 41 Senior High Schools (SMA) located across Serdang Bedagai Regency. Data were collected through several methods, including direct observation at schools, official documentation from the Serdang Bedagai Regency Department of Education, and the distribution of questionnaires to school representatives to obtain information that was not publicly available. Five main attributes were used as the basis for the clustering process: school status, accreditation, number of teachers, achievements, and facilities. School status indicates whether the school is public or private. Accreditation reflects the most recent accreditation score obtained by the school. The number of teachers refers to the total number of active teaching staff. Achievements record the number of accomplishments attained by the school in various competitions or educational events. Meanwhile, facilities refer to the completeness of infrastructure and supporting resources for teaching and learning activities. These five attributes were selected because they are considered to represent the core aspects of educational quality assessment at the senior high school level and are relevant indicators for determining the overall quality of a school.

2. Data Preprocessing

Before the clustering process was carried out, the dataset underwent several preprocessing steps to ensure consistency and accuracy. First, categorical attributes such as school status and accreditation were transformed into numerical format using label encoding, where qualitative values were systematically assigned numeric codes. This step was essential to allow mathematical processing in the K-Means algorithm. Following transformation, the dataset was normalized using the Min-Max Normalization technique, which scaled all attribute values into a range between 0 and 1. This normalization ensured that no single attribute would dominate the distance calculation during the clustering process, allowing all variables—status, accreditation, number of teachers, achievements, and facilities—to contribute proportionally. These preprocessing steps were critical in preparing the data for the clustering analysis and ensuring that the results reflected real patterns rather than scale biases among variables.

3. Data Mining

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

Data mining is a systematic process of extracting information and hidden patterns from large-scale data [13], [14]. In this study, data mining is utilized to cluster senior high schools (SMA) in Serdang Bedagai Regency based on attributes such as accreditation, school status, number of teachers, achievements, and facilities, with the aim of providing an objective overview to support decision-making in efforts to improve the quality of education.

4. Clustering

Clustering is a technique in data mining used to group a set of data into several clusters based on similarity in characteristics [15]. Data within the same cluster exhibit high similarity, while data between clusters are significantly different. This process is performed automatically using algorithms and is widely applied in various fields such as market segmentation, behavior analysis, and data mapping to uncover hidden structures within large datasets[16].

5. Elbow Method

The Elbow Method is a technique used to determine the optimal number of clusters in clustering analysis by examining a sharp drop in a plotted graph that forms an “elbow” shape. This elbow point indicates the most appropriate number of clusters, as further increases in the number of clusters result in only minimal reductions in intra-cluster variation. This method helps to establish the value of K by balancing data variation and model complexity [17].

6. K-Means Clustering Algorithm

K-Means is a non-hierarchical clustering method that partitions data into several groups based on the similarity of numerical attributes. This algorithm uses a partitioning approach to separate data into a predefined number of clusters and is known for its efficiency in handling large datasets and detecting outliers. Each data point is assigned to the nearest cluster and the clustering process is updated iteratively until the results converge and become stable [18] Steps in Data Grouping Using the K-Means Clustering Algorithm

1. Determine the number of clusters (K) to be created.
2. Initialize the centroids randomly.
3. Calculate the distance between each data point and each centroid using the Euclidean Distance formula:
   $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (1)
4. Assign each data point to the cluster whos’e centroid is closest based on the calculated distance.
5. Recalculate the centroids by computing the average of all data points assigned to each cluster using the formula:

   $c_k = \left(\frac{1}{n_k}\right)\sum_{i=1}^{n_k} x_i$ (2)

6. Recalculate the distance from each data point to the new centroids. Repeat the process until the centroids no longer change significantly, indicating that the clustering has converged.

3. RESULT AND DISCUSSION

3.1 Data Transformation and Normalization

Categorical variables such as school status and accreditation were first converted into numerical format using label encoding. This transformation was necessary to allow the K-Means algorithm to process the data mathematically. Subsequently, Min-Max normalization was applied to rescale all attribute values within the range of 0 to 1, ensuring that no single variable disproportionately influenced the clustering outcome due to differing value ranges.

To ensure consistency and comparability across all observations, the transformation and normalization processes were conducted prior to clustering. The selected attributes school status, accreditation, number of teachers, achievements, and facilities were chosen as representative indicators of educational quality. The processed dataset, ready for analysis, is shown in the following table:

Table 1. Research Dataset

No	School Name	Status	Accreditation	Teachers	Achievements	Facilities
1	SMAN 1 Bandar Khalipah	Public	A	33	0	3

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

2	SMAN 1 Bintang Bayu	Public	A	23	2	4
3	SMAN 1 Dolok Masihul	Public	A	41	2	4
4	SMAN 1 Dolok Merawan	Public	A	21	4	4
5	SMAN 1 Kotarih	Public	B	23	1	4
…	…	…	…	…	…	…
40	SMAS Yapim Syah Bandar	Private	A	8	0	4
41	SMAS Yapim Taruna	Private	B	15	0	4

The transformation was carried out by converting categorical data into numerical values using the label encoding method. Subsequently, the data were normalized using the Min-Max Normalization method to standardize the scale across attributes, ensuring that distance calculations in the algorithm are not biased toward any particular attribute.

Table 2. Normalizedtation Data

No	School Name	Status	Accreditation	Teachers	Achievements	Facilities
1	SMAN 1 Bandar Khalipah	1	0,65	0	0	0,65
2	SMAN 1 Bintang Bayu	1	0,4	0,333333333	0,5	0,4
3	SMAN 1 Dolok Masihul	1	0,85	0,333333333	0,5	0,85
4	SMAN 1 Dolok Merawan	1	0,35	0,666666667	0,5	0,35
5	SMAN 1 Kotarih	0,5	0,4	0,166666667	0,5	0,4
…	…	…	…	…	…	…
40	SMAS Yapim Syah Bandar	0	1	0,025	0	0,5
41	SMAS Yapim Taruna	0	0,5	0,2	0	0,5

3.2 Determining Number of Clusters

The Elbow Method is used to determine the optimal number of clusters in the K-Means algorithm by observing the "elbow point" on the inertia graph, where the rate of decrease begins to slow down, indicating the most suitable number of clusters, as illustrated in the following figure.

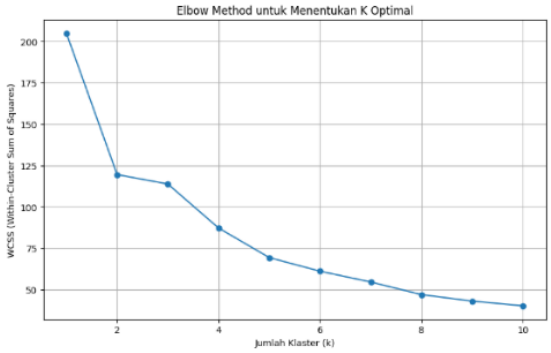

Figure 2. Elbow Curve

The Elbow graph shows a significant decrease in inertia up to K = 3, after which it stabilizes. Therefore, K = 3 is selected as the optimal number of clusters. The three resulting clusters are: C1 = Excellent, C2 = Good, and C3 = Fairly Good.

3.3 Implementation of the K-Means Clustering

The following steps represent the stages applied by the author in the data grouping process using the K-Means Clustering algorithm.

1. Input the normalized data of all senior high schools in Serdang Bedagai Regency for the year 2024.
2. Initial centroid determination is performed randomly for three points, in accordance with the number of clusters previously established based on the data. The initial centroids are presented in the following table:

Table 3. Initial Centroid

Cluster	Status	Accreditation	Teachers	Achievements	Facilities

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

C1 =	1	1	1	0,5	0,5
C2 =	1	1	0,6	0	0,5
C3 =	0	1	0,4	0,5	0,5

3. Next, the distance of each data point to the predetermined cluster centroids was calculated using the Euclidean Distance formula. Below is an example of the distance calculation for each data point in the cluster:

$$d(1,1) = \sqrt{(1+1)^2 + (1+1)^2 + (0{,}65+1)^2 + (0+0{,}5)^2 + (0+0{,}5)^2} = 0{,}788986692$$

$$d(2,1) = \sqrt{(1+1)^2 + (1+1)^2 + (0{,}65+0{,}6)^2 + (0+0)^2 + (0+0{,}5)^2} = 0{,}502493781$$

$$d(3,1) = \sqrt{(1+0)^2 + (1+1)^2 + (0{,}65+0{,}4)^2 + (0+0{,}5)^2 + (0+0{,}5)^2} = 1{,}25$$

Table 5. Distance Calculation Results of Each Data Point to Cluster Centroids in Iteration 1.

Table 4. Iteration 1

No	C1	C2	C3	Distance	Class
1	0,788986692	0,502493781	1,25	0,502493781	C2
2	0,622718056	0,388730126	1,013793755	0,388730126	C2
3	0,224227068	0,416666666	1,109178876	0,224227068	C1
4	0,671027405	0,712000312	1,015025999	0,671027405	C1
5	0,849182613	0,563717818	1,166666667	0,563717818	C2
…	…	…	…	…	…
40	1,48345037	1,1535272	0,625	0,625	C3
41	1,462873884	1,187434209	0,734846923	0,734846923	C3

The process of calculating distances and updating clusters was carried out iteratively to adjust the centroid positions based on the average of the data within each cluster. Changes in data classification were still observed during the initial iterations. However, after five iterations, the system no longer exhibited any changes in data grouping. Therefore, iteration 5 was considered the final iteration, as the centroids had converged and the clustering results were deemed stable.

Table 5. Iteration 5

No	C1	C2	C3	Distance	Class
1	0,779979248	0,434732822	1,306910526	0,434732822	C2
2	0,357292248	0,449051164	1,143869673	0,357292248	C1
3	0,234856872	0,443448924	1,298101792	0,234856872	C1
4	0,416003588	0,661119541	1,251779662	0,416003588	C1
5	0,674402138	0,390210497	1,023507823	0,390210497	C2
…	…	…	…	…	…
40	1,306646838	1,235934583	0,500813374	0,500813374	C3
41	1,322783624	1,14823215	0,106699054	0,106699054	C3

4. In the fifth iteration, the results were identical to those in the fourth iteration, indicating that the clustering process had converged and the computation was terminated. Using the K-Means Clustering method, the data were divided into three clusters: Cluster 1 and Cluster 2 each contained 9 data points, while Cluster 3 contained 23 data points. The final clustering results are presented in the following table.

Table 6. Clustering Result Percentage

No	Cluster Type	Number of Schools	Percentage
1	Excellent (C1)	9	22,0%
2	Good (C2)	9	22,0%
3	Fairly Good (C3)	23	56,1%
Total		41	100%

3.4 Visualization of K-Means Clustering Results

The results of the K-Means Clustering process are visualized using a scatter plot where each data point represents a high school (SMA) in Serdang Bedagai Regency. The plot is constructed based on two principal components

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 6, Nomor 4, Juni 2025
e-ISSN 2685-998X
DOI 10.30865/json.v6i4

derived from dimensionality reduction (such as PCA) to effectively project multi-attribute data into a two-dimensional space for interpretation. Each cluster is represented by a distinct color:

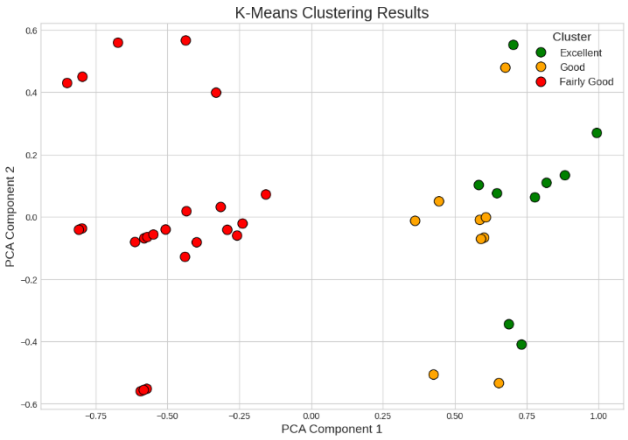

Figure 3. K-means Clustering Results

To clarify the clustering results, a Principal Component Analysis (PCA) plot was created, as shown in Figure 2. This visualization projects the multi-dimensional dataset onto two principal components, providing a clearer view of the distribution of schools within each cluster. The plot reveals that schools classified as "Excellent" (represented by green dots) tend to cluster tightly in the upper right quadrant, indicating strong similarities in their characteristics and high performance across the evaluated indicators. This cluster consists of 7 schools, suggesting that only a small portion of the schools have achieved top-tier quality.

In contrast, schools in the "Fairly Good" category (red dots) are more widely scattered, especially in the left quadrant, possibly indicating greater variability in quality within this group. This is the largest cluster, comprising 22 schools. Meanwhile, schools categorized as "Good" (orange dots) are positioned between the two extremes in both location and quality, with a total of 9 schools falling into this group. The relatively clear separation between clusters in the PCA space supports the validity of the K-Means clustering results and demonstrates that the selected features effectively capture differences in school quality. This graphical representation also enhances the readability of the clustering outcomes, making it easier for stakeholders to understand the relative positioning of each school based on overall performance.

4. CONCLUSION

This study concludes that the K-Means clustering algorithm is an effective method for classifying 41 Senior High Schools (SMA) in Serdang Bedagai Regency into three quality categories: Excellent, Good, and Fairly Good. The classification was based on five key indicators school status, accreditation, number of teachers, achievements, and facilities which were selected to represent the core dimensions of educational quality. Prior to clustering, data were transformed and normalized using the Min-Max method to ensure balanced contributions of each variable in distance calculations. The Elbow Method was employed to determine the optimal number of clusters, which resulted in three distinct groupings. The final clustering outcome revealed that the majority of schools fall into the Fairly Good category, while fewer schools were categorized as Good or Excellent. These findings indicate that K-Means not only enables objective, data-driven categorization but also provides valuable insights into the distribution of school quality within the region. Consequently, this approach can assist policymakers and educational stakeholders in formulating more targeted and equitable strategies for improving educational standards, ensuring that resources and interventions are directed toward schools that need them most.

REFERENCES

[1] Nurdin, Bustami, Rini Meiyanti, Amalia Fahada, and Marleni, “Application of the K-Means Method for Clustering Capture Fisheries Products in North Aceh with A Data Mining Approach,” *Journal of Advanced Zoology*, vol. 44, no. 4, pp. 39–49, Oct. 2023, doi: 10.17762/JAZ.V44I4.1358.

[2] Warisa and N. Nurahman, “Perbandingan Performa Cluster Model Algoritma K-Means Dalam Mengelompokkan Penerima Bantuan Program Keluarga Harapan,” *J. Sistem Info. Bisnis*, vol. 13, no. 1, pp. 20–28, Jun. 2023, doi: 10.21456/vol13iss1pp20-28.

[3] R. Iman, B. Rahmat, and A. Junaidi, “Implementasi Algoritma K-Means dan Knearest Neighbors (KNN) Untuk Identifikasi Penyakit Tuberkulosis Pada Paru-Paru,” *Repeater : Publikasi Teknik Informatika dan Jaringan*, vol. 2, no. 3, pp. 12–25, Jun. 2024, doi: 10.62951/repeater.v2i3.77.

[4] Nurdin, Bustami, R. Meiyanti, and A. Fahada, "Clustering Types Of Capture Fisheries Products Using The K-Means Clustering," *J Theor Appl Inf Technol*, vol. 15, no. 17, 2024, [Online]. Available: www.jatit.org

[5] D. A. Manalu and G. Gunadi, "Implementasi Metode Data Mining K-Means Clustering Terhadap Data Pembayaran Transaksi Menggunakan Bahasa Pemrograman Python Pada Cv Digital Dimensi," *Infotech: Journal of Technology Information*, vol. 8, no. 1, pp. 43–54, Jun. 2022, doi: 10.37365/jti.v8i1.131.

[6] D. Oktario Dacwanda and Y. Nataliani, "Implementasi k-Means Clustering untuk Analisis Nilai Akademik Siswa Berdasarkan Nilai Pengetahuan dan Keterampilan," *AITI: Jurnal Teknologi Informasi*, vol. 18, no. Agustus, pp. 125–138, 2021.

[7] S. Kurniawan, A. M. Siregar, and H. Y. Novita, "Penerapan Algoritma K-Means dan Fuzzy C-Means Dalam Mengelompokan Prestasi Siswa Berdasarkan Nilai Akademik," vol. IV, no. 1, 2023.

[8] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, Mar. 2022, doi: 10.30812/matrik.v21i2.1411.

[9] . Hanniva, A. Kurnia, S. Rahardiantoro, and A. A. Mattjik, "Penggerombolan Kabupaten/Kota di Indonesia Berdasarkan Indikator Indeks Pembangunan Manusia Menggunakan Metode K-Means dan Fuzzy C-Means," *Xplore: Journal of Statistics*, vol. 11, no. 1, pp. 36–47, Jan. 2022, doi: 10.29244/xplore.v11i1.855.

[10] S. Salamah, D. Abdullah, and N. Nurdin, "Comparative Analysis of K-Means and K-Medoids to Determine Study Programs," *International Journal of Engineering, Science and Information Technology*, vol. 5, no. 1, pp. 167–176, 2025, doi: 10.52088/ijesty.v5i1.673.

[11] T. Salsabila, N. Nurdin, and S. Retno, "Comparison of K-Medoids and K-Means Result for Regional Clustering of Capture Fisheries in Aceh Province," *International Journal of Engineering, Science and Information Technology*, vol. 5, no. 2, pp. 282–289, Mar. 2025, doi: 10.52088/ijesty.v5i2.829.

[12] L. S. Riza, R. A. Rosdiyana, A. Wahyudin, and A. R. Pérez, "The k-means algorithm for generating sets of items in educational assessment," *Indonesian Journal of Science and Technology*, vol. 6, no. 1, pp. 93–100, 2021, doi: 10.17509/ijost.v6i1.31523.

[13] M. Fikry *et al.*, "Data Mining for Processing of Research and Community Service by Lecturer Using Decision Tree Method," 2020.

[14] N. Nurdin and D. Astika, "Penerapan Data Mining Untuk Menganalisis Penjualan Barang Dengan Menggunakan Metode Apriori Pada Supermarket Sejahtera Lhokseumawe," *TECHSI - Jurnal Teknik Informatika*, vol. 7, no. 1, pp. 132–155, 2015, doi: 10.29103/TECHSI.V7I1.184.

[15] M. Faisal and Z. Fitri, "Information and Communication Technology Competencies Clustering for students for Vocational High School Students Using K-Means Clustering Algorithm," 2022, doi: 10.52088/ijesty.v1i4.318.

[16] N. K. Zuhal, "Study Comparison K-Means Clustering Dengan Algoritma Hierarchical Clustering." Accessed: Dec. 14, 2024. [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/stains/article/view/1495/1220

[17] N. Syahfitri, E. Budianita, A. Nazir, and I. Afrianty, "KLIK: Kajian Ilmiah Informatika dan Komputer Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid," *Media Online*, vol. 4, no. 3, pp. 1668–1675, 2023, doi: 10.30865/klik.v4i3.1525.

[18] F. Handayani, "Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokan Mahasiswa Berdasarkan Gaya Belajar." Accessed: Dec. 13, 2024. [Online]. Available: https://ojs.unikom.ac.id/index.php/jati/article/view/6733/2965