Volume 9 No 1 June (2026)

e-ISSN: 2615-3270
p-ISSN: 2615-3599

Eigen Mathematics Journal

homepage: https://eigen.unram.ac.id/index.php/eigen

Comparison of Cluster Average Linkage and K-Means Analysis Methods for Poverty Grouping in The Nusa Tenggara Area

Muhammad Alimuddin[1], Lisa Harsyiah[2*], Zulhan Widya Baskara[2]

[1] *Department of Mathematics, University of Mataram, Indonesia*
[2] *Department of Statistics, University of Mataram, Indonesia*
[*] *Corresponding author: lisa_harsyiah@unram.ac.id*

A B S T R A C T

Poverty is a problem that often occurs and is a fundamental problem in almost all developing countries, including Indonesia. The Nusa Tenggara region consists of two administrative regions, namely the Provinces of West Nusa Tenggara (NTB) and East Nusa Tenggara (NTT) which have high poverty rates. The increase in the number of poor people was caused by several indicators such as environmental conditions, education, income, health, access to goods and services, and others. The purpose of this research is to determine the best method in the process of classifying poverty with the cluster analysis method. The methods used in this study are the average linkage and K-Means cluster analysis methods, as well as the silhouette index method in terms of cluster validation to obtain the best cluster analysis method. The data used is poverty data for the Nusa Tenggara Region in 2021 which includes four poverty sectors, namely employment, education, health, and housing and the environment. Based on the research results, the best method for grouping is the K-Means cluster analysis method by forming three clusters where the first cluster consists of 3 districts/cities, the second cluster consists of 22 districts/cities, and the third cluster consists of 7 districts/cities. The K-Means cluster analysis method is the best method with the highest silhouette index value of 0.28, higher than the average linkage method which obtained a silhouette index value of 0.27.

Keywords: Average Linkage, Cluster Analysis, K-Means, Poverty

Received : 09-08-2024;
Revised : 20-01-2026;
Accepted : 03-02-2026;
Published : 02-03-2026;

DOI: https://doi.org/10.29303/emj.v9i1.243

1. Introduction

Poverty is one of the fundamental problems in almost all countries, particularly in developing countries, including Indonesia. According to the Statistics Indonesia, poverty is an economic inability to meet basic food and non-food needs (measured in terms of expenditure), where a person is said to be poor if the average per capita expenditure per month is below the poverty line.

The Nusa Tenggara region consists of two administrative regions, namely West Nusa Tenggara Province which consists of 10 districts/cities and East Nusa Tenggara Province which consists of 22 districts/cities. Based on data from Statistics Indonesia, it is known that the number of poor people in West Nusa Tenggara Province in 2021 will reach 746.66 thousand people or 14.14% of the total population in West Nusa Tenggara Province. Meanwhile, for East Nusa Tenggara Province itself, in 2021 the number of poor people will reach 1,169,310 people or 20.99% of the total population in East Nusa Tenggara Province.

The increase in the number of poor people is caused by several interconnected poverty problems such as geographic location, environmental conditions, education, income, health, access to goods and services, and so on. Apart from that, there are also several cases regarding the distribution of social assistance that is not on target, both in terms of target recipients and distribution areas. The problem of poverty is what causes the slow pace of the community's economy which leads to an increase in the percentage of poor people in the Nusa Tenggara Region [1].

Therefore, we need a method that cn group each district/city into groups with poverty problems that are similar to other group members. By grouping districts/cities, it is hoped that it can help the government or related agencies in dealing with poverty problems appropriately and quickly in accordance with the poverty problems in the districts/cities.

Cluster analysis is used to analyze data by grouping objects or individuals into several groups, such that those within the same group share relatively homogeneous characteristics [2]. In cluster analysis, it is divided into two methods, namely hierarchical and non-hierarchical methods. Hierarchical clustering methods include single linkage, complete linkage, average linkage, centroid, Ward, and median clustering. Meanwhile, non-hierarchical cluster analysis is generally referred to as the K-Means method [3].

The hierarchical method used in this research is the Average Linkage method because it has several advantages, namely it does not require determining how many clusters there are first, it can provide a graphical representation in the form of a dendrogram and can detect various shapes and sizes of clusters. Meanwhile, the non-hierarchical method or K-Means also has several advantages, namely low complexity, fast calculations, can handle large data, and adjustable cluster members [3].

Apart from the advantages possessed by each method, the use of this method is also based on previous research related to this research, namely that conducted by [3] where in this research they compared which one is better to use between K-Means and Average Linkage for poverty clustering in Central Java Province. The results of their research show that the Average Linkage method better with a Silhouette Coefficient value of 0.35, where this value is higher when compared to the K-Means method with a Silhouette Coefficient value of 0.2. However, a study conducted by [4] on the clustering of poor population data in Cianjur Regency using the K-Means and Hierarchical Clustering methods found that K-Means has advantages in terms of speed and efficiency, particularly for large datasets with uniform data distributions

Based on the explanation above, this research was conducted to compare the best method between the two cluster analysis methods , namely the Average Linkage method and the K-Means method based on poverty indicators in the Nusa Tenggara Region. Then, to obtain the best results from the district/city grouping, a cluster validation process is carried out using one of the methods, namely Silhouette Index, where the best results can be seen from the silhouette coefficient value which is higher when compared with other values.

2. Research Methods

The data used is poverty data for 2021 which can be obtained through book publications by Statistics Indonesia of West and East Nusa Tenggara Province. This research aims to compare the best method between the two cluster analysis methods, namely the Average Linkage method and the K-Means method based on poverty indicators in the Nusa Tenggara Region. These research variables will, among others, be presented in the following Table 1.

Table 1. Research variables

Variable	Information
x_1	Percentage of Poor Population
x_2	Percentage of Open Unemployment Rate
x_3	Percentage of Labor Force Participation Rate
x_4	Percentage of Illiterate Rates of population aged 15 years and over by district/city
x_5	Percentage of School Participation Rates of population aged 16–18 years by district/city
x_6	Percentage of Population Who Had Health Complaints During the Last Month by Regency/City
x_7	Percentage of Population Who Have Health Complaints and Seek Outpatient Treatment During the Last Month by Regency/City
x_8	Percentage of Population Who Have BPJS Health Insurance Recipients of Contribution Assistance According to Regency/City
x_9	Percentage of Population Who Have BPJS Health Insurance Non-Contribution Assistance Recipients According to Regency/City
x_{10}	Percentage of Households based on Use of Own Defecation Facilities by Regency/City
x_{11}	Percentage of Households based on Owned Residential Building Control Status by Regency/City
x_{12}	Percentage of Households that Have Access to Adequate Sanitation by Regency/City
x_{13}	Percentage of Households that Have Access to Adequate Drinking Water Sources by Regency/City

The steps of this research are divided into several stages as follows.

1. Collect and input data related to poverty problems that occur in districts/cities in the Nusa Tenggara region, then determine the research variables.
2. Carrying out several tests, namely the first is the sample adequacy test, namely by carrying out the Keiser Mayer Olkin (KMO) test which has a range value of 0.5-1.0. The KMO value can be found using Equation (1) [5].

$$\text{KMO} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} \tag{1}$$

where r_{ij}^2 is correlation coefficient between variables i and j, term a_{ij}^2 is partial correlation coefficient between variables i and j, and $i = 1, 2, 3, ..., n$ and $j = 1, 2, 3, ..., n$ with $i \neq j$. Next, carry out a multicollinearity test with the aim of finding out whether there is a correlation between the independent variables. The way to detect multicollinearity is to use the Variance Inflation Factor (VIF) price. N is the VIF value > 10 shows that there is multicollinearity between the independent variables, where the VIF value can be found using the Equation (2) [6]:

$$\text{VIF}(x_j) = \frac{1}{1 - R_j^2}, \qquad j = 1, 2, 3, \ldots, n \tag{2}$$

with R_j^2 is the coefficient of determination.

3. Determine the optimal number of clusters using the silhouette method with the aim of obtaining optimal values before carrying out cluster analysis.
4. Cluster analysis using two methods, namely the Average Linkage method using Equation (3) [7]

$$d_{(AB)C} = \frac{\sum_i \sum_k d_{ik}}{N_{(AB)}\, N_C} \tag{3}$$

where $d_{(AB)C}$ is distance between cluster (AB) and cluster C, term d_ik means distance between objects i in the cluster (AB) and objects k in cluster C, the number of objects in the cluster (AB) is represented by N_{AB}, and N_C is number of objects in cluster C. The K-Means method is using Equation (4) [8]

$$C_{kj} = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n}. \tag{4}$$

In Equation (4), C_{kj} is the value of the centroid of the i-th variable j, and n is the amount of data. The distance between clusters is calculated first before carrying out the clustering process using the Euclidean distance which is formulated using Equation (5) [9]:

$$d_{(l,m)} = \sqrt{\sum_{k=1}^{n} (x_{lk} - x_{mk})^2} \tag{5}$$

where $d_{(l,m)}$ is the distance between object l and object m in terms of value poverty variable, the x_{lk} means the object value l in variable k, object value m in the variable k is represented by x_{mk}, and n is the number of variables observed.

5. Carrying out cluster validation which aims to determine which method is better between the two methods used by using the Silhouette Index method with Equation (6) [10]

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{6}$$

with s_i is Silhouette Coefficient value, a_i is the average distance of the object i to other objects is in one cluster, and b_i is the minimum value of the average data distance i with the amount of data in other clusters

6. Form a cluster plot from the best clustering results with the aim of providing an overview of the clustering results formed.

3. Result and Discussion

3.1. Sample Adequacy Test

Before carrying out cluster analysis, you must first know whether the data or sample that will be used is sufficient to represent the population or not, namely by carrying out the KMO test. Before that, first look for simple correlation values and partial correlations between the research variables used. Using Equation (1), the KMO value is obtained as follows.

$$\begin{aligned}\text{KMO} &= \frac{(-0.45)^2 + 0.42^2 + \cdots + 0.59^2}{(-0.45)^2 + \cdots + 0.59^2 + \left(0.14^2 + \cdots + 0.46^2\right)} \\ &= \frac{10.837}{10.837 + 4.764} \\ &= 0.695.\end{aligned}$$

Based on the calculation results for the KMO test results, it is 0.695, which means the KMO value > 0.5. This indicates that the sample data used can represent the existing population and is worthy of further analysis.

3.2. Multicollinearity Test

Obtained is first calculated by looking for a multiple linear regression model. The R_j^2 coefficient of determination value obtained was 0.778622, which was then used to find the VIF value with the following Equation (2).

$$\begin{aligned}VIF(x_j) &= \frac{1}{1 - R_j^2} \\ VIF(x_1) &= \frac{1}{1 - R_1^2} \\ &= \frac{1}{1 - 0.778622} \\ &= 4.517152.\end{aligned}$$

Calculations in the same way can be done for all variables until the VIF value of each research variable is found. Multicollinearity tests can also be carried out using statistical software. The VIF value for each research variable that has been searched for can be seen in the Table 2.

Table 2. Multicollinearity test results

Variable	VIF	Variable	VIF
x_1	4.517152	x_8	2.30254
x_2	4.166641	x_9	7.304781
x_3	2.412425	x_{10}	2.791276
x_4	3.091029	x_{11}	8.075657
x_5	1.559189	x_{12}	3.920284
x_6	1.82088	x_{13}	3.525421
x_7	1.567494		

Based on Table 2, it can be seen that all research variables have a VIF value of ≤ 10, so it can be said that there is no correlation between the independent variables.

3.3. Optimal Clusters k

Determining clusters optimal k using the silhouette method with the reason that this method is the easiest method, namely by just looking at the k value in which cluster is the highest. The goal is to measure how well an object or individual is positioned in a cluster to see the quality and strength of the cluster the . The results obtained can be presented in graphical form in the Figure 1.

It is known in Figure 1 that during the cluster k value 3 is the highest value, so it is a cluster k The optimum is 3 clusters, so when carrying out K-Means and average linkage cluster analysis, grouping is carried out with 3 clusters for each method.

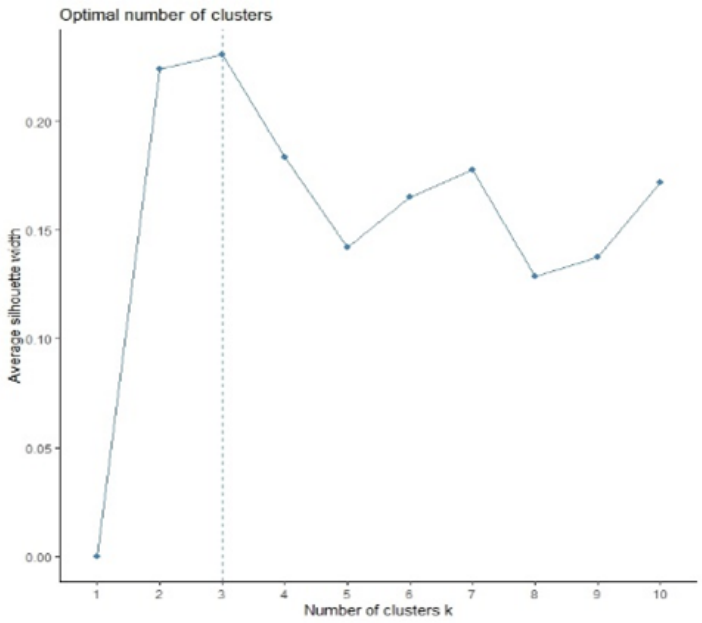

Figure 1. Optimal results of k the silhouette method

3.4. Average Linkage

Cluster formation in this method is done by combining two data that have the closest distance to other data. Calculation of the distance between data is carried out using the Euclidean distance calculation as in Equation (5) with the value $(i, j) = 1, 2, 3, \ldots, 32$.

$$\begin{aligned} d_{(1,2)} &= \sqrt{(14.47 - 13.44)^2 + \cdots + (91.76 - 92.52)^2} \\ &= 24.231. \end{aligned}$$

Calculation of the distance between data using Euclidean distance calculations is carried out on each data to obtain the minimum distance to combine the corresponding data. The minimum distance is 14.749 which is the distance between the 4^{th} data and the 28^{th} data.

The next step is to calculate the distance between clusters $(4, 28)$ and other data using Equation (3), so the results are obtained in Equation (7).

$$d_{(4,28)1} = \frac{32.186 + 31.450}{2 \times 1} = 31.818. \tag{7}$$

Distance calculations between clusters are carried out on each data to obtain the desired results, namely forming 3 clusters. The process of forming these 3 clusters was carried out until the 30^{th} grouping with the following results.

Table 3. Results of forming 3 clusters

No	Regency/City	Cluster	No	Regency/City	Cluster
1	West Lombok	1	17	Alor	1
2	Central Lombok	1	18	Lembata	1
3	East Lombok	1	19	East Flores	1

No	Regency/City	Cluster	No	Regency/City	Cluster
4	Sumbawa	1	20	Sikka	1
5	Dompu	1	21	Ende	1
6	Bima	1	22	Ngada	1
7	West Sumbawa	1	23	Manggarai	1
8	North Lombok	2	24	Rote Ndao	1
9	Mataram City	3	25	West Manggarai	1
10	Bima City	3	26	Central Sumba	2
11	West Sumba	2	27	Southwest Sumba	2
12	East Sumba	1	28	Nagekeo	1
13	Kupang	1	29	East Manggarai	2
14	South Central Timor	2	30	Sabu Raijua	2
15	North Central Timor	1	31	Malaka	1
16	Sikka	1	32	Kupang City	3

Based on Table 3, it is known that the cluster with the most members is the first cluster with 22 districts/cities, then the second cluster with 7 districts/cities, while the third cluster has the least members with 3 districts/cities.

3.5. K-Means

The first step of the K-Means method is to determine the initial cluster center (centroid). The centroid is determined randomly, because it will form 3 clusters, the centroids are C1, C2, and C3. The centroid that has been selected is then calculated using the Euclidean distance formula which can be seen in Equation (5). Calculation of the distance from the data nto the centroid C1, namely

$$\begin{aligned} d_{(1,1)} &= \sqrt{(14.47 - 8.88)^2 + \cdots + (91.76 - 99.93)^2} \\ &= 39,98601005. \end{aligned}$$

The same thing is done for each data at each centroid that has been determined. The next step is I carry out calculations for the second iteration or repetition using the new centroid . Centroid The new one can be obtained by calculating the cluster average as in Equation (4). The calculation of the new centroid C1 is

$$x_1 = \frac{14.88 + 8.65 + 8.88 + 13.35 + 9.17}{5} = 10.986.$$

The same thing is done for each data at each centroid to obtain a new centroid. Repetition or iteration is carried out using the same formula, namely the Euclidean distance, until in the end there is no change in the clustering process. The data did not change after the 3rd iteration with the results in Table 4.

Table 4. Results of the 3 cluster K-Means method

No.	Regency/City	Cluster	No.	Regency/City	Cluster
1	West Lombok	2	17	Alor	2
2	Central Lombok	2	18	Lembata	2
3	East Lombok	2	19	East Flores	2
4	Sumbawa	2	20	Sikka	2

No.	Regency/City	Cluster	No.	Regency/City	Cluster
5	Dompu	2	21	Ende	2
6	Bima	2	22	Ndao	2
7	West Sumbawa	2	23	Manggarai	2
8	North Lombok	2	24	Rote Ndao	2
9	Mataram City	1	25	West Manggarai	2
10	Bima City	1	26	Central Sumba	3
11	West Sumba	3	27	Southwest Sumba	3
12	East Sumba	3	28	Nagekeo	2
13	Kupang	2	29	East Manggarai	3
14	South Central Timor	3	30	Sabu Raijua	3
15	North Central Timor	2	31	Malacca	2
16	Sikka	2	32	Kupang City	1

As seen in Table 4, it is known that the second cluster is the cluster with the most members, namely 22 districts/cities, then the third cluster with 7 districts/cities, and finally the first cluster with 3 districts/cities.

3.6. Cluster Validation

Cluster validation method used is the silhouette method index. The silhouette index method is an evaluation method that aims to test the accuracy of a cluster that has been formed from the clustering process.

1. Average Linkage Method
   Calculating the silhouette coefficient value based on Equation (6), the results are obtained in Equation (8).

$$s_1 = \frac{46.52386 - 29.93233}{\max(29.93233; 46.52386)} = 0.356624. \tag{8}$$

   After knowing the value s_i for each district/city, the average of all values is calculated s_i to obtain the silhouette index 3 cluster average linkage value . The results obtained is Equation (9).

$$\begin{aligned} \bar{s}_i &= \frac{0.356624 + 0.30078 + \cdots + 0.341542}{32} \\ &= 0,266733 \\ &\approx 0,27. \end{aligned} \tag{9}$$

   So from calculating the value of the silhouette index for 3 cluster average linkage, the result is 0.266733 or close to 0.27.
2. K-Means Method
   Calculating the silhouette coefficient value based on Equation (6), the following are the calculation results:

$$s_1 = \frac{42.6371756 - 28.5215024}{\max(28.5215024; 42.6371756)} = 0.331065. \tag{10}$$

   After knowing the value s_i of each district/city, the average of all values is calculated s_i to obtain the silhouette index value of 3 cluster K-Means. The results obtained are as follows:

$$\begin{aligned} \bar{s}_i &= \frac{0.331065 + 0.177639 + \cdots + 0.137828}{32} \\ &= 0.275652 \\ &\approx 0.28. \end{aligned} \tag{11}$$

So from calculating the silhouette index value for 3 cluster K-Means, the result is 0.275652 or 0.28.

3.7. Cluster Plots

The cluster validation that has been carried out, the best results were obtained, namely the 3 cluster K-Means method with silhouette values index of 0.28. Clustering results These can be presented in cluster form plot which aims to provide an easy overview of the clustering results that are formed. The cluster plot can be seen in Figure 2.

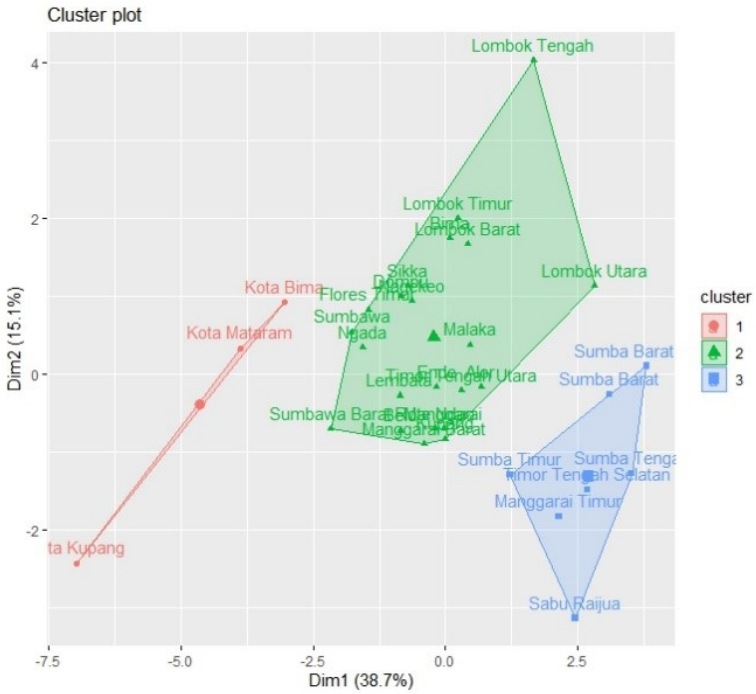

Figure 2. Cluster plot K-Means 3 clusters

Based on Figure 2, it can be seen that the clustering results obtained are consistent with the calculations previously performed, where the first cluster, characterized by a low poverty level, consists of 3 districts/cities; the second cluster, with a moderate poverty level, consists of 22 districts/cities; and the third cluster, characterized by a high poverty level, consists of 7 districts/cities.

The limitation of this study is that it only compares two clustering analysis methods, namely Average Linkage and K-Means; therefore, the clustering results cannot be compared with other clustering methods. In addition, cluster validation is conducted using only the Silhouette Index, so the evaluation of cluster quality does not yet cover a variety of validation measures. Based on these limitations, future research is recommended to apply and compare other clustering methods, such as Fuzzy C-Means or machine learning–based models, to improve the accuracy and robustness of the clustering results

4. Conclusions

Based on the results and discussions previously explained, the conclusion was obtained that the best results obtained from comparing the grouping of districts/cities in the Nusa Tenggara Region based on poverty indicators were the 3 cluster K-Means method with a silhouette index value of 0.28.

As for each cluster, the first cluster consists of 3 districts/cities, namely Mataram City, Bima City and Kupang City. The second cluster consists of 22 districts/cities, namely West Lombok, Central Lombok, East Lombok, Sumbawa, Dompu, Bima, West Sumbawa, North Lombok, Kupang, North Central Timor, Belu, Alor, Lembata, East Flores, Sikka, Ende, Ngada, Manggarai, Rote Ndao, West Manggarai, Nagekeo, and Malacca. The third cluster consists of 7 districts/cities, namely West Sumba, East Sumba, South Central Timor, Central Sumba, Southwest Sumba, East Manggarai and Sabu Raijua.

REFERENCES

[1] Badan Pusat Statistik, *Statistik Indonesia 2020*, vol. 53. Springer, 2022. https://www.bps.go.id/id/publication/2020/04/29/e9011b3155d45d70823c141f/statistik-indonesia-2020.html.

[2] M. W. Talakua, Z. A. Leleury, and A. W. Taluta, "Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan kabupaten/kota di provinsi maluku berdasarkan indikator indeks pembangunan manusia tahun 2014," *BAREKENG : Journal of Mathematics and Its Applications*, vol. 11, no. 2, pp. 119–128, 2017. https://doi.org/10.30598/barekengvol11iss2pp119-128.

[3] D. Widyadhan, R. B. Hastuti, I. Kharisudin, and F. Fauzi, "Perbandingan analisis klaster k-means dan average linkage untuk pengklasteran kemiskinan di provinsi jawa tengah," *Prosiding Seminar Nasional Matematika Jurusan Matematika Fakultas MIPA Universitas Negeri Semarang*, vol. 4, pp. 584–594, 2021. https://journal.unnes.ac.id/sju/prisma/article/view/45032.

[4] I. P. Putra and A. Fadhillah, "Perbandingan metode k-means dan hierarchical clustering dalam pengelompokan data penduduk miskin di kabupaten cianjur," *Jurnal Inovasi dan Tren*, vol. 3, no. 1, pp. 227–234, 2025. https://doi.org/10.35870/ljit.v3i1.4028.

[5] W. Alwi and M. Hasrul, "Analisis klaster untuk pengelompokkan kabupaten/kota di provinsi sulawesi selatan berdasarkan indikator kesejahteraan rakyat," *Jurnal Matematika dan Statistika serta Aplikasinya*, vol. 6, no. 1, pp. 35–42, 2018. https://doi.org/10.24252/msa.v6i1.4782.

[6] Z. R. Putri, "Perbandingan analisis cluster hierarki aglomeratif dengan menggunakan metode single linkage, complete linkage dan average linkage," *Jurusan Statistika, Fakultas Matematika Dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta*, 2017. https://dspace.uii.ac.id/handle/123456789/31493.

[7] S. Pratiwi, T. Widiharih, and A. Hakim, "Analisis klaster metode ward dan average linkage dengan validasi dunn index dan koefisien korelasi cophenetic (studi kasus: Kecelakaan lalu lintas berdasarkan jenis kendaraan tiap kabupaten/kota di jawa tengah tahun 2018)," *Jurnal Gaussian*, vol. 8, no. 4, pp. 486–495, 2019. https://doi.org/10.14710/j.gauss.8.4.486-495.

[8] E. Prasetyo, *Data mining : konsep dan aplikasi menggunakan MATLAB*, vol. 1. CV Andi Offset, 2012. https://elibrary.bsi.ac.id/readbook/200350/data-mining-konsep-dan-aplikasi-menggunakan-matlab.

[9] B. Ruswandi, *Analisis Klaster K-Means dan K-Median Pada Data*. https://repository.uinjkt.ac.id/dspace/bitstream/123456789/21826/1/FEBRIYANA-FST.pdf.

[10] R. Handoyo, R. Mangkudjaja, and S. M. Nasution, "Perbandingan metode clustering menggunakan metode single linkage dan k - means pada pengelompokan dokumen," *Jurnal SIFO Mikroskil*, vol. 15, pp. 73–82, 10 2014. https://doi.org/10.55601/jsm.v15i2.161.