Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 1 2023, 1 - 6
DOI: *https://doi.org/10.22487/27765660.2023.v3.i1.16435*
eISSN: 2776-5660

GROUPING OF POVERTY IN INDONESIA USING K-MEANS WITH SILHOUETTE COEFFICIENT

Gustriza Erda[1*], Chairani Gunawan[1], Zulya Erda[2]
[1]Universitas Riau
[2]Poltekkes Kemenkes Tanjung Pinang

***e-mail**: *gustrizaerda@lecturer.unri.ac.id*

ABSTRACT

Poverty is an enormous problem in numerous nations including Indonesia. Poverty can be measured using several indicators, including the unemployment rate, the percentage of poor people, expenditures per capita, and the poverty line. The purpose of this study is to categorize Indonesian provinces based on poverty indicators in 2021 using K-Means with the Silhouette Coefficient approach. Based on the silhouette coefficient approach, there are two clusters that are created. The first cluster is a high-poverty-rate regional group that includes the provinces of Aceh, Bengkulu, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Gorontalo, Maluku, West Papua, and Papua. On the other hand, the second cluster is an association of regions with a low poverty rate, and it includes 25 provinces. The greater number of provinces in the low poverty rate cluster implies that the poverty rate in Indonesia in 2021 is included in the low category.

Keywords: *Cluster, K-Means, Poverty, Silhouette Coefficient.*

Cite: Erda, G., Gunawan, C., & Erda, E. (2023). *Grouping of Poverty in Indonesia Using K-Means with Silhouette Coefficient*. *Parameter: Journal of Statistics*, *3*(1), 1-6, https://doi.org/10.22487/27765660.2023.v3.i1.16435.

[*]Received: Jun 2023; Reviewed: Jun 2023; Published: Jun 2023

INTRODUCTION

Poverty is an economic inability to meet basic food and non-food needs as measured by expenditure. Poverty has an impact on various sectors of people's lives such as the economic and social sectors. Disruption to this sector of people's life will be directly related to the level of poverty of the people in a region (BPS, 2021).The increase in the poverty rate can have an impact on increasing the unemployment rate. The percentage of poor people, poverty line, unemployment, and per capita expenditure can provide an overview of poverty in an area.

In 2021, the Central Statistics Agency Noted that the open unemployment rate in August 2021 in Indonesia was 6.49%. The Open Unemployment Rate is an indicator used to measure labor that is not absorbed by the labor market. The increase in unemployment that occurred was caused by layoffs carried out by most companies. In addition, the Covid-19 pandemic has also caused some residents to lose or stop working and become unemployed (BPS, 2021). An increase in the unemployment rate can result in an increase in the poverty rate.

Indonesia consists of 34 provinces, each province has its own characteristics that support the level of welfare of the people in its territory. The economic and social conditions of each region have different figures. Characteristics based on circumstances that affect poverty in each province can be grouped based on their resemblance to one another. Grouping each province is useful for seeing how the poverty rate is in the region so that the government can implement good policies in national development activities, especially on problems caused by the Covid-19 pandemic. The process of grouping these regions can be done by cluster analysis.

Cluster analysis is an important data mining method for finding information in multidimensional data (Kassambara, 2017). Cluster analysis is a grouping of objects or cases into smaller groups where each group contains objects that are similar to one another (Talakua et al., 2017). The purpose of cluster analysis is to identify patterns or groups of similar objects in the data set used. Cluster analysis is a method of grouping objects based on the results of information from data between these objects. Data grouping is done by selecting similar objects in a group from each cluster that represents the type or category of objects. In cluster analysis, the k-means method is a method for dividing data into several clusters. When the variables used have a high degree of similarity, they can be grouped into a cluster (Rahman et al., 2017).

Clustering methods are generally divided into two, namely hierarchical clustering and non-hierarchical clustering. Hierarchical clustering performs grouping of data from two or more objects that have the closest similarity. Then the process is forwarded to another object that has a second closeness and so on so that clusters will form a kind of tree where there is a hierarchy (clear level) between objects, whereas in non-hierarchical clustering, determine the desired number of clusters in advance. After the number of clusters is known, the cluster process will be carried out without following the hierarchical process (Agarwal, 2013). The hierarchical clustering method carries out a grouping process through tiered stages, while the non-hierarchical clustering method groups n objects as many as c clusters that have been determined beforehand. The non-hierarchical clustering method that is often used is K-Means Clustering (Izzadin, 2020).

Research using the k-means method has been carried out in grouping provinces in Indonesia based on the food poverty line which produces 3 clusters namely cluster 1 consisting of 6 provinces as a group that has a high level of food poverty line, cluster 2 produces 16 provinces as a group that has a high level of food poverty. moderate food poverty line, and cluster 3 produces 12 provinces as groups that have low levels of food poverty line (Aprilia & Sembiring, 2021). In addition, cluster analysis has also been studied in grouping regencies/cities in West Java Province based on poverty indicators using the k-means algorithm which produces 5 clusters (Febianto & Palasara, 2019). On the other hand, Setiawan & Zahra (2023) used time series-based clustering to classify poverty in Indonesia in which there are 3 clustering groups were created, that are low, medium and high poverty categories.

Based on the background above, by developing various ideas and results from previous research, this research will conduct a cluster analysis of provinces in Indonesia based on poverty indicators in 2021 using the k-means method with silhouette coefficient. Research to classify provinces based on the unemployment rate, the percentage of poor people, expenditures per capita, and the poverty line using a k-means with a silhouette coefficient has never been done before, so it is a novelty in this research. This research is expected to be able to see the relationship between the variables used to reduce poverty rates and provide information about the clusters of each province in Indonesia which are seen based on poverty indicators in 2021.

MATERIALS AND METHODS

The data used is secondary data obtained from the publication of the Central Bureau of Statistics. There are four variables used as a measure of poverty in each province in Indonesia, namely the open unemployment rate, percentage of poor population, per capita expenditures, and the poverty line. Data were analyzed using k-means clustering using the silhouette coefficient method. K-Means Clustering is a non-hierarchical clustering method that forms data in one or more clusters. The grouping process is based on determining the initial number of groups by defining the initial centroid value. The k-means analysis uses an iterative process to obtain cluster results (Fathia et al., 2016). This method attempts to divide the data into groups so that data with the same characteristics is included in one group, while data with different characteristics is included in another group. In addition, silhouette coefficient is a method used to see the quality and strength of clusters, how well an object is placed in a cluster, the optimal k value in this method can be seen based on the highest point formed on the graph.

The initial stage in the k-means analysis is to determine the value of k to produce clusters. The k-means method will randomly choose the pattern k as the centroid starting point. The number of iterations to reach the cluster centroid will be affected by the initial cluster centroid value randomly if the position of the new centroid does not change. The value of k which is chosen to be the initial center, will be calculated using the Euclidean Distance formula, which is to find the shortest distance between centroid points. Data that has the shortest or closest distance to the centroid will form a cluster. The steps of the k-means clustering method according to Dwitri et al., (2020) are as follows:

a) Determine the number of clusters (k) using silhouette coefficient;
b) Determine the centroid randomly in the first stage;
c) Calculate the Euclidean distance using the following formula:

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (1)$$

Where,
$d(x,y)$: Data distance to x to cluster center y
x_i : data x on the ith observation
y_i : center point to y the i observation
n : the number of observations

d) Recalculating the centroid with cluster membership formed by calculating the average value of all data in a cluster using the following formula:

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj} \quad (2)$$

Where,
V_{ij} : average centroid in the ith cluster for the jth variable
N_i : the number of members of the i cluster
i, k : index of cluster
j : variable index
X_{kj} : data value for the k for the j variable in the cluster

e) Recalculate each object using the new centroid. If the cluster members do not change anymore, then the clustering process is declared complete.

RESULTS AND DISCUSSION

Descriptive statistics

The distribution of poverty indicator data in the form of the open unemployment rate, percentage of poor population, expenditure per capita, and the 2021 poverty line from 34 provinces is explained in Figure 1. It can be seen from the figure 1 that the distribution of data on the open unemployment rate, percentage of poor population, and poverty line tends to be symmetrical because the median values for the three indicators are in the middle of 1st and 3rd quartiles. Meanwhile, per capita expenditure is skewed to the right, indicating that the data on the average distribution of per capita expenditure is greater than the median value.

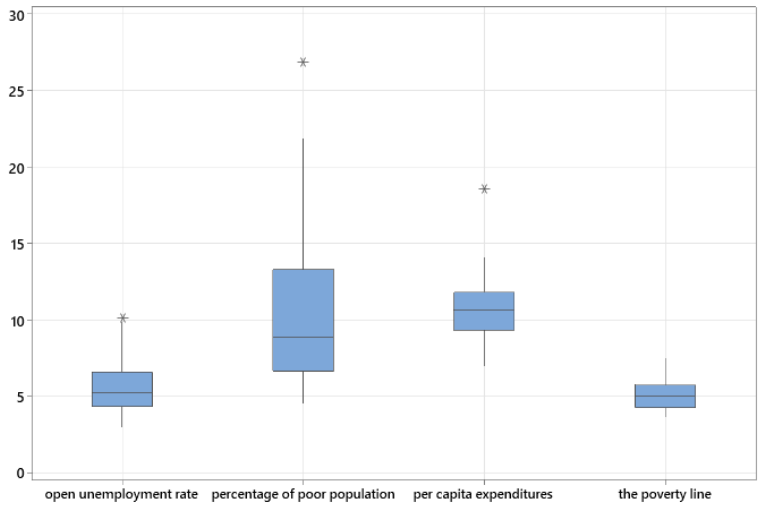

Figure 1. Poverty Indicator Data Boxplot in Indonesia

Figure 1 also shows that there are upper outliers for all poverty indicators used. The open unemployment rate indicator has outlier data originating from the Riau Islands Province and the variable percentage of poor people has outlier data originating from the Papua Province. On the other hand, the per capita expenditure variable has outlier data originating from DKI Jakarta Province, and the poverty line variable has outlier data originating from Bangka Belitung.

Clustering

K-Means Clustering is a non-hierarchical clustering method that forms data in one or more clusters. The grouping process is based on determining the initial number of groups by defining the initial centroid value. The k-means analysis uses an iterative process to get cluster results (Fathia & Rahmawati, 2016). This study uses the silhouette coefficient method to estimate the optimal number of clusters to be formed. The determination of the value of *k* can be seen based on the highest line formed in the resulting plot. The graphical results of the silhouette coefficient method can be seen in Figure 2. In Figure 2 it can be seen that the optimal cluster formed based on the highest graph is when the value $k = 2$. Therefore, using the silhouette coefficient method, the optimal *k* value is obtained when it is at $k = 2$.

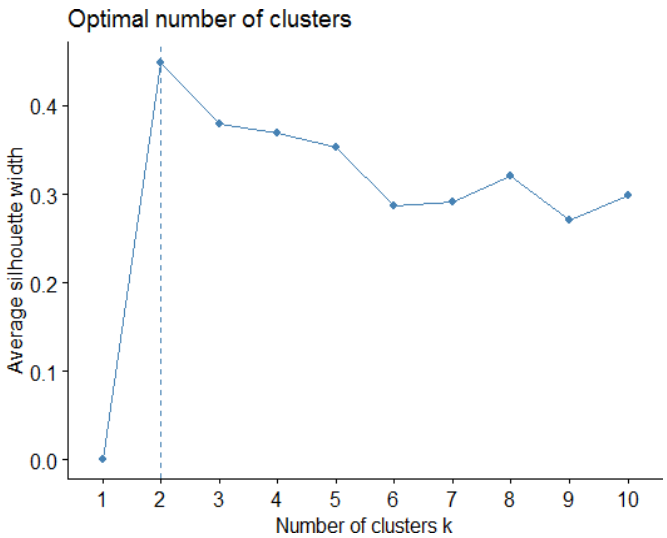

Figure 2. Silhouette Coefficient Graph

By using the k-means method, it is found that the grouping of provinces based on poverty indicators in 2021 produces 2 groups. Cluster 1 consists of 9 provinces and cluster 2 consists of 25 provinces. The results of clustering k-means can be seen in Table 1. Table 1 shows that cluster 1 is a cluster consisting of two provinces in the western region and 7 others in the eastern region. None of the provinces in cluster 1 originate from the island of Java.

Table 1. Group of Province based one K-Means Clustering

Cluster	Provinces
1	Aceh, Bengkulu, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Gorontalo, Maluku, West Papua and Papua.
2	North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Lampung, Kep. Bangka Belitung, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi and North Maluku.

In order to facilitate the interpretation of the clustering results, the characteristics of each cluster formed are described based on the average value of each of the variables studied. The average value based on the unemployment rate, the percentage of poor people, spending per capita, and the poverty line can be is seen in Table 2. Based on the total average of the four poverty indicators, it is found that the total average in cluster 1 greater than cluster 2, so that cluster 1 is categorized as a group with a high poverty rate and cluster 2 is categorized as a group with a low poverty rate. In detail, the interpretation of the two clusters is as follows:

1. Cluster 1 (High Poverty Rate)
   The results of cluster 1 are the grouping of poverty indicators based on high poverty rates. Based on the average value of cluster 1 in table 4.3, the unemployment rate variable has an average value of 4.00, the percentage of poor people is 17.33, per capita expenditure is 8.44, and the poverty line is 4.55. Therefore, cluster 1 is categorized as a cluster with a high poverty rate.
2. Cluster 2 (Low Poverty)
   The results of cluster 2 are the grouping of poverty indicators based on low poverty levels. Based on the average value of cluster 2 in table 4.3, the unemployment rate variable has an average value of 5.52, the percentage of poor people is 7.68, per capita expenditure is 10.96, and the poverty line is 4.64. Therefore, cluster 2 is categorized as a cluster with a low poverty rate.

Table 2. The Average Value of Cluster Results in Each Variable

Indicator	***Cluster* 1**	***Cluster 2***
open unemployment rate	4,00	5,52
percentage of poor population	17,33	7,68
per capita expenditures	8,44	10,96
poverty line	4,55	4,64

Furthermore, a visualization of the grouping of provinces based on cluster results using K-Means is described in Figure 3.

Figure 3. Visualization of Cluster

The red plots are members of cluster 1 which consists of 9 provinces and the blue parts of the plot are members of cluster 2 which consists of 25 provinces. Based on the cluster results obtained, it can be seen that most of the data gathered in cluster 2 with the diversity value generated by dimension 1 of 54% and dimension 2 of 23.1%. Overall, the variance that can be explained by these two dimensions is 77.1% of the total actual data variance.

CONCLUSION

Based on the findings of the k-means clustering approach, it can be inferred that by utilizing the silhouette coefficient, two groups are established, namely the group with a high level of poverty and the group with a low level of poverty. Aceh, Bengkulu, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Gorontalo, Maluku, West Papua, and Papua are the provinces with the highest poverty rates. Meanwhile, 25 additional provinces are among those with low poverty levels. The increased number of members joining the low poverty rate cluster indicates that the poverty rate of Indonesia in 2021 will be in the low category.

REFERENCES

Agarwal, S. (2013). *Data Mining: Data Mining Concepts and Techniques. 2013 International Conference on Machine Intelligence and Research Advancement*, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Aprilia, K., & Sembiring, F. (2021). *Analisis Garis Kemiskinan Makanan Menggunakan Metode Algoritma K-Means Clustering. Seminar Nasional Sistem Informasi Dan Manajemen Informatika*, 1–10.

BPS. (2021). Keadaan Ketenagakerjaan Indonesia Agustus 2021. BPS.

Dwitri, N., Tampubolon, J. A., Prayoga, S., R.H Zer, F. I., & Hartama, D. (2020). *Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Indonesia*. *Jurnal Teknologi Informasi*, 4(1), 128–132. https://doi.org/10.36294/jurti.v4i1.1266

Fathia, A., Rahmawati, R., & Tarno, R. (2016). *Analisis Klaster Kecamatan Di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward Dan Single Linkage*. Jurnal Gaussian, 5(4), 801–810.

Febianto, N. I., & Palasara, N. (2019). *Analisa Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018*. Jurnal Sisfokom (Sistem Informasi Dan Komputer), 8(2), 130–140. https://doi.org/10.32736/sisfokom.v8i2.653

Izzadin, F. (2020). *Optimasi Jumlah Cluster K-Means dengan Metode Elbow dan Silhouette Pada Produktivitas Tanaman Pangan di Provinsi Jawa Tengah Tahun 2018*. Liquid Crystals, 21(1).

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (1st ed.)*. STHDA.

Rahman, A., Wiranto, & Anggrainingsih, R. (2017). *Coal Trade Data Clusterung Using K-Means (Case Study PT. Global Bangkit Utama)*. ITSMART: Jurnal Ilmiah Teknologi Dan Informasi, 6(1), 24–31.

Setiawan, D., & Zahra, A. (2023). *Pengelompokan Kemiskinan di Indonesia Menggunakan Time Series Based Clustering*. Inferensi, 6(1), 83. https://doi.org/10.12962/j27213862.v6i1.14969

Talakua, M. W., Leleury, Z. A., & Taluta, A. W. (2017). *Analisis Cluster Dengan Menggunakan Metode K-Means Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014*. Barekeng: Jurnal Ilmu Matematika Dan Terapan, 11(2), 119–128. https://doi.org/10.30598/barekengvol11iss2pp119-128