International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparison of the K-Means and DBSCAN Methods in the Classification of Poverty Data of Districts/Cities in Kalimantan and Eastern Indonesia in 2022

Intan Dewi Difasari [a], *Iut Tri Utami* [b*], *Deby Fakhriyana* [c]

[a,b,c]*Statistics Department, Univeritas Diponegoro, Semarang, 50275, Indonesia*
[*]*Email adress: triutami.iut@gmail.com*

ABSTRACT

The poverty rate in eastern Indonesia is higher than in western Indonesia. The highest percentage of poor people is in Maluku and Papua Islands, reaching 19.89 percent. Poverty must be eradicated as soon as possible. If not handled, it can cause political and social issues. An essential factor to support poverty control is to map the characteristics of poverty in each regional group. Grouping is intended to make it easier for the government to develop poverty alleviation plans. This research groups poverty data for districts/cities in Kalimantan and eastern Indonesia in 2022 using the K-Means and DBSCAN methods. K-Means is an algorithm that divides data based on distance into groups and operates only on numeric attributes. K-Means is a simple grouping method that is easy to use. DBSCAN is a data clustering algorithm based on density and can find arbitrary group shapes. K-Means and DBSCAN group data into groups that do not overlap. The selection of optimal group results by comparing the highest Silhouette Index value between the K-Means and DBSCAN methods. Based on the research result, DBSCAN is the best grouping method which produces two groups with a Silhouette Index value of 0.585.

Keywords: Poverty; Grouping; K-Means; DBSCAN; Silhouette Index

1. INTRODUCTION

Poverty is a fundamental, complex, and multidimensional problem. Many countries are striving to overcome poverty [1]. Poverty needs to be addressed as soon as possible. Unaddressed poverty can lead to serious political and social unrest [2]. Based on data from the National Socioeconomic Survey [3], the poverty rate in eastern Indonesia is higher than the poverty rate in western Indonesia. The highest proportion of poor people is found in Maluku and Papua, reaching 19.89 percent. The next highest proportion of poor people is found on the islands of Bali and Nusa Tenggara at 13.35 percent, Sulawesi at 10.02 percent, and Kalimantan at 5.82 percent.

An important factor in supporting poverty control is mapping the characteristics of poverty in each regional group, particularly in districts/cities in Kalimantan and eastern Indonesia. Clustering is intended to facilitate government planning related to poverty alleviation. Cluster analysis can be applied to clustering.

Cluster analysis is a grouping that describes objects in the data and their relationships based on the facts presented in the data [4]. There are many methods of data clustering, including K-Means and DBSCAN (Density-Based Spatial Clustering Algorithm with Noise). K-Means and DBSCAN are non-hierarchical clustering methods that divide data sets into groups that do not overlap with one another. These methods are faster than hierarchical methods and are suitable for use with large data sets [5].

K-Means is an algorithm that divides data into clusters based on distance and operates only on numerical attributes [6]. K-Means is a simple and easy-to-use clustering method. DBSCAN is a clustering algorithm based on data density [7]. This method can find arbitrary cluster shapes that cannot be handled by K-Means.

The K-Means method has been compared with the DBSCAN method in previous studies. Jatipaningrum et al. [8] used the K-Means and DBSCAN methods to cluster districts and cities in East Java Province based on welfare levels. The clustering results showed that DBSCAN was the best method with a Davies-Bouldin Index value of 0.284 and produced 2 clusters with 5 noise points. Budiman [9] studied the comparison between the K-Means method and the DBSCAN method in clustering student boarding houses in Tembalang Village, Semarang. The K-Means method was the best with 2 clusters formed (k = 2) and a Silhouette Index value of 0.463. Based on Qadrini's research, the clustering of basic data on ITS laboratory competencies in 2017 using the K-Means and DBSCAN methods showed that the best clustering was achieved using the DBSCAN method with 4 groups formed, 0 noise, a Silhouette Index of 0.72, and a Dunn Index of 0.57 [9].

This study aims to find the optimal group between the K-Means method and the DBSCAN method in a case study of poverty data in regencies/cities in Kalimantan and eastern Indonesia in 2022. The selection of the optimal group was done by comparing the Silhouette Index values between the two methods. The group with the highest Silhouette Index value will be selected as the optimal group and will be profiled to determine the characteristics of the formed group.

2. LITERATURE REVIEW

Data mining is a procedure for collecting useful information from large databases. With data mining, it is possible to find new patterns that are useful and understandable from large databases. This new model will provide useful data analysis that can be studied by other decision support tools. One data mining technique is clustering. Clustering is a mechanism for grouping data in a certain way so that data in the same group have maximum similarity and data between groups have minimum similarity [4].

Data pre-processing is essential in data mining so that data sets can be processed quickly, and accurate conclusions can be reached. Normalization is part of data pre-processing that scales the data of a variable into a specified range. Min-max normalization is a normalization technique that builds a linear transformation on the original data into new data in the minimum and maximum value range [11]. Min-max normalization is expressed in equation (1).

$$x_i' = \frac{x_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \qquad (1)$$

where

x_i' = the normalized value, where i=1,2,...,*n*

x_i = the value of *the i-th* object in the original data

min_A = minimum value in the original data

max_A = maximum value in the original data

new_min_A = minimum value in the new data *range*

new_max_A = maximum value in *the* new data *range*

Cluster analysis is the grouping of data that describes objects and their relationships based on the facts presented in the data [4]. Cluster analysis aims to group similar objects into one cluster and distinguish them from other clusters [7].

There are two assumptions that must be met before grouping, namely representative samples and non-multicollinearity [12]. The representative sample test is conducted to prove that the selected sample can represent the existing population. The test is conducted using the Kaiser Meyer Olkin (KMO) test. A sample is said to be representative or able to represent the population if the KMO value is greater than 0.5 [13]. The second assumption is non-multicollinearity. Multicollinearity testing is conducted to identify linear relationships between independent variables. According to Gujarati [14], multicollinearity can be identified by using the Variance Inflation Factor (VIF) value. Data does not contain multicollinearity if the VIF value is less than 10, while a VIF value greater than 10 indicates multicollinearity and requires corrective measures such as Principal Component Analysis (PCA).

K-Means is an algorithm that divides data into groups based on distance and operates only on numerical attributes [6]. Data with similar characteristics will be grouped into one cluster, and data with different characteristics will be grouped into another cluster. K-Means is a simple and easy-to-use clustering method [7]. K-Means clustering is performed using the following algorithm [15]:

Determine the value of *k*

1. Randomly determine the initial cluster centers
2. Calculate the *Euclidean* distance of the object to each cluster center using equation (2)

$$D_{(x_i,C_k)} = \|x_i - C_k\|_2 = \sqrt{\sum_{t=1}^{z}\left(x_{i,t} - C_{k,t}\right)^2} \qquad (2)$$

where

x_i = object whose distance will be calculated

C_k = *the centroid* of group *k*, where k=1,2,...,*K*

$x_{i,t}$ = value of object *i* on variable *t*

$C_{k,t}$ = value of *the centroid of* group *k* on variable *t*

3. Group objects to the nearest cluster center. If the distance of an object to cluster center *k* is smaller than the distance to other cluster centers, then the object becomes a member of cluster *k*.
4. Determine the new cluster center using equation (3).

$$C_{k,t} = \frac{1}{M}\sum_{i=1}^{M} x_{i,t} \qquad (3)$$

where M is the number of objects in group *k*.

5. The algorithm continues to repeat until there are no changes in the members of each cluster.

DBSCAN is a density-based clustering algorithm [7]. Density in DBSCAN refers to the number of data points within a radius of Eps (the maximum distance between two data points in a group). Data is included in the specified density level if the total data within the Eps radius (including the data itself) is greater than or equal to MinPts (the minimum number of data within the Eps radius). The concept of density in data forms three types of states for each data point, namely core , and noise [7]. DBSCAN can find arbitrary cluster shapes that cannot be handled by K-Means [7].

The optimal Eps and MinPts values in DBSCAN are selected based on the k-dist graph by observing the range of Eps values from various k values. The k-dist graph is created by sorting the calculated distance values between the kth objects in descending order. The x-axis represents the order of objects and the y-axis represents the distance values of the objects. The Eps value is selected when there is a sharp change point on the k-dist with the k value as MinPts [17.

Clustering with the DBSCAN method according to [18] is carried out with the following algorithm:

1. Determine the Eps and MinPts parameters.
2. Determine the initial point *p* randomly.
3. Calculate the *Euclidean* distance of objects to *p* using the formula in equation (4).

$$D_{(x_i,p_i)} = \|x_i - p_i\|_2 = \sqrt{\sum_{t=1}^{z}\left(x_{i,t} - p_{i,t}\right)^2} \qquad (4)$$

with

x_i = object whose distance is to be calculated

p_i = the *i-th* object selected as point *p*

$p_{i,t}$ = value of the *i-th* object selected as point *p* in variable *t*

4. A cluster is formed with point *p* as its center if the number of objects within radius Eps exceeds the specified MinPts.
5. If point *p* is a boundary point and all points adjacent to point *p* have been visited, the process continues to another point.

The algorithm continues to repeat until all points have been processed.

The results of cluster analysis need to be validated to obtain the partition that best fits the data. This study uses Silhouette Index validation to select the best distance in K-Means and DBSCAN clustering. The Silhouette Index analyzes the position of each object in each cluster by comparing the average distance between objects within a cluster and with different clusters [19]. The Silhouette Index value ranges from -1 to 1. The quality of the formed clusters improves as the value approaches 1 [20].

The initial step in calculating the Silhouette Index according to Prasetyo [7] is to calculate a(i), which is the average distance of object i to all objects within a group in equation (5). The process continues by calculating b(i), which is the minimum value of the average distance of object i to all objects in other groups in equation (6). The final step is to calculate the Silhouette value of each object using equation (7) and calculate the Silhouette Index (SI) value, which is defined as the average ofs(i) using equation (8).

$$a(i) = \frac{1}{n_k - 1}\sum_{h \in Cl_k, ih \neq i} d(i,h) \qquad (5)$$

$$b(i) = min\left(\frac{1}{n_v}\sum_{g \in Cl_v, ig \neq i} d(i,g)\right) \qquad (6)$$

$$s(i) = \frac{b(i) - a(i)}{max\left(a(i), b(i)\right)} \qquad (7)$$

$$\text{SI} = \frac{1}{n}\sum_{i=1}^{n} s(i) \qquad (8)$$

with

$a(i)$ = the average distance of object *i* to all objects within a group, where i=1,2,...,n_{ik}

n_k = the number of objects in group *k*

h = other objects in group *k*, where h=1,2,..., n_{ik}

Cl_k = set of objects in group *k*

$d(i,h)$ = distance between object *i* and object *h*

$b(i)$ = the minimum value of the average distance between object *i* and all objects in other group *v*, where v=1,2,...,K

n_v = number of objects in group *v*

g = other objects in group *v*

Cl_v = the set of objects in group *v*, where *k≠v*

$d(i,g)$ = distance between object *i* and object *g*

$s(i)$ = *Silhouette* value of each data point

The interpretation of the *Silhouette Index* value according to Kaufman and Rousseeuw (1990) is given in Table 1.

Table 1. Silhouette Index Value Categories and Their Interpretation

Silhouette Index Value	Interpretation
$0.70 < \text{SI} \le 1.00$	Clustering with very strong bonds (*strong structure*)
$0.50 < \text{SI} \le 0.70$	Grouping with moderately good bonding (*medium structure*)
$0.25 < \text{SI} \le 0.50$	Grouping with weak bonds (*weak structure*)
$\text{SI} \le 0.25$	No bonds in the formed group

Poverty is a condition in which an individual or household faces difficulties in meeting basic needs, while the surrounding community does not provide the means to improve their welfare [21]. The Central Statistics Agency [13] applies an approach based on the ability to meet basic needs in estimating poverty. Poverty through this approach is seen as financial powerlessness to meet basic food and non-food needs, which are estimated based on expenditure.

3. RESEARCH METHOD

The type of data used in this study is secondary data from Statistics Indonesia publications based on data from the 2022 National Socioeconomic Survey. The data consists of 241 districts/cities in Kalimantan and eastern Indonesia, where there are 18 provinces covering Kalimantan, Bali, the Nusa Tenggara Islands, Sulawesi, the Maluku Islands, and Papua.

Based on Indonesia's 2022 poverty profile data, the research variables used are the percentage of poor people (X1 , in percent), the poverty depth index (X2), and the poverty line (X3 , in rupiah/capita/month). The percentage of poor people is the percentage of people living below the poverty line. People with an average per capita expenditure per month below the poverty line are classified as poor. The poverty line is the minimum amount of expenditure on food and other items that must be met in order not to be categorized as poor. Meanwhile, the poverty depth index is the average level of expenditure gap of each poor person relative to the poverty line [3].

The K-Means and DBSCAN methods were applied to data analysis using RStudio software version 4.2.0. The following are the stages of data analysis that were carried out:

1. Collecting data and preprocessing data by performing min-max normalization.
2. Testing whether the sample represents the population using the normalized data with the KMO test.
3. Performing a non-multicollinearity test using the normalized data by checking the VIF value. If multicollinearity is found, PCA is performed.
4. Performing clustering using the K-Means method by first determining the optimal k value based on the Silhouette curve.
5. Performing clustering using the DBSCAN method by first determining the optimal values of Eps and MinPts based on the k-dist graph.
6. Calculate the Silhouette Index value from the K-Means and DBSCAN clustering results.
7. Select the optimal cluster based on the highest Silhouette Index value.
8. Performing profiling of the optimal clusters formed.

4. RESULT AND DISCUSSION

Before processing the data, descriptive analysis was conducted to obtain a general description of the research data. The following is descriptive statistics of poverty data for districts/cities in Kalimantan and eastern Indonesia in 2022.

Table 1 . Descriptive Statistics of Poverty Data for Districts/Cities in Kalimantan and Eastern Indonesia in 2022

Variable	N	Minimum	Maximum	Mean	Standard Deviation
X_1	241	2.45	42.03	13.6240	9.22105
X_2	241	0.17	13.90	2.5150	2.44063
X_3	241	264,666	1099019	488,279.80	130,970.157

The preprocessing technique performed on the data is normalization using min-max normalization. The normalization results are presented in Table 3.

The KMO test is used to test whether the sample used can represent the population. The hypothesis used in the KMO test, with H0 is that the sample represents the population and H1 is that the sample does not represent the population. Based on the KMO test results, KMO = 0.52 was obtained, which is greater than 0.50. This value indicates that H0fails to be rejected, meaning that the data can represent the existing population.

Multicollinearity is identified by examining the VIF value. Data is said to have no multicollinearity between variables if the VIF value is less than 10. Based on Table 4, the VIF value for all variables is less than 10. Therefore, it can be concluded that there is no multicollinearity in the data and the grouping process can be carried out.

lustering using the K-Means method begins with determining the optimal number of k based on the Silhouette curve. The highest Silhouette Index value based on Figure 1 is when k = 2, so it can be concluded that the optimal number of K-Means clusters is 2. The clustering results using the K-Means method form 2 clusters with 191 districts/cities in cluster 1 and 50 districts/cities in cluster 2.

Table 3. Min-Max Normalization Results

District/City	X_1	X_2	X_3
Jembrana	0.072	0.041	0.232
Tabanan	0.069	0.036	0.309
Badung	0.002	0.009	0.442
Gianyar	0.057	0.038	0.203
Klungkung	0.091	0.026	0.114
⋮	⋮	⋮	⋮
Puncak	0.854	0.585	0.587
Dogiyai	0.683	0.430	0.393
Intan Jaya	1,000	0.371	0.606
Deiyai	0.957	0.497	0.459
Jayapura City	0.219	0.170	1.000

2 . VIF Values

Variable	VIF
X_1	5.949736
X_2	6.083331
X_3	1.063484

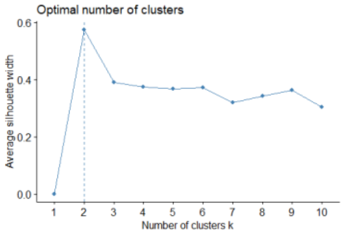


Figure 1. Silhouette Curve

Next, clustering will be performed using the DBSCAN method. Before performing clustering, the optimal Eps and MinPts parameters will be determined using the k-dist graph. The computation is performed using k = 2, 3, 4, and 5, because k values that are too large will create small clusters and mislabel noise.

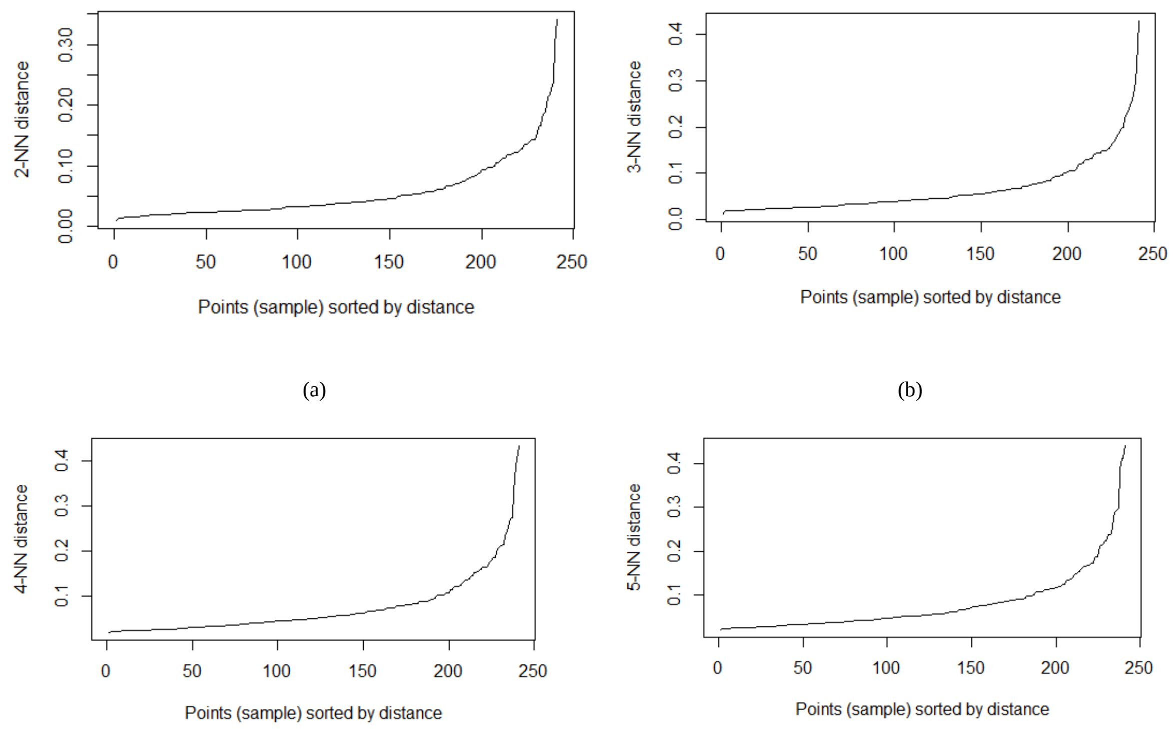
Based on trial and error to determine the optimal DBSCAN parameters based on the plots on the k-dist graph, the optimal parameters for forming groups are Eps = 0.24 and MinPts = 2. The clustering results using the DBSCAN method formed 2 groups with 237 districts/cities in group 1 and 4 districts in group 2.

The validation calculation of the clustering results of the K-Means method and the DBSCAN method with the Silhouette Index is presented in Figure 3.

Based on the highest Silhouette Index values obtained from grouping poverty data for districts/cities in Kalimantan and eastern Indonesia in 2022 using the K-Means and DBSCAN methods, DBSCAN was found to be the best grouping method. The Silhouette Index value for the DBSCAN method was 0.585, indicating that the resulting clusters were of fairly good quality (medium structure).

After selecting the best clustering method, the profiling process was carried out to examine the characteristics of the formed clusters. The average of each research variable in cluster k can represent the characteristics of that cluster. The following is the profiling of the formed clusters based on the average research variables.

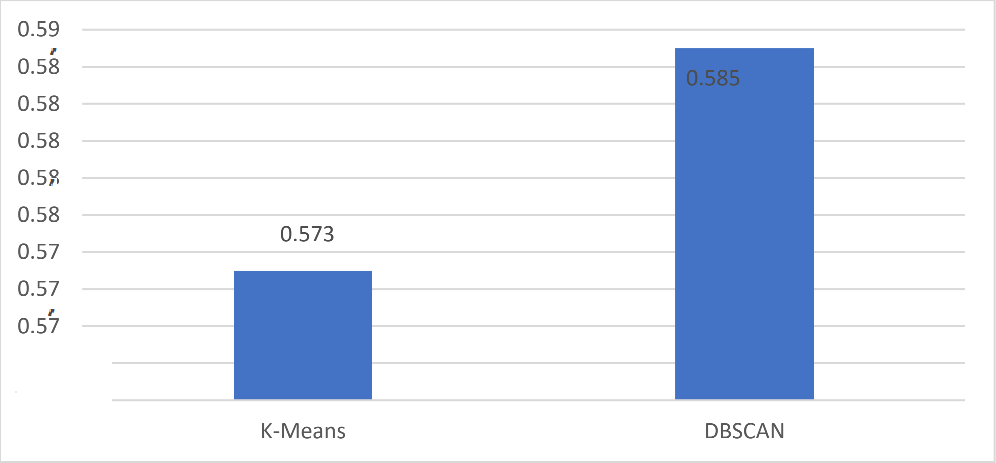
The highest average variable based on Table 5 is in group 2. This means that districts/cities in group 2 have a higher percentage of poor people, poverty depth index, and poverty line compared to districts/cities in group 1. Thus, it can be said that members of group 1 are districts/cities with low poverty levels, while members of group 2 are districts/cities with high poverty levels. The districts included in group 2 are Jayawijaya District, Puncak Jaya District, Supiori District, and Lanny Jaya District, which are located in Papua Province. Based on this, it can be concluded that members of group 2 are districts that must be given more attention by the Indonesian government so that these districts can develop further and reduce poverty levels in the coming years.



(a) (b)

(c) (d)

1 Figure. k-dist graph: (a) k = 2, (b) k = 3, (c) k = 4, (d) k = 5



3. Silhouette Index graph

Table 5. Average Variables of Poverty Data in Regencies/Cities in Kalimantan and Eastern Indonesia in 2022

Variable	Group	
	1	2
Percentage of Poor Population	13,221	**36,890**
Poverty Depth Index	2,340	**12,893**
Poverty Line	486,562.637	**590,021.75**

5. CONCLUSION

The grouping of poverty data for districts/cities in Kalimantan and eastern Indonesia using the K-Means and DBSCAN methods resulted in DBSCAN being the best grouping method with a Silhouette Index value of 0.585. This value indicates that the grouping results are of fairly good quality (medium structure). The optimal parameters for DBSCAN are Eps = 0.24 and MinPts = 2, with 2 clusters formed and 0 noise. There were 237 regencies/cities in group 1 and 4 in group 2. Meanwhile, the K-Means method produced a lower Silhouette Index value than the DBSCAN method, namely 0.573. The K-Means method produced 2 clusters (k = 2) with 191 regencies/cities in group 1 and 50 in group 2.

The results of optimal group profiling using the DBSCAN method produced 2 groups, where group 1 consisted of districts/cities with low poverty levels and group 2 consisted of districts/cities with high poverty levels. Members of group 2, Jayawijaya Regency, Puncak Jaya Regency, Supiori Regency, and Lanny Jaya Regency, located in Papua Province, are regencies that require more attention from the Indonesian government so that they can develop further and reduce poverty levels in the coming years.

The DBSCAN method is better than the K-Means method in this study, but it does not guarantee that the DBSCAN method is always better than the K-Means method. Therefore, in future studies, a simulation study can be conducted to compare the DBSCAN and K-Means methods. Poverty indicators used in subsequent studies can use other indicators such as education and employment. In addition, a Graphical User Interface (GUI) needs to be created to facilitate optimal clustering analysis using the K-Means and DBSCAN algorithms.

REFERENCE:

[1] V. Latumahina, “Peran Gereja dalam Menanggapi Kemiskinan,” *Jurnal Teologi Biblika*, vol. 6, no. 1, pp. 29–36, 2021.

[2] F. Rizal and H. Mukaromah, “Kebijakan Pemerintah Indonesia dalam Mengatasi Masalah Pengangguran Akibat Pandemi Covid-19,” *Al-Manhaj: Jurnal Hukum dan Pranata Sosial Islam*, vol. 3, no. 1, pp. 35–66, 2021.

[3] Badan Pusat Statistik, *Profil Kemiskinan di Indonesia Maret 2022*. 2022. [Online]. Available: https://www.bps.go.id/pressrelease/2022/07/15/1930/persentase-penduduk-miskin-maret-2022-turun-menjadi-9-54-persen.html. [Accessed: Apr. 14, 2023].

[4] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. New York: Pearson Addison Wesley, 2006.

[5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Prentice Hall International, 2007.

[6] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[7] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi, 2012.

[8] M. T. Jatipaningrum, B. A. Prasetyo, N. B. Sari, and R. W. Sari, "Pengelompokan Kabupaten Dan Kota Di Provinsi Jawa Timur Berdasarkan Tingkat Kesejahteraan Dengan Metode K-Means Dan Density-Based Spatial Clustering Of Applications With Noise (DBSCAN)," Jurnal Derivat, vol. 9, no. 1, pp. 1-13, Jul. 2022.

[9] S. A. D. Budiman, D. Safitri, and D. Ispriyanti, "Perbandingan Metode K-Means dan Metode DBSCAN pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang," JURNAL GAUSSIAN, vol. 5, no. 4, pp. 757-762, 2016.

[10] L. Qadrini, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, vol. 13, no. 2, pp. 5-11, 2020.

[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann, 2012.

[12] S. Nugroho, *Statistika Multivariat Terapan*. Bengkulu: UNIB Press, 2008.

[13] S. Yamin and H. Kurniawan, *SPSS for Windows untuk Analisis Data Statistik dan Penelitian*. Jakarta: Salemba Infotek, 2014.

[14] D. Gujarati, *Dasar-dasar Ekonometrika*, vol. 2. Jakarta: Erlangga, 2009.

[15] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, 2014.

[16] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. New York: John Wiley & Sons, 1990.

[17] U. Y. Purwanto, “Penggerombolan Spasial Hotspot Kebakaran Hutan dan Lahan Menggunakan DBSCAN dan ST-DBSCAN,” Tesis, Sekolah Pascasarjana Institut Pertanian Bogor, 2012.

[18] N. M. A. S. Devi, I. K. G. D. Putra, and I. M. Sukarsa, “Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan,” *Lontar Komputer*, vol. 6, no. 3, pp. 185–191, 2015.

[19] F. Aini, S. Palgunadi, and R. Anggrainingsih, “Clustering Business Process Model Petri Net dengan Complete Linkage,” *Jurnal ITSMART*, vol. 3, no. 2, pp. 47–51, 2014.

[20] M. A. Nahdliyah, W. Tatik, and P. Alan, “Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C Index,” *Jurnal Gaussian*, vol. 8, no. 2, pp. 161–170, 2019.

[21] A. Cahyat, *Mengkaji Kemiskinan dan Kesejahteraan Rumah Tangga*. Bogor: Center for International Forestry Research, 2007.