Interdisciplinary Journal of Global and Multidisciplinary
IJGAM
+62852-3430-9925 https://jurnal-ijgam.or.id/ admin@jurnal-ijgam.or.id
Vol. 2 No. 2 April 2026 | Page. 313 - 321

CLUSTERING OF DISTRICTS/CITIES IN INDONESIA BASED ON POVERTY LEVEL USING THE K-MEANS METHOD WITH THE ELBOW METHOD APPROACH (CASE STUDY: BPS 2025 DATA)

Dinda Syafitri[1], Sahara Lani Lestari[2], Kayla Amelia Putri[3], Frengki Alfredo Matondang[4]
Computer Science, Faculty Of Mathematics and Natural Sciences,
State University of Medan, Indonesia
e-mail : dindasyafitrii06@gmail.com, lestarisahara28@gmail.com,
kaylaamelia499@gmail.com, fr4nkelblue.689@gmail.com

Keywords:

Poverty,
K-Means Clustering,
Elbow Method,
Silhouette Score,
Central Statistics Agency (BPS)

Abstrak

Poverty remains a significant structural issue in Indonesia's development. The disparity in poverty levels across regions highlights the need for a data-driven analytical approach to comprehensively understand its distribution patterns. This study aims to cluster districts and cities in Indonesia based on the number of poor residents using the K-Means Clustering method optimized via the Elbow Method. The data used consists of secondary data from the Central Statistics Agency (BPS) for the year 2025, covering 514 districts/cities across 38 provinces. The analysis process began with a data preprocessing stage, including data cleaning, outlier detection using the Interquartile Range (IQR) method, and normalization using StandardScaler. The optimal number of clusters was determined using the Elbow Method and evaluated using the Silhouette Score. The analysis results show that the optimal number of clusters is K = 3 with a Silhouette Score of 0.6987, which falls into the "good" category. The classification resulted in three groups: Low Poverty (369 regions or 71.8%), Moderate Poverty (113 regions or 22.0%), and High Poverty (32 regions or 6.2%). Although the number of regions in the High Poverty group is relatively small, this group accounts for 26.4% of the total national poor population and is dominated by regions with high population density on the island of Java. These findings are expected to serve as a basis for the government in formulating more targeted and data-driven poverty alleviation policies.

INTRODUCTION

Etymologically, the word poverty derives from a state of lacking material possessions or being in a condition of deprivation. According to the Ministry of Social

Affairs and the Central Statistics Agency (BPS), poverty is defined as the inability of individuals to meet the minimum basic needs for a decent life, such as food, clothing, shelter, education, and health (Salmin, 2023). Poverty is commonly found in developing countries and represents one of the most complex issues to resolve, including in Indonesia (Putra & Anggrawan, 2021).

One of the most pressing social issues in Indonesia is poverty. Despite various government efforts to reduce poverty rates, the problem in numerous districts and cities remains a significant challenge (Amelia et al., 2025). In Indonesia, poverty measurement is conducted by BPS through the establishment of minimum standards for both food and non-food needs that must be met on a daily basis (Madaliyah & Rohmah, 2024).

To understand poverty distribution patterns more systematically, a data analysis approach capable of processing large volumes of information is needed. Data mining is a branch of computer science used to process large and complex datasets. In the context of poverty, clustering techniques can be employed to group regions based on similarity in data characteristics (Prasetyo et al., 2025). Therefore, it is necessary to cluster the welfare levels of districts and cities to understand the degree of disparity and equity in welfare across regions (Matdoan et al., 2024).

Clustering is a data analysis technique used to group objects or data into several clusters based on shared characteristics (Tawakal et al., 2025). This method aims to identify hidden patterns or structures in data that were previously unknown. Clustering is a widely used data mining technique that separates a set of data points into several groups or clusters (Rahmadani & Nursyahira, 2025).

One of the most popular and widely applied clustering algorithms across various fields is the K-Means algorithm. K-Means is one of the most frequently used clustering methods due to its ability to efficiently group numerical data (P. Sari et al., 2024). K-Means is a non-hierarchical clustering method used to partition data into several groups (clusters). Data with similar characteristics are grouped into the same cluster, while data with different characteristics are placed in different clusters (Suhartini & Yuliani, 2021).

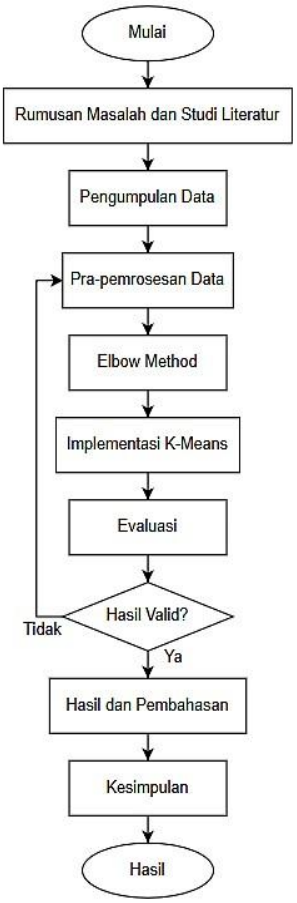
In data mining, this method divides data into specific clusters by calculating the distance between data points, rather than based on specific variables. Each cluster has a central point (centroid), and a data point is assigned to the cluster whose centroid is closest (Sianturi et al., 2025). To support the determination of the appropriate number of clusters, the Elbow Method is one of the most commonly used approaches. This method works by calculating the variation within the data (inertia) for several cluster counts and visualizing the results in a graph to find the optimal "elbow" point (S. N. Sari et al., 2024). Thus, the combination of K-Means and the Elbow Method can produce more objective and measurable groupings.

Several previous studies have attempted to apply the K-Means method to cluster regions based on poverty levels. (Alfiansyah et al., 2022) conducted clustering on 22 sub-districts in Blitar Regency using five poverty variables and successfully formed two regional groups; however, the results applied only to a single regency and therefore could not represent poverty conditions across a broader area (R. Sari et al., 2025). also applied a similar method in Pagar Alam City, producing five population groups, but the data used were individual-level and covered only one city. Meanwhile, (Dasa et al., 2024) K-Means to 18 villages in Tana Righu Sub-district; however, the very small dataset made the clustering results insufficiently representative to depict regional poverty conditions.

Based on the above review, no study has yet conducted a comprehensive clustering of poverty across all districts/cities in Indonesia using the most recent BPS data. Therefore, this study seeks to fill that gap by clustering all districts/cities in Indonesia based on BPS poverty data for 2025, with the aim of producing a more complete poverty map that can serve as a basis for national policy considerations.

RESEARCH METHOD

This study is descriptive quantitative in nature with a computational data mining-based approach. The analytical technique applied is K-Means Clustering optimized using the Elbow Method, which belongs to the group of unsupervised machine learning methods. Rather than testing hypotheses, this research is directed at discovering hidden patterns and structures in national-level poverty data. The data source used is from BPS 2025, covering all districts and cities in Indonesia. The research process was carried out in stages as illustrated in Figure 1.



[Figure 1. Research Stages]

Data Collection

The data used in this study are secondary data obtained from the official publication of the Central Statistics Agency (BPS) of the Republic of Indonesia for 2025,

accessible via www.bps.go.id. BPS was selected as the data source because it is the official institution authorized to collect and present statistical data, ensuring high levels of accuracy and reliability. The unit of analysis in this study covers 514 districts/cities spread across 38 provinces in Indonesia. The dataset used is titled "Number of Poor Population (Thousands) by District/City, 2025." The variable used in this study is the number of poor residents in units of thousands of people, which serves as the basis for the regional clustering process to understand the pattern of poverty distribution in Indonesia.

Data Preprocessing

Before analysis, the data must go through a preprocessing stage to ensure more accurate results. This stage involves three main processes. First, data cleaning, which involves checking and handling missing, incomplete, or inconsistent data to prevent its impact on the analysis results. Second, outlier handling, which identifies extreme values that differ significantly from other data using the Interquartile Range (IQR) method. These outliers need to be addressed because they can cause clustering results to become unrepresentative. Third, data normalization using StandardScaler, a process of standardizing the data scale so that each variable has equal influence. This is important because the K-Means algorithm is sensitive to scale differences; without normalization, variables with larger values may dominate the clustering results.

Determination of Optimal Number of Clusters Using the Elbow Method

One of the challenges in using K-Means is determining the optimal number of clusters (K). For this purpose, the Elbow Method is used by running K-Means for several values of K (e.g., K = 1 to K = 10) and calculating the inertia value for each trial. The inertia value indicates the level of data compactness within a cluster, where a smaller value indicates better clustering quality. The results are then visualized in a graph to identify the "elbow" point, where the decrease in inertia begins to slow. This point indicates the most efficient number of clusters, as adding more clusters beyond this point does not yield significant improvement. Based on the analysis, the optimal number of clusters is K = 3.

K-Means Implementation with K = 3

Once the optimal K value is established, the K-Means algorithm is run with three clusters. To produce more stable clustering results and reduce the influence of random initialization, the k-means++ method is used to determine the initial centroids. This method selects center points that are far apart from each other, allowing the clustering process to begin from a better starting condition and achieve convergence more quickly. Subsequently, each district/city data point is assigned to one of the clusters based on the shortest Euclidean distance to the centroid.

Evaluation of Clustering Results

The evaluation of clustering results is performed using several indicators, namely the Silhouette Score and the inertia value. The Silhouette Score measures the quality of the clustering by assessing how well a data point fits within its own cluster compared to other clusters. The inertia value is used to determine the compactness of data within each cluster relative to its centroid. In addition, descriptive analysis is conducted for each cluster, covering the number of members, mean values, and regional distribution. The

clustering results are then assigned interpretive labels—low poverty, moderate poverty, and high poverty—to facilitate understanding of the grouping outcomes.

Visualization and Interpretation of Results

To facilitate interpretation of the clustering results, all findings are presented in several forms of visualization. The Elbow Method graph and Silhouette Score are displayed together to show the process of determining the optimal number of clusters, where the elbow point at K = 3 supported by a Silhouette Score of 0.699 provides the basis for selection. Scatter plots and histograms are used to depict the distribution of poor population data within each cluster, making the differences in characteristics between groups clearly visible. Boxplots are used to compare value distributions and detect the presence of outliers in each cluster. Additionally, horizontal bar charts display the top ten districts/cities with the highest number of poor residents in each category, making it easier to identify regions that require greater attention.

RESULTS AND DISCUSSION

This section describes the results of the clustering analysis of districts/cities in Indonesia based on BPS 2025 poverty data using the K-Means Clustering method. The discussion proceeds in stages, from the data preparation process, determination of the optimal number of clusters, algorithm application, to evaluation of the clustering results obtained.

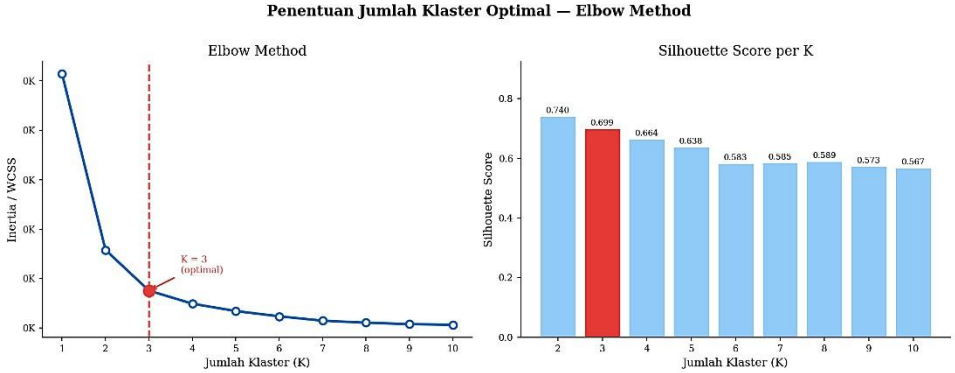
Data Preprocessing Results

The initial stage of this study involved understanding and preparing the data prior to the clustering process. The data used were sourced from BPS 2025, containing the number of poor residents (in thousands) across 514 districts/cities in 38 provinces of Indonesia, with two main variables: regional name and number of poor residents. Inspection revealed no missing data, allowing the dataset to be used directly. Statistically, the average number of poor residents is 46.41 thousand people, with a standard deviation of 51.26 thousand people. The minimum value recorded is 1.59 thousand people and the maximum is 401.86 thousand people, with the first quartile (Q1) at 13.83 and the third quartile (Q3) at 61.98 thousand people. This indicates a considerable variation in poverty conditions across regions.

During the outlier detection stage using the IQR method, the IQR value was 48.15 with an upper bound of 134.20 thousand people. From these results, 34 districts/cities were identified as outliers—regions with a number of poor residents significantly higher than average. Among them are Bogor Regency (401.86 thousand people), Tangerang Regency (265.90 thousand people), Brebes Regency (257.29 thousand people), Garut Regency (252.56 thousand people), and Bandung Regency (236.06 thousand people). Despite being classified as outliers, this data was retained as it reflects actual conditions in the field. The final step was data normalization using StandardScaler to ensure uniform data scale. After this process, the data had a mean of 0 and a standard deviation of 1, making it ready for the clustering process with more accurate results.

Determination of Optimal Number of Clusters

Determining the appropriate number of clusters is a critical step in K-Means analysis, as selecting too few or too many clusters can result in suboptimal groupings. In this study, the best number of clusters was determined using two methods: the Elbow Method and the Silhouette Score, as shown in Figure 2.



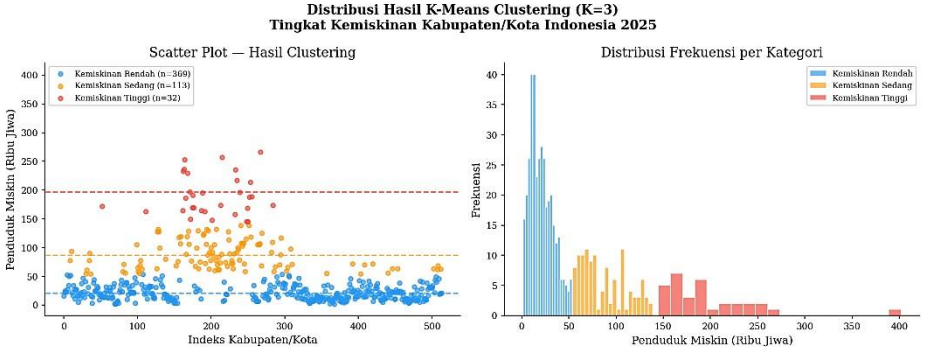
[Figure 2. Determination of Optimal Number of Clusters: Elbow Method and Silhouette Score]

Based on the Elbow Method graph in Figure 2, the inertia value shows a significant decrease from K = 1 to K = 2 of 69.3%, and from K = 2 to K = 3 of 52.1%. After K = 3, the decrease in inertia tends to flatten, indicating that adding more clusters no longer provides meaningful improvement. This point is identified as the elbow point and is located at K = 3.

This finding is supported by the Silhouette Score graph, where the value at K = 2 is 0.740 and at K = 3 is 0.699. Although the highest value is at K = 2, the selection of K = 3 is considered more representative as it forms three more informative poverty categories—low, moderate, and high. Moreover, the Silhouette Score at K = 3 still falls within the "good" category. Therefore, the optimal number of clusters used in this study is K = 3.

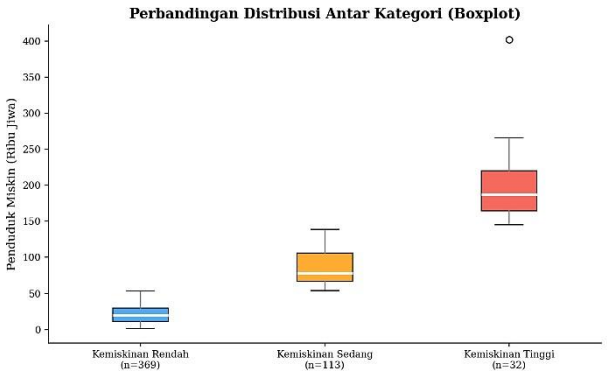
K-Means Clustering Implementation Results

After the optimal number of clusters was set at K = 3, the K-Means algorithm was run using k-means++ initialization to obtain more optimal initial centroids. The clustering process reached convergence after 10 iterations with a final inertia value of 75.72. Cluster labels were then sorted based on the average number of poor residents, from lowest to highest, yielding Cluster 1 (Low Poverty), Cluster 2 (Moderate Poverty), and Cluster 3 (High Poverty). The distribution of clustering results is presented in Figure 3 in the form of a scatter plot and histogram depicting the spread of data in each cluster.



[Figure 3. Distribution of K-Means Clustering Results (K=3): Scatter Plot and Frequency Histogram]

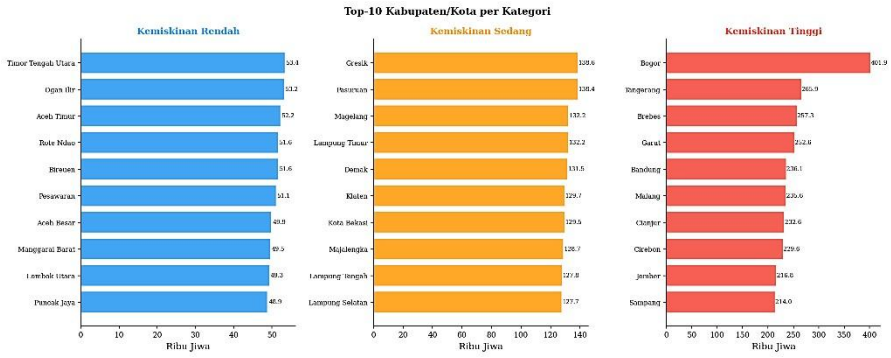
The scatter plot in Figure 3 (left) shows that the three clusters are formed with fairly clear separation. Cluster 1 (blue) dominates the lower portion of the graph with values below 53.44 thousand people. Cluster 2 (orange) is in the intermediate range, approximately 54 to 138 thousand people, while Cluster 3 (red) occupies the upper portion with significantly higher values. The histogram in Figure 3 (right) reinforces this pattern. Cluster 1 has a distribution that tends to be right-skewed with the highest frequency at low values, while Clusters 2 and 3 show a more spread distribution. To examine the differences in distribution between clusters more clearly, a boxplot visualization is displayed in Figure 4.



[Figure 4. Comparison of Distribution Across Categories (Boxplot)]

The boxplot in Figure 4 shows clear distributional differences across the three clusters. Cluster 1 (low poverty) has a narrow spread with a median of 19.45 thousand people, indicating that the values among its members are relatively uniform. Cluster 2 (moderate poverty) has a wider range with a median of 77.93 thousand people, reflecting greater data variation. Meanwhile, Cluster 3 (high poverty) shows the widest spread with a median of 186.94 thousand people, and contains one extreme outlier—Bogor Regency with a value of 401.86 thousand people. The absence of overlap between boxplots indicates that the three clusters have significant differences.

Details of regions with the highest number of poor residents in each category are presented in Figure 5.



[Figure 5. Top-10 Districts/Cities per Poverty Category]

Based on Figure 5, in the Low Poverty category, the regions with the highest poor population count (approaching the cluster upper boundary) include Timor Tengah Utara Regency (53.44 thousand people), Ogan Ilir Regency (53.21 thousand people), East Aceh Regency (52.23 thousand people), Rote Ndao Regency (51.64 thousand people), and Bireuen Regency (51.60 thousand people). In the Moderate Poverty category, regions

with the highest values are dominated by regencies in Java and Sumatra, such as Gresik Regency (138.55 thousand people), Pasuruan Regency (138.43 thousand people), Magelang Regency (132.21 thousand people), East Lampung Regency (132.17 thousand people), and Demak Regency (131.47 thousand people). Meanwhile, in the High Poverty category, Bogor Regency holds the highest position with a value far exceeding other regions (401.86 thousand people), followed by Tangerang Regency (265.90 thousand people), Brebes Regency (257.29 thousand people), and Garut Regency (252.56 thousand people). The dominance of Java-based regions in this cluster reflects the high poverty pressure in areas with large population densities.

CONCLUSION

This study successfully mapped poverty levels across all districts/cities in Indonesia by clustering 514 regions into three categories using the K-Means Clustering method based on BPS 2025 data. The optimal number of clusters was determined through a combination of the Elbow Method and Silhouette Score, yielding K = 3 with a Silhouette Score of 0.6987 (good category) and a final inertia value of 75.72. The clustering results revealed three significantly distinct categories: Low Poverty with 369 regions (71.8%), Moderate Poverty with 113 regions (22.0%), and High Poverty with 32 regions (6.2%). Although the number of regions in the High Poverty category is relatively small, this group contributes 26.4% of the total national poor population and is dominated by regions with high population density on the island of Java.

These findings demonstrate that data-driven clustering approaches can provide a clear picture of poverty patterns and thus serve as a basis for more targeted policy formulation. For future research, it is recommended to incorporate additional variables such as the poverty depth index, unemployment rate, and access to basic services, and to consider the use of multivariate clustering methods so that the mapping results become more comprehensive.

BIBLIOGRAPHY

Alfiansyah, D. N., Rahmayanti, V., Nastiti, S., & Hayatin, N. (2022). *Penerapan Metode K-Means pada Data Penduduk Miskin Per Kecamatan Kabupaten Blitar*. *4*(1), 49–58.

Amelia, M., Faqih, A., & Rinaldi, A. R. (2025). *PENERAPAN METODE K-MEANS CLUSTERING DALAM PEMETAAN KEMISKINAN KABUPATEN / KOTA DI TEPAT*. *13*(2).

Dasa, A. U., Pati, G. K., Ege, E. D., Stella, U., & Sumba, M. (2024). *Penerapan Algoritma K-Means Clustering Data Penduduk Miskin Berdasarkan Desa di Kecamatan Tana Righu*. *2*(6), 1–7.

Madaliyah, M., & Rohmah, S. (2024). *Upaya pengentasan kemiskinan di indonesia*. *3*(2), 269–273.

Matdoan, M. Y., Igo, L., Rumeon, R., Fadhilah, R., & Laamena, N. S. (2024). *Penerapan Algoritma K- Means Untuk Klusterisasi Kabupaten / Kota Berdasarkan Tingkat Kemiskinan di Kepulauan Maluku dan Papua Abstrak*. *10*(1), 1–9.

Prasetyo, T. L., Ramadhan, M. R., Fadhil, M. R., Wicaksono, D. M., Nurhakim, I. L., & Supiyan, D. (2025). *Klastering Data Kemiskinan Diindonesia Dari Tahun 2007-*

2017 , Menggunakan Kmeans Dan Decision Tree Python. *4*(2), 5062–5066.

Putra, L. G. R., & Anggrawan, A. (2021). *Pengelompokan Penerima Bantuan Sosial Masyarakat dengan Metode Grouping of Recipients of Community Social Assistance with the K-Means Method*. *21*(1), 205–214. https://doi.org/10.30812/matrik.v21i1.1554

Rahmadani, A., & Nursyahira. (2025). *Implementation of the K-Means Algorithm for Inventory Data Clustering Implementasi Algoritma K-Means Untuk Clustering Data Inventori*. *5*(1), 1–11.

Salmin. (2023). *Efektivitas KKS Dalam Penanggulangan Kemiskinan di Desa Setanggor*. *7*(1), 856–863. https://doi.org/10.58258/jisip.v7i1.4290/http

Sari, P., Efan, & Syahri, R. (2024). *ANALISIS CLUSTERING DATA PENDUDUK MISKIN MENGGUNAKAN ALGORITMA K-MEANS*. *8*(2), 2194–2199.

Sari, R., Yasin, M., & Asahan, U. (2025). *Penerapan Data mining untuk Clustering Kondisi Sosial Ekonomi Berdasarkan Kepemilikan Jaminan Kesehatan Menggunakan Algoritma untuk merumuskan kebijakan yang lebih efektif . Setelah kelompok-kelompok masyarakat*. *September*.

Sari, S. N., Pratama, B. G., & Prastowo, R. (2024). *Penggunaan Metode Elbow untuk Pemilihan Jumlah Klaster dalam Identifikasi Bahan Material Shelter Modular*. *2024*(November), 157–163.

Sianturi, M. W., Sirait, K. F., Nurfadillah, D., Shopia, Lubis, M., Muliani, F., & Fazrina, S. (2025). *Penerapan Analisis Cluster K-Means untuk Pengelompokan Kabupaten/Kota di Provinsi Sumatera Utara Berdasarkan Indikator Kemiskinan Tahun 2023*. *5*, 280–288.

Suhartini, & Yuliani, R. (2021). *Infotek : Jurnal Informatika dan Teknologi Penerapan Data Mining untuk Mengcluster Data Penduduk Miskin Menggunakan Algoritma K-Means di Dusun Bagik Endep Sukamulia Timur Infotek : Jurnal Informatika dan Teknologi Pendahuluan masalah kemiskanan belum bis*. *4*(1), 39–50.

Tawakal, Q., Effendi, M. M., & Majid, A. M. (2025). *ANALISIS TINGKAT KEMISKINAN DENGAN ALGORITMA K-MEANS MENGGUNAKANRAPIDMINER DITINGKAT KOTA KABUPATENDI JAWA TENGAH*. *7*(1), 112–119.