Jurnal JEETech: Vol. 7 No. 1 Tahun 2026
Journal Of Electrical Engineering And Technology
https://ejournal.ft-undar.ac.id/index.php/jeetech
DOI: https://doi.org/10.32492/jeetech.v7i1.7115
e-ISSN: 2722-5321
p-ISSN: 2964-7320

K-Means Clustering for Primary Education Inequality in Serang City

[1*] Ari Sopiandi, [2] Muawan Bisri, [3] Eko Aziz Apriadi, [4] Agus Komarudin
[1,2,3,4] Department of Informatics, Universitas Indonesia Mandiri, Lampung Selatan
[1] arisopiandi1810@gmail.com, [2] muawan.bisri@gmail.com, [3] ekoazizapriadi72@gmail.com, [4] aguskomarudin689@gmail.com

Article Info

Article history:

Received March 23, 2026
Revised April 29, 2026
Accepted May 17, 2026

Keyword:

Data mining
K-Means
Clustering
Educational inequality

ABSTRACT

Educational inequality remains a persistent issue in various regions, including Serang City. Differences in access to and quality of education across areas are influenced by several factors such as the number of schools, students, and teaching staff. However, available educational data are generally presented only in the form of descriptive statistics, which are not sufficient to provide in-depth insights into the patterns of inequality. Therefore, a computational-based approach is needed to analyze educational data more effectively. This study aims to analyze trends in educational development and map educational inequality in Serang City using a data mining approach with the K-Means algorithm. The data used in this study consist of educational data obtained from the Serang City Education Office, Dapodik, and BPS, including the number of schools, the number of students, and the number of teaching staff. The analysis process is carried out through data preprocessing, normalization, and data clustering using the K-Means algorithm with the help of RapidMiner software. The results show that educational data can be grouped into several clusters representing the level of educational conditions across regions. Each cluster has different characteristics, which can be used to identify areas with high and low levels of educational inequality. Thus, the data mining approach is able to provide a more systematic overview of educational conditions and support data-driven decision-making

Corresponding Author:
Ari Sopiandi
Department of Informatics, Universitas Indonesia Mandiri, Lampung Selatan
Email: arisopiandi1810@gmail.com

1. Introduction

Educational inequality remains a major concern in many countries, including Indonesia. In Serang City, Banten Province, disparities in access to and quality of education still exist between urban and suburban areas. These inequalities are influenced by economic, social,

and geographical factors, which affect educational quality and may lead to future social disparities [1].

With the advancement of information technology, the use of data in various sectors, including education, has significantly increased. Educational data such as the number of schools, students, and teaching staff continue to grow each year and hold great potential for analysis in supporting decision-making. However, in practice, such data are often presented only as descriptive statistics and are not fully utilized to uncover deeper patterns. Therefore, a computational-based approach is needed to process educational data more effectively.

One approach that can be used is data mining, which is the process of analyzing large datasets to discover patterns, relationships, and useful information for decision-making [2]. Data mining techniques can extract knowledge from large datasets using various analytical methods [3]. In addition, it can identify hidden patterns within data, producing more meaningful insights, particularly in the field of education [4].

One commonly used technique in data mining is clustering, which groups data based on similarity in certain characteristics. This technique allows similar data to be grouped together, making analysis easier. One of the widely used algorithms is K-Means, which groups data based on the distance to cluster centers (centroids).

Previous studies have shown that data mining, especially clustering, is effective for educational data analysis. The K-Means algorithm is widely used in data mining to group similar data based on characteristics [5]. Furthermore, clustering methods can provide a more systematic and data-driven overview of educational conditions [6]. Other studies also indicate that data mining can help analyze patterns and trends of educational inequality [7].

In the context of Indonesia, several recent studies have also demonstrated the effectiveness of K-Means clustering in analyzing educational data. A study by [8] shows that clustering can be used to group regions based on similarities in educational indicators. Furthermore, research conducted by [9] indicates that K-Means is effective in classifying educational quality based on variables such as students and teachers. In addition, [10] explains that clustering can identify patterns of educational distribution across regions.

Previous studies state that data mining techniques are widely used to analyze student performance [11], while [12] highlights their role in identifying educational characteristics and supporting decision-making. However, the application of data mining in mapping educational inequality at the regional level, particularly in Serang City, remains limited. There is a research gap in applying data mining methods specifically to analyze and map educational inequality in this area. Therefore, this study aims to provide a more comprehensive and systematic analysis of educational data.

Based on these issues, this study aims to analyze educational development trends and map educational inequality in Serang City using a data mining approach with the K-Means algorithm. This research is expected to provide a more accurate overview of educational

conditions across regions and support data-driven decision-making in educational policy formulation.

2. Methodology

A. Method

This study uses a quantitative approach with an exploratory method to analyze educational inequality in Serang City. The quantitative approach is used because the study processes numerical data, while the exploratory approach aims to identify patterns in educational data using data mining techniques, especially using the K-Means clustering algorithm introduced by MacQueen [13]. The study was conducted during 2025–2026 in Serang City, Banten Province.

The data used in this study were obtained from the Serang City Education Office, the Central Bureau of Statistics (BPS), and the Basic Education Data (Dapodik). The dataset includes the number of schools, students, teaching staff, and other educational indicators. This study focuses on primary education to maintain data consistency.

The research variables include the number of study groups (rombel), number of students, number of teaching staff, and educational facilities. These variables are used as the basis for clustering to identify patterns of educational inequality across regions.

Data collection was conducted through documentation methods using official sources. The collected data were then processed through preprocessing stages, including data cleaning, handling missing values, and normalization to ensure consistent scales across variables.

Data analysis was performed using a data mining approach with clustering techniques using the K-Means algorithm. The optimal number of clusters (K) was determined using the Elbow method. The clustering process includes initializing centroids, calculating distances using Euclidean Distance, assigning data to the nearest cluster, and updating centroids until convergence is reached.

The use of the K-Means algorithm in this study is supported by previous research in Indonesia. According to [14], K-Means is effective in grouping data based on similarity and is widely used in data analysis. Furthermore, [15] states that the Elbow method can be used to determine the optimal number of clusters by evaluating the clustering results.

Additionally, descriptive statistical analysis was conducted to provide an overview of the educational data. Data processing and analysis were carried out using RapidMiner software to facilitate analysis and visualization.

B. Figures and Tables

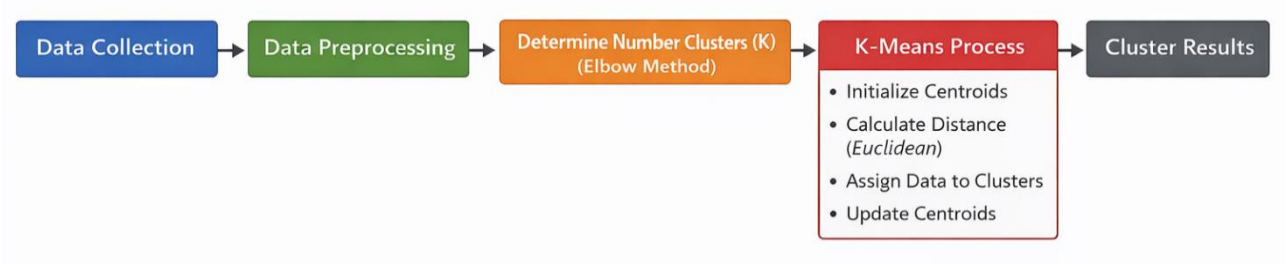

Figure 1. Research Methodology Flowchart

The process begins with Data Collection, where relevant data is gathered from various sources. In the context of educational analysis, this may include the number of students, teaching staff, study groups, and school distribution data. The quality and completeness of this stage are crucial, as the accuracy of the final clustering results depends heavily on the data collected.

Next is Data Preprocessing, which involves cleaning and preparing the data before analysis. This step may include handling missing values, removing duplicates, normalizing data, and transforming variables into a suitable format. Proper preprocessing ensures that the dataset is consistent and ready for clustering, reducing noise that could affect the outcome.

The third stage is Determining the Number of Clusters (K) using the Elbow Method. In this step, different values of K are tested to identify the optimal number of clusters. The Elbow Method works by plotting the within-cluster sum of squares (WCSS) and finding the point where the rate of decrease sharply changes, forming an "elbow." This point indicates a balance between minimizing error and avoiding excessive cluster complexity.

After determining the optimal K, the K-Means Process is carried out. This involves several iterative steps, starting with initializing centroids randomly. Then, the algorithm calculates the distance between each data point and the centroids, typically using Euclidean distance. Each data point is assigned to the nearest cluster, and the centroids are updated based on the mean of the assigned points. This process repeats until the centroids stabilize.

Finally, the process produces Cluster Results, which represent grouped data based on similarity. These clusters can be interpreted to identify patterns, such as disparities in educational distribution or performance across regions. The results provide meaningful insights that can support decision-making and policy development.

3. Results and Discussion

The results of this study were obtained by applying the K-Means clustering algorithm to educational data in Serang City. A total of 92 data points were analyzed, including variables such as number of students, study groups, and teaching staff. Before clustering, the data were normalized to ensure equal scaling.

The results show that the data can be grouped into five clusters. Each data point was successfully assigned to a cluster based on similarity.

Table 1. Data Distribution

No	Cluster	Number of Data	Percentage
1	Cluster 1	41	44.6%
2	Cluster 2	36	39.1%
3	Cluster 3	11	12.0%
4	Cluster 4	2	2.2%
5	Cluster 5	2	2.2%

The distribution of data shows that Cluster 1 contains the largest proportion (44.6%), followed by Cluster 2 (39.1%), while the remaining clusters have smaller proportions. This indicates that most data are concentrated in two main groups.

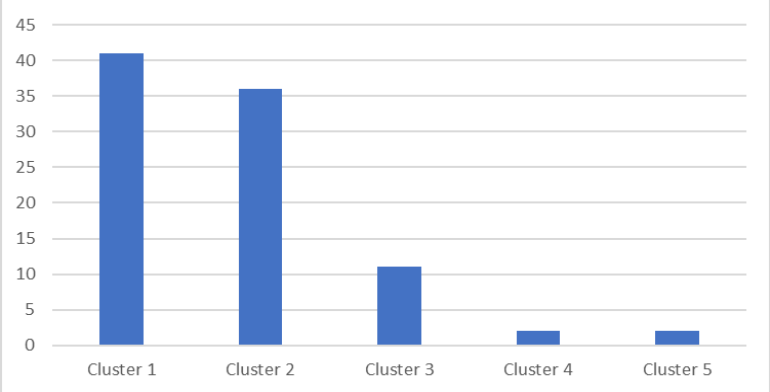

Figure 2. Data Distribution in Each Cluster

Based on the bar chart, Cluster 1 contains the highest number of data points, with a total of approximately 40, indicating that most data are grouped within this cluster. Cluster 2 follows as the second largest group, with around 35 data points, showing a relatively similar but slightly smaller concentration compared to Cluster 1.

In contrast, Cluster 3 has a significantly lower number of data points, around 10, suggesting a moderate grouping. Meanwhile, Cluster 4 and Cluster 5 contain the fewest data points, each with only about 2–3 entries, indicating that these clusters represent smaller or more specific data patterns.

Overall, the distribution shows an uneven clustering result, where the majority of data are concentrated in Clusters 1 and 2, while the remaining clusters contain relatively fewer data points. This suggests that the dataset has dominant patterns captured by the larger clusters, with smaller clusters representing outliers or less frequent characteristics.

Table 2. Average Value of Variables in Each Cluster

No	Cluster	Average Teachers	Average Sync Count	Average Students	Average Study Groups
1	Cluster 1	-0.701	-0.322	-0.733	-0.751
2	Cluster 2	1.935	0.173	1.996	1.819
3	Cluster 3	-0.774	4.338	-0.481	-0.622
4	Cluster 4	0.124	0.051	0.334	0.155
5	Cluster 5	2.271	0.401	-1.485	3.215

Based on Table 2, the average values for each cluster indicate significant differences in characteristics among clusters. These values are derived from a data normalization process, which ensures comparability across variables. In this context, positive values indicate that a variable is above the overall mean, while negative values indicate that it is below the average. This interpretation helps in understanding the relative position of each cluster compared to the entire dataset. As a result, the clusters can be analyzed more objectively based on their standardized characteristics.

Cluster 1 has negative average values across all variables, including the number of students, study groups, and teaching staff. This indicates that this cluster tends to have characteristics below the overall average. Such conditions suggest that regions within this cluster may face limitations in educational resources and infrastructure. The consistently low values across variables highlight a general pattern of underdevelopment in the education sector. Therefore, Cluster 1 can be categorized as representing areas with relatively low educational conditions.

Cluster 2 shows relatively high positive average values across all variables, particularly in the number of students and teaching staff. This suggests that this cluster has better educational conditions compared to other clusters. The availability of more teaching staff and larger student populations indicates stronger educational capacity and accessibility. These characteristics reflect regions that are more developed in terms of education services. Consequently, Cluster 2 can be seen as representing areas with above-average educational performance.

Cluster 3 has a very high average value in the sync count variable but relatively low values in other variables. This indicates an imbalance in the characteristics within this cluster. The high sync count suggests a strong level of data activity or reporting frequency, while the lower values in teaching staff, students, and study groups indicate limited educational resources. This disparity reflects an inconsistency between administrative activity and actual educational capacity. Therefore, Cluster 3 represents regions with high synchronization or reporting intensity but relatively low availability of educational resources.

Cluster 4 has average values close to zero across all variables, indicating that its characteristics are near the overall average. This suggests that regions in this cluster have relatively balanced educational conditions. There are no extreme shortages or surpluses in terms of students, study groups, or teaching staff. Such stability indicates a moderate level of development in the education sector. Therefore, Cluster 4 can be interpreted as representing areas with average and stable educational conditions.

Cluster 5 shows relatively high average values in study groups and teaching staff but low values in the number of students. This reflects a different pattern of characteristics compared to other clusters. The availability of resources appears sufficient or even excessive relative to the number of students. This condition may indicate inefficiencies in

resource allocation within these regions. As a result, Cluster 5 highlights areas where educational resources are not optimally utilized.

The clustering results using the K-Means algorithm reveal significant differences in the characteristics of educational data in Serang City. Based on the analysis, the data are grouped into five clusters with distinct characteristics. Cluster 1 is characterized by below-average values across all variables, indicating limited educational resources. In contrast, Cluster 2 shows above-average values, reflecting better educational conditions. Cluster 4 represents relatively stable conditions, as its values are close to the overall average. Meanwhile, Clusters 3 and 5 exhibit imbalanced patterns, where certain variables are significantly higher or lower than others. These differences highlight variations in the distribution of students, teaching staff, and study groups across regions and indicate disparities in educational conditions.

4. Conclusion

The results of this study demonstrate that the K-Means algorithm is effective in identifying patterns of educational inequality at the primary school level in Serang City. The clustering process reveals that educational conditions vary significantly across regions, indicating that the distribution of students, teaching staff, and study groups is not evenly balanced. These disparities highlight the existence of both resource gaps and inefficiencies in allocation.

Furthermore, the findings confirm that educational inequality is influenced not only by the availability of resources but also by how they are distributed across regions. The clustering results provide a clear and structured overview of these conditions, which can support data-driven decision-making. Therefore, this study can serve as a reference for policymakers in determining priority areas and improving the equity of educational resource distribution.

References

[1] Badan Pusat Statistik, *Statistik Pendidikan Kota Serang*. Serang: BPS Kota Serang, 2022.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2021.

[3] A. B. M. Shawkat Ali et al., "Data Mining Techniques in Education: A Review," *IEEE Access*, vol. 9, 2021.

[4] A. Saxena et al., "A Review of Clustering Techniques and Developments," *Neurocomputing*, vol. 267, pp. 664–681, 2021.

[5] X. Wu et al., "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, vol. 63, no. 1, pp. 1–37, 2022.

[6] S. Sharma and K. Jain, "Applications of Data Mining in Education: A Review," *Education and Information Technologies*, vol. 27, 2022.

[7] M. Kaur and U. Kaur, "Educational Data Mining for Student Performance Analysis Using Clustering Techniques," *IEEE Access*, vol. 10, 2022.

[8] S. K. Mohamad et al., "Educational Data Mining: A Review," *Procedia - Social and Behavioral Sciences*, vol. 97, pp. 320–324, 2021.

[9] M. Romero et al., "Educational Data Mining and Learning Analytics: An Updated Survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2021.

[10] Y. Xing, X. Chen, and J. Sun, "Application of K-Means Clustering in Educational Data Analysis," *Expert Systems with Applications*, vol. 186, 2022.

[11] R. Baker and K. Yacef, "The State of Educational Data Mining in 2023: A Review," *Journal of Educational Data Mining*, vol. 15, no. 1, pp. 1–25, 2023.

[12] A. Peña-Ayala, "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2022.

[13] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 2021.

[14] T. Kanungo et al., "An Efficient K-Means Clustering Algorithm: Analysis and Implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[15] N. Maori, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Jurnal Simetris*, 2023.