NUANSA INFORMATIKA ***p-ISSN :1858-3911* , *e-ISSN : 2614-5405***
Volume 19 Nomor 1, Januari 2025 https://journal.fkom.uniku.ac.id/ilkom

USING K-MEANS FOR DISTRICT-CITY POVERTY CLUSTERING IN INDONESIA

Abdul Mukhyidin*[1], Ahmad Faqih[2], Ade Rizki Rinaldi[3]
[1,2,3]STMIK IKMI Cirebon, Indonesia
E-mail: *[1]**abdulmukhyidin.19@gmail.com**

Abstract

Poverty is one of the main challenges faced by the government in its efforts to improve people's welfare. Identifying regions based on the poverty line level is an important step to ensure well-targeted interventions. This study aims to categorize districts/cities based on poverty levels using the K-Means Algorithm, so that it can be a guide in data-based policy making. The research method starts with data collection, data selection process to handle missing values using the replacement method. Determination of the optimal number of clusters was done using Within Sum of Squares (WSS) to ensure that each region was grouped into clusters based on their level of similarity, which showed that three clusters were the ideal number. An evaluation of the clustering results was conducted to ensure the stability and accuracy of the clustering. The results show that the districts/municipalities are divided into three clusters based on the poverty line level, namely cluster 0 with a high poverty line level (241 regions), cluster 1 with a medium poverty line level (247 regions), and cluster 2 with a low poverty line level (90 regions). This study concludes that the K-Means Algorithm is effective in clustering regions based on poverty levels, providing a strong basis for data-driven decision-making. Future research is recommended to use more diverse data and cover more indicators, such as education level, access to health services, or infrastructure quality.

Keywords: Poverty, K-Means Algorithm, Within Sum of Squares (WSS), Clustering.

Submission: 9 December 2024 Accepted: 31 January 2025 Published: 31 January 2025

1. INTRODUCTION

Poverty is one of the main problems faced by many developing countries, including Indonesia. It affects various aspects of people's lives, such as health, education, and access to basic services. The determination of the poverty line is often used as a reference in various social assistance programs. However, although the government has implemented a number of policies to reduce the poverty rate, the effectiveness of social assistance distribution remains a challenge, especially due to the heterogeneity of economic conditions in different regions.

Along with the development of information technology and data science, the K-Means clustering method has developed into one of the effective tools for clustering regions based on poverty lines. Research by [1] showed the success of this algorithm in clustering patient disease data based on similar characteristics. Other studies by [2] and [3] highlighted the ability of K-Means in analyzing socio-economic data, including public health data. These results support the potential use of K-Means in helping the government optimize the allocation of social assistance programs more effectively.

The K-Means method works by dividing a dataset into a number of clusters based on data similarity, using the centroid as the cluster center. Research by [4] shows that this algorithm is effective in improving the accuracy of social assistance distribution by grouping regions according to their economic characteristics. In addition, a study by [5] highlighted the use of K-Means to analyze population density in Jakarta, providing insights for city policy planning.

In a broader context, the application of K-Means is also successfully used for various analyses. Research by [6] used K-Means for clustering student learning outcomes, while research by [7] focused on traffic accident data. [8], demonstrated the application of K-Means in clustering sales patterns, which provided inspiration for its application in socio-economic analysis. In addition, research by [9] demonstrated data clustering for student major selection relevant to education policy planning.

Other studies have shown the success of K-Means in more specific fields. For example, research by [10] analyzed violence-prone areas in West Java, while [11] integrated K-Means with CRISP-DM method for data clustering.

cake sales. Research by [12] shows the application of K-Means in determining the amount of tuition fees based on the economic group of students. In addition, a study by [13] applied K-Means in analyzing retail sales patterns to assist strategic planning.

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

Despite having advantages such as computational efficiency and the ability to process large datasets, the K-Means method has challenges. Research by [14] mentioned that this algorithm is vulnerable to outliers and uneven data distribution. To overcome this, other methods such as K-Medoids and Fuzzy C-Means can be used as a comparison. Fuzzy C-Means, as stated by [15] provides flexibility by allowing one data to be in more than one cluster. The study by [16] even shows a comparison of this algorithm with Naïve Bayes to identify optimal use cases.

This study aims to develop a poverty line clustering model based on districts or cities using the K-Means method, in order to optimize the distribution of social assistance programs. The data used is the poverty line in Rupiah per capita per month from various districts/cities in Indonesia for the years 2022 to 2024. This data will be processed using RapidMiner software, with a focus on identifying regions with similar economic conditions. The research results are expected to contribute to the effectiveness of social assistance programs as well as the development of data-based decision support systems.

2. RESEARCH METHODS

This research uses a quantitative method based on secondary data obtained from the Central Bureau of Statistics (BPS), specifically district/city poverty line data for the period 2022 to 2024. This research aims to cluster districts/cities in Indonesia based on the poverty line using the K-Means Clustering method, in order to support the optimization of social assistance programs. The following are the stages of the methodology carried out in this research:

Figure 1: Stages of Research

a. Finding the Problem
The initial stage of research begins with the identification of the main problem to be solved. In the context of this research, the problem identified is the inequality in the distribution of social assistance caused by the lack of a comprehensive understanding of the variations in poverty lines in various districts/cities in Indonesia. This inequality has the potential to cause social assistance programs to be less well-targeted.

b. Literature Study
Before entering the analysis stage, a literature review was conducted to understand the approaches that have been used in previous studies. Some of the key literature used include:
a. Journal that discusses the application of the K-Means Algorithm in regional clustering [17].
b. A journal that compares the K-Means method with K-Medoids in data clustering [14]. These literatures help provide guidance in the use of the K-Means Algorithm, including in determining the optimal number of clusters and how to evaluate the clustering results.

c. Data Collection
The data used in this study comes from the official portal of BPS (Central Bureau of Statistics), which includes information on the poverty line for each district/city for three years (2022-2024). This data is downloaded in CSV or Excel format and provides a comprehensive picture of the variation in poverty lines between regions in Indonesia.

d. Data Pre-Processing
The downloaded data will then be processed through a pre-processing stage. This includes data cleaning to remove outliers or missing data, normalization to equalize the scale between variables, as well as interpolation or average filling methods to overcome missing data.

e. Use of RapidMiner Tools
The modeling process is done using RapidMiner software which facilitates the implementation of K-Means and helps in selecting parameters as well as evaluation of clustering results.

f. K-Means Method
The method used in this research will use K-Means to cluster the data.

g. Conclusion and Recommendation

p-ISSN :1858-3911 **,** ***e-ISSN : 2614-5405***
https://journal.fkom.uniku.ac.id/ilkom

The conclusion of this research will include recommendations for government policies in optimizing social assistance programs based on the results of poverty line clustering.

3. RESEARCH RESULTS

3.1 Data selection

At this stage, it is possible to select missing data, delete columns or label the data that is needed.

1	ACEH	579227	627534	661227
2	Simeulue	493303	538693	576505
3	Aceh Singkil	518951	568691	609322
4	Aceh Selatan	446224	494565	528243
5	Aceh Tenggara	430825	471301	496074
6	Aceh Timur	491550	539934	557943
7	Aceh Tengah	533810	584863	626090
8	Aceh Barat	558638	616091	644009
9	Aceh Besar	519320	564431	586860
10	Pidie	535713	579450	607117
11	Bireuen	451163	486667	494866
12	Aceh Utara	420615	454361	473719
13	Aceh Barat Daya	430297	474127	516947
14	Gayo Lues	470221	514836	557201
15	Aceh Tamiang	508599	555387	583294
16	Nagan Raya	544300	594374	635820
17	Aceh Jaya	468854	515244	536092
18	Bener Meriah	512111	551868	591421

Figure 2. Data selection

3.2 Preprocessing

1. The first step is to prepare Read CSV to import the data to be tested..

Figure 3. Read CSV

2. Next is to input data by double-clicking on Read CSV to find the data to be tested.

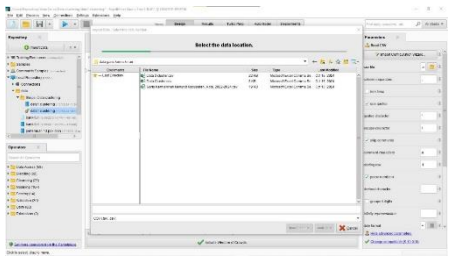

Figure 4. Input Data

3. At this stage, it is optional if there are empty values or data and the data should not be deleted, then it can be done by adding Replace Missing Values to replace

the missing values.
Figure 5. Replacing missing values

4. Before continuing data testing, it is necessary to test the number of clusters by testing WSS cluster 2 to cluster 6, to find out the optimal number of clusters.

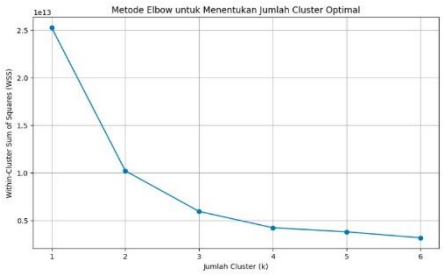

Figure 6. WSS

5. Setting on the K-Means Clustering parameter menu, the k value is set, where k becomes the value that will be used to determine the number of clusters that will be created. Here, the number of clusters that will be created is as many as 3 clusters, namely the results of high poverty lines, medium poverty lines, and low poverty lines.

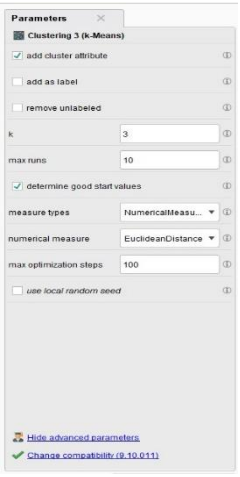

Figure 7. K-Means Parameters

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

6. Then add Cluster Distance Performance to see the performance results of the tested data.

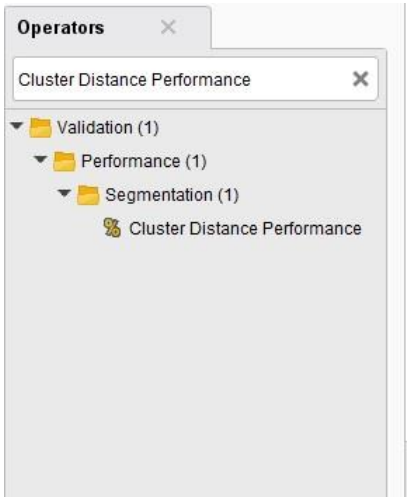

Figure 8. Distance Performance

3.3 Transformation

In this study, the data transformation stage was not carried out because the data used was already in a suitable format and ready for analysis. The data has gone through preliminary processes such as cleaning and validation to ensure there are no missing or invalid values. Data transformation is usually done to transform data into a format that is easier to understand or process by certain algorithms, such as normalization, standardization, or recoding. However, in this study, the characteristics of the data did not require further changes, so the transformation step could be skipped. The decision not to transform the data is also based on the consideration that the K-Means algorithm used is already quite effective in handling data in its original form. Thus, the focus of the research is more directed towards clustering analysis and interpretation of the results without modifying the original data, to ensure that the patterns and information contained in the data remain accurate and relevant.

3.4 Data Mining

Next, connect all K-Means and Performance clustering data to see the output towards the result.

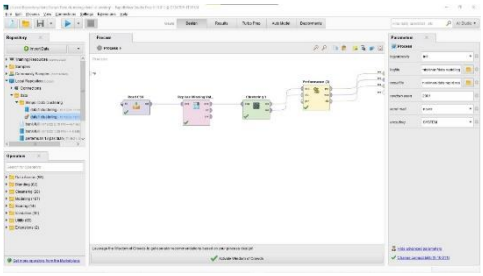

Figure 9. Process circuit

a. Cluster Model

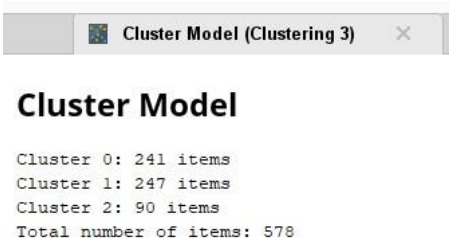

Figure 10. Cluster Model

Cluster model displays the results of the number of each cluster, cluster 0 totals 241, cluster 1 totals 247, cluster 2 totals 90 and displays the total amount of data..

b. Cluster Distance Performance

Figure 11. Distance Performance

Cluster Distance Performance displays a performance of the test data results and the average value in the distance between cluster centers centroid. The smaller the distance between cluster centers the better, because it shows that the data in one cluster is more homogeneous or similar to each other.

3.5 Evaluation

From the entire output produced, the data can be further explored by displaying visualizations, statistics through various models, making it easier to understand and analyze the data, with visualizations, statistics as follows:

p-ISSN :1858-3911 **,** ***e-ISSN : 2614-5405***
https://journal.fkom.uniku.ac.id/ilkom

a. Visualisasi Pie

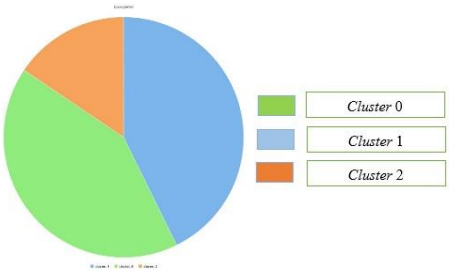

Figure 12. Pie Chart

This pie chart shows the size comparison of cluster 0, cluster 1 and cluster 2, where cluster 0 is marked in green, cluster 1 is marked in blue, and cluster 2 is marked in orange.

b. Visualisasi Bar

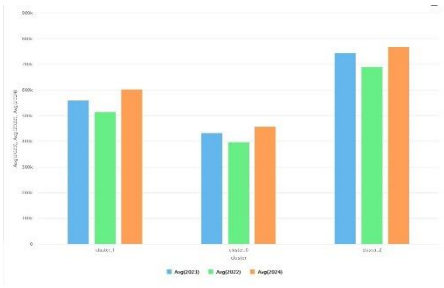

Figure 13. Bar (Column)

Figure 13 shows the results of the clustering analysis illustrating the changes in average income values across the three clusters from 2022 to 2024.DISCUSSION

In this process, the K-Means Algorithm is used to group districts/cities based on the poverty line level. The first step is to prepare data in the form of information on income levels, the number of poor people, and other relevant indicators. After the data is collected, checking is done to ensure that there are no empty values or inconsistent data. If found, the data is supplemented with a method of replacing missing values.

Once the data was ready, the number of clusters was analyzed using the elbow method to determine the optimal number of clusters. This analysis showed that k 3 was the ideal number, with three categories: high poverty, medium poverty, and low poverty.

The K-Means process starts by randomly selecting three initial centroids. Then, each region is calculated for its distance to the centroid using the Within Sum of Squares (WSS) method. In this case, WSS measures how close the data in the cluster is to its cluster center. Regions are grouped into the cluster with the smallest WSS value. This process is repeated until the centroid position stabilizes and the WSS value in each cluster does not change significantly.

The clustering results are evaluated to ensure the clustering is appropriate. This evaluation is done by calculating the average WSS, which describes the extent to which data in a cluster are close to each other and separate from other clusters.

Cluster 0 includes 241 areas with high poverty rates. These areas have a large number of poor people and a low average income.

Cluster 1 consists of 247 regions with a medium poverty rate. Economic conditions in these areas are more stable than in Cluster 0, but still require attention.

Cluster 2 includes 90 regions with a low poverty rate. These regions tend to have high average incomes.

These results are visualized through various methods:

1. Pie Chart

Pie charts are used to visualize the proportion of each cluster, making it easier to understand the distribution of data based on predetermined categories. This visualization provides a clear picture of the comparison of the number of regions in each cluster, so that it can help in identifying the groups that have the largest and smallest proportions.

Using pie charts, we can quickly see how the data is spread between clusters. For example, a cluster with a high poverty rate may have the largest proportion, indicating that more areas need special attention. Conversely, a cluster with a low poverty rate that has a smaller proportion illustrates that only a few areas are in this category. In addition, this diagram also helps in conveying information in a simple and intuitive manner, making it an effective tool for data presentation to decision-makers or others who do not have a technical background. Thus, pie charts play an important role in presenting analysis results in a visual and easy-to-understand manner.

1. Bar (Column)

The results of the clustering analysis visualized in the form of bar/column diagrams illustrate the changes in average income values across the three cluster groups from 2022 to 2024. Overall, the graphs show an improving economic trend characterized by an increase in community income in each cluster.

4. CONCLUSIONS

In this study, based on the results and discussion, it can be concluded that

1. Application of K-Means Algorithm for Regional Clustering

Using the Rapidminer tool, the K-Means algorithm was successfully applied to cluster districts/municipalities based on the poverty line, with the process of analyzing the number of clusters using the Within Sum of Squares (WSS) method, three optimal clusters were determined, reflecting regions with high, medium, and low poverty rates. The clustering process using the K-Means method allows for fast and accurate clustering with an evaluation that ensures that the regions within the clusters are highly similar. The results of this clustering have mapped the regions based on the cluster categories, making decision-making easier.

2. Visualization and integration benefits

The clustering results are visualized using pie and bar charts, to provide a clearer picture of the data distribution. The pie charts help understand the proportion of regions in each cluster, while the bar charts show the change in average income from 2022 to 2024, reflecting both the trend of economic improvement and inequality between clusters. This visualization helps future research or related parties in determining the priority of social assistance programs.

5. SARAN

This research provides important insights into the grouping of regions based on the poverty line, which is expected to be a reference for knowing the economic development of each region from year to year. The suggestions given aim to ensure that the results of data analysis do not stop at statistical understanding but are also applied in real policies. That way, socio-economic disparities can be reduced, and people's welfare can be improved equally.

Future research is recommended to use more diverse data and cover more indicators, such as education levels, access to health services, or infrastructure quality.

Local governments are expected to use the clustering results to design specific programs for each cluster, so that the programs are more effective and targeted. Conduct regular monitoring and evaluation of the programs implemented in each cluster to ensure their sustainability and impact on the community.

By involving data in policy planning and implementation, the government can create more inclusive and equitable economic growth, and ensure that prosperity is felt by all levels of society.

With this, it is hoped that inequality can be reduced, and all communities, especially in low-income areas, can feel the benefits of more inclusive and sustainable economic growth.

REFERENCES

[1] R. Anggraini, E. Haerani, J. Jasril, and I. Afrianty, "Pengelompokkan Penyakit Pasien Menggunakan Algoritma *K-Means*," *Jurnal Riset Komputer*, vol. 9, no. 6, p. 1840, Dec. 2022, doi: 10.30865/jurikom.v9i6.5145.

[2] M. Faisal, N. Fajriana, and Z. Fitri, "Information and Communication Technology Competencies *Clustering* for students for Vocational High School Students Using *K-Means Clustering* Algorithm," *International Journal of Engineering, Science & InformationTechnology (IJESTY)*, vol. 2, pp. 111–120, 2022, doi: 10.52088/ijesty.v1i4.318.

[3] T. Wahyudi and T. Silfia, "Implementation Of Data Mining Using *K-Means Clustering* Method To Determine Sales Strategy In S&R Baby Store," *Journal of Applied Engineering and Technological Science*, vol. 4, no. 1, pp. 93–103, 2022.

[4] B. Baskoro, A. Gunaryati, and A. Rubhasy, "Klasifikasi Penduduk Kurang Mampu Dengan Metode *K-Means* untuk Optimalisasi Program Bantuan Sosial," *Jurnal Informatika, Manajemen dan Teknologi*, vol. 25, no. 1, pp. 41–48, Jun. 2023, doi: 10.23969/infomatek.v25i1.7271.

[5] F. Handayanna and S. Sunarti, "Penerapan Algoritma *K-Means* Untuk Mengelompokkan Kepadatan Penduduk Di Provinsi DKI Jakarta," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 50–55, Mar. 2024, doi: 10.52158/jacost.v5i1.477.

[6] S. Anwar, T. Suprapti, G. Dwilestari, and I. Ali, "Pengelompokkan Hasil Belajar Siswa Dengan Metode *Clustering K-Means*," *Jurnal Sistem Informasi dan*

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

Teknologi Informasi), vol. 4, no. 2, pp. 60–72, 2022.

[7] T. Kurniawan and M. Jajuli, "*Clustering* Data Kecelakaan Lalu Lintas di Kecamatan Cileungsi Menggunakan Metode *K-Means*," *Generation Journal*, vol. 6, no. 1, pp. 2580–4952, 2022.

[8] S. Pujiono, R. Astuti, and F. M. Basysyar, "Implemetasi Data Mining Untuk Menentukan Pola Penjualan Produk Menggunakan Algoritma *K-Means Clustering*," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 1, pp. 615–620, 2024.

[9] J. Jemakmun and R. A. D. S. Purboyo, "Data *Clustering* Recommendations For Selection Student Majors To Higher Edication Using The *K-Means* Method (Case Study of SMAN 2 Palembang)," *Journal Of Informatics And Telecommunication Engineering*, vol. 6, no. 2, pp. 367–377, Jan. 2023, doi: 10.31289/jite.v6i2.7911.

[10] R. Rahma and R. Mufidah, "Pengelompokan Daerah Rawan Kekerasan Terhadap Perempuan Dan Anak Di Jawa Barat Menggunakan Algoritma *K-Means*," *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 7, pp. 850–857, 2022.

[11] M. R. Muttaqin, T. I. Hermanto, and M. A. Sunandar, "Penerapan *K-Means Clustering* Dan Cross-Industry Standard Process For Data Mining (Crisp-Dm) Untuk Mengelompokan Penjualan Kue," *Jurnal Ilmiah Ilmu Komputer dan Matematika*, vol. 19, no. 1, pp. 38–53, 2022, [Online]. Available: https://journal.unpak.ac.id/index.php/komputasi

[12] K. Haris, D. Sarjon, and S. Sumijan, "Data Mining Menggunakan Metode *K-Means Clustering* Untuk Menentukan Besaran Uang Kuliah Tunggal," *Journal of Applied Computer Science and Technology*, vol. 1, no. 2, pp. 80–89, Dec. 2020, doi: 10.52158/jacost.v1i2.102.

[13] C. A. Sugianto and T. P. O. R. Bokings, "*K-Means* Algorithm For *Clustering* Poverty Data in Bangka Belitung Island Province," *Journal of Computer Networks, Architecture, and High-Performance Computing*, vol. 3, no. 1, pp. 58–67, Feb. 2021, doi: 10.47709/cnahpc.v3i1.934.

[14] A. Supriyadi, A. Triayudi, and I. D. Sholihati, "Perbandingan Algoritma *K-Means* Dengan K-Medoids Pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas," *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 6, pp. 229–240, 2021.

[15] N. N. Hasanah and A. S. Purnomo, "Implementasi Data Mining Untuk Pengelompokan Buku Menggunakan Algoritma *K-Means Clustering* (Studi Kasus : Perpustakaan Politeknik LPP Yogyakarta)," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 4, no. 2, pp. 300–311, Jul. 2022, doi: 10.47233/jteksis.v4i2.499.

[16] N. Nurhachita and E. S. Negara, "A Comparison Between Naïve Bayes and The *K-Means Clustering* Algorithm for The Application of Data Mining on The Admission of New Students," *Int J Comput Appl*, vol. 17, no. 8, pp. 43–48, Mar. 2020, doi: 10.5120/2237-2860.

[17] F. Sembiring, O. Octaviana, and S. Saepudin, "Implementasi Metode *K-Means* Dalam Pengklasteran Daerah Pungutan Liar Di Kabupaten Sukabumi (Studi Kasus : Dinas Kependudukan Dan Pencatatan Sipil)," *Jurnal Tekno Insentif*, vol. 14, no. 1, pp. 40–47, Apr. 2020, doi: 10.36787/jti.v14i1.165.