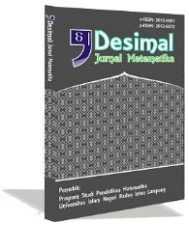
Desimal: Jurnal Matematika Vol 7 No 3 (2024) 489-500

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



Comparison of agglomerative hierarchical clustering (AHC) algorithm and k-means algorithm in poverty data clustering in north sumatra

Wilia Usna*, Rima Aprilia

Universitas Islam Negeri Sumatera Utara Medan, Indonesia

ARTICLE INFO

Article History
Received : 28-08-2024
Revised : 29-09-2024
Accepted : 07-10-2024
Published : 06-11-2024

Keywords:
Clustering; Agglomerative Hierarchical Clustering Algorithm; K-Means Algorithm; Davies Bouldin Index Validation.

*Correspondence: E-mail:
wilia0703201058@uinsu.ac.id

Doi:
10.24042/djm.v7i3.24373

ABSTRACT

North Sumatra had the 17th lowest rate of poverty in 2023 out of 34 provinces, with 1,239.71 thousand people, or 8.15 percent, living there. Although there has been a decline in the poverty rate in 2023 compared to previous years, there are still many districts and cities in North Sumatra with significant rates of poverty; thus, this cannot be disregarded. The government must act to address this by providing the community with various forms of aid and increasing the number of job openings. To overcome this, one must first identify the cities or districts with the lowest to highest rates of poverty. This can be avoided with data mining, namely by applying the clustering technique. The Agglomerative Hierarchical Clustering (AHC) algorithm and the K-Means algorithm were the clustering techniques employed in this investigation. The Davies Bouldin Index (DBI) will then be used to validate the clustering results in order to ascertain which technique yields the best cluster. Three clusters were created using the AHC method: cluster 1 had 31 districts/cities, cluster 2 had one district/city, and cluster 3 had one district/city. Using the k-means approach, three clusters were identified: cluster 1, which included 22 districts/cities, had the lowest poverty rate; cluster 2, which included 10 districts/cities, had a moderate poverty rate; and cluster 3, which included 1 district/city, had the highest poverty rate. It was discovered through clustering validation that the k means method with a DBI value of 0.45 was the most effective approach for this investigation.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

One of the objectives of the state of Indonesia is to advance general welfare, as stated clearly in the fourth paragraph of the 1945 Constitution's preamble. Community welfare, on the other hand, refers to the degree of economic and social welfare of the community, which is

examined from eight angles, including poverty (Saputri & Arianto, 2023). When an individual or group of individuals cannot satisfy their fundamental needs, such as food, clothes, health care, and education, they are said to be in poverty (Hidayat, Putra, Alfitrah, & Widodo, 2023). The problem of poverty is usually in the midst of a society that is generally unemployed. There are various complex factors that can lead to poverty, namely social injustice, economic inequality, lack of educational opportunities, and increasing unemployment (Sachrrial & Iskandar, 2023). Because of this factor, poverty basically has a negative impact on people's lives, so this problem needs more attention from the government. Poverty in society is a major problem that is of concern to governments in various countries, including Indonesia (R, Anggraeni, & Enri, 2022). The government has tried multiple ways to overcome existing poverty by assisting, such as Poor Rice, the Smart Indonesia Program, the Family Hope Program, the National Health Insurance Program, and numerous other forms of support (Manurung, Sari Ramadhan, & Suryanata, 2020). Despite the many tactics adopted, the population's economic conditions in North Sumatra Province are defined by a high level of poverty.

According to Badan Pusat Statistik (2021), the number of poor people in 2023 is 1,239.71 people with a percentage of 8.15%. Although the poverty rate lowers year after year, the rate of reduction is insufficient, so a program is required to attain a low poverty level and prevent it from rising again. According to the districts/cities in North Sumatra Province, the government must determine which places have high, medium, and low poverty levels. Thus, data mining can be utilized to identify which locations will be prioritized based on the greatest level of poverty (Luchia, Handayani, Hamdi, Erlangga, & Octavia, 2022).

The process of looking through a big database and locating information is called data mining (Amna et al., 2023). Data mining is a technique that uses statistical methodologies, mathematics, artificial intelligence, and machine learning to extract and uncover significant information (Luthfi & Wijayanto, 2021). Clustering will be employed in data mining to group districts and cities with high poverty rates in North Sumatra Province. The goal of the data analysis technique known as clustering is to organize items or data according to similar attributes. Using cluster analysis, objects can be grouped into substantially similar groups where items in each category are typically distinct from one another and identical to each other (Asyfani et al., 2024). The Agglomerative Hierarchical Clustering (AHC) and K-Means approaches are employed in this work.

The AHC method is a hierarchical clustering technique that has the ability to progressively combine clusters with the highest similarity based on preset metrics or distances after merging n clusters into a single cluster (Sachrrial & Iskandar, 2023). According to Tuhpatussania, Erniwati, & Mutaqin (2024), a hierarchical structure is formed from bottom to top by grouping many data sets based on commonalities. K-Means clustering is a non-hierarchical technique for data clustering that seeks to partition, or divide, existing data into two or more groups, to group data that share similar features and data that differ from each other (Aprilia et al., 2022; Manurung et al., 2020). The two approaches will be compared to determine which produces the best clustering results.

The best cluster value can be found by applying the DBI validity test (Davies Bouldin-Index) to the analysis of the optimal algorithm's output. David L. Davies and Donald W. Bouldin established the Davies-Bouldin index (DBI) metric in 1979 with the intention of assessing

clusters (Tempola, Muhammad, & Mubarak, 2020). The best quality cluster results are those with lower DBI values (Riska & Farokhah, 2023).

Prior studies in this area, like those conducted by Sachrrial & Iskandar (2023), compared the AHC and K-Medoids Complete Linkage approaches for classifying poverty data in Indonesia. The study's findings indicate that using both approaches, three clusters with distinct cluster locations are generated. Luchia et al. (2022) examined two approaches to classifying poverty data in Indonesia: k-means and k-medoids and found that k-means outperforms k-medoids in the research, with the best DBI value of 0.041 and k = 8 in k-means.

This study focuses on the AHC and K-Means methodologies for clustering poverty data in North Sumatra, as well as the validation of the Davies Bouldin Index (Sujjada, Insany, & Noer, 2024). Thus, the purpose of this study is to determine the results of the clustering analysis carried out using the Agglomerative Hierarchical Clustering and K-Means algorithms to group the number of poor people based on districts/cities. The results obtained can be used as a benchmark for the government in reducing poverty rates in North Sumatra by prioritizing districts/cities with the highest poverty rates by opening job vacancies or distributing other government assistance. After the clustering was carried out, the validity test of the Davies Bouldin Index was carried out. The Davies Bouldin Index is used because this test is a validation metric used to evaluate the clustering model. In addition, this test is very effective compared to other tests. This test is flexible and can be used for several clusters. Unlike the Silhouette Score evaluation metric, there are no assumptions about the shape of the cluster, and this test is easy to use and intuitive.

METHOD

The research location is in North Sumatra Province and was conducted through BPS North Sumatra. The research began in February 2024 and lasted until completion. The essence of the research method is a scientific method for collecting data and information in a manner that is in accordance with current conditions and for a specific purpose (Yusri, 2020). The type of research in this study is quantitative, where this type of research is generally in the form of numbers or diagrams with the aim of collecting and analyzing data in the form of diagrams to identify numbers, trends, or relationships. This research is also relative research, where this type of research is used to compare two or more groups, variables, or situations to identify differences or similarities. This research uses a quantitative research system for its research. Because it can be used to compile and develop new technical wisdom by using research data in the form of numbers and statistical analysis, the quantitative system is known as a discovery system (Saputra, 2021). The study utilized report data, which is a sort of secondary data that was obtained from the BPS of North Sumatra. The researcher additionally studies relevant books and journal articles in order to perform a literature review on the subject of the investigation. The open severance rate (X_1), the number of impoverished people (X_2), the probability of becoming poor people (X_3), the poverty depth indicator (X_4), and the poverty inflexibility indicator (X_5) are the variables that were employed (Fikri, Mushardiyanto, Laudza'Banin, Maureen, & Patria, 2021; Latupeirissa, Lewaherilla, & Hiariey, 2022). The methodology utilized in this study begins with data collection via BPS North Sumatra. The number of clusters to be formed is then decided, followed by a clustering analysis utilizing the AHC (Agglomerative Hierarchical Clustering)

and K-Means methodologies. In order to compare the two styles employed and draw conclusions, the final stage is to validate the Davies Bouldin Index (DBI) and identify the stylish cluster. The steps taken in this research can be seen in Figure 1.

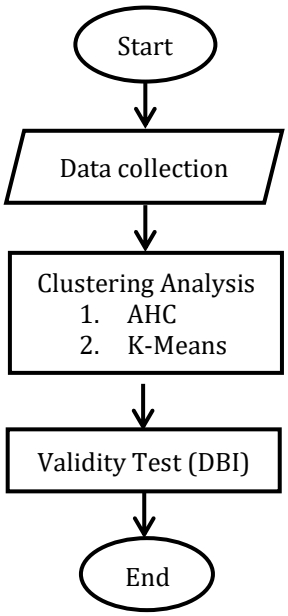


Figure 1. Research Flow

The AHC (Agglomerative Hierarchical Clustering) process begins by measuring the distance between each pair of data with the formula Euclidean Distance (Faran & Aldisa, 2023) $d(x,y\ ) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|^2}$. Then merging the two clusters that have the closest distance, forming a new cluster with the formula single linkage mode $d_{(ab)c} = min\{d_{a,c}; d_{b,c}\}$. This process continues by merging larger clusters until all data is combined into one main cluster.

In K-Means the first step is to determine the number of clusters, then determine the center value (centroid) randomly, then calculate the closest distance using Euclidean Distance with the formula $d(x,y\ ) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|^2}$ (Umagapi et al., 2023), and then group the distance to the nearest centroid. After all the processes are complete, it will be continued by determining the new centroid with the formula $V_{ij} = \frac{1}{n}\sum_{k=0}^{Ni} X_{kj}$, the same steps are carried out as before until the cluster formed does not change.

The process of determining the DBI value begins by finding the SSW value using the formula (Fathurrahman, Harini, & Kusumawati, 2023) $SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_i)$, then finding the SSB value using the formula $SSB_{i,j} = d(c_i, c_j)$, then the R-value using the formula using the equation $R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$, until finally determining the DBI value using the formula (Septiani, Fauzan, & Huda, 2022) $DBI = \frac{1}{K}\sum_{i=1}^{K} max_{i=j}(R_{i,j})$.

RESULTS AND DISCUSSION

In grouping poverty data for districts/cities in North Sumatra, the first thing to do after all the required data is complete is to conduct a descriptive analysis. The following are the overall results of each variable carried out using SPSS software in Table 1.

Table 1. Results of Descriptive Statistical Calculations

	X_1	X_2	X_3	X_4	X_5
Minimum	0.45	4.01	3.44	0.34	0.13
Maximum	8.67	187.28	754	3.04	0.84
Mean	4.5530	37.5679	32.5455	1.3176	0.3536
Standard Deviation	2.55136	34.29833	129.58089	0.60101	0.19591

A multicollinearity test is then performed to determine which similarity metrics can be applied. To establish whether multicollinearity occurs, consider the tolerance and variance inflation factor, or VIF. According to the tolerance value decision-making guideline, multicollinearity does not occur if the

tolerance value exceeds 0.10. When the tolerance value is less than 0.10, multicollinearity exists. According to Riswanda, Kusnandar, & Imro'ah (2023), if the VIF value is less than 10, multicollinearity does not exist. If the VIF value is greater than 10, multicollinearity does arise. The multicollinearity test results can be seen in Table 2.

Table 2. Multicollinearity Test Results

Variable	Tolerance	VIF
X_1	0.623	1.604
X_2	0.754	1.325
X_3	0.567	1.764
X_4	0.543	1.840
X_5	0.536	1.865

Table 2 shows that each variable's VIF value is less than 10 and that the tolerance value is more than 0.10. Thus, it may be said that these variables do not exhibit multicollinearity. Euclidean distance can therefore be applied to clustering using the k-means approach.

Algoritma Agglomerative Hierarchical Clustering (AHC)

Agglomerative Hierarchical Clustering (AHC) is the first technique applied. The first step is to use the formula to determine the Euclidean distance.

$d(Nias,Nias) = \sqrt{(2.31-2.31)^2 + (21.99-21.99)^2 + (15.1-15.1)^2 + (1.95-1.95)^2 + (0.4-0.4)^2} = 0$

$d(Nias,Mandailing\ Natal) = \sqrt{(2.31-7.54)^2 + (21.99-41.04)^2 + (15.1-8.86)^2 + (1.95-1.48)^2 + (0.4-0.44)^2} = 20.70$

The calculation was carried out until the last district, namely Mount Sitoli. If the results of the distance calculation above are included in a table, it will form a matrix as in Table 3.

Table 3. Euclidean Distance Calculation AHC Method

District/City	Nias	Mandailing Natal	Tapanuli Selatan	Tapanuli Tengah	...	...	Gunung Sitoli
Nias	0	20.70	8.46	25.95	...	...	1.41
Mandailing Natal	20.70	0	21.41	6.61	...	...	20.27
...	...	...	...	...	...	...	...
Gunung Sitoli	1.41	20.27	8.10	25.62	...	...	0

Based on the calculation of the Euclidean distance that has been carried out, it can be known that the smallest value is located in the district/city of Nias and Mount Sitoli with a value of 1.41. So, the next step is to group districts/cities using the single linkage mode.

d(Nias, Gunung Sitoli)Nias, Gunung Sitoli = min {0 ; 1.41} = 0

d(Nias, Gunung Sitoli)Mandailing Natal = min {20.70; 20.27} = 20.27

d(Nias, Gunung Sitoli)Tapanuli Selatan = min {8.46; 8.10} = 8.10

The calculation is carried out as in the previous step until the cluster is formed into 3 clusters. The final results of grouping using the AHC method are obtained in the following table 4. In Table 4, it can be seen that the cluster results are uneven; there is a very high difference in numbers. This is because the AHC (Agglomerative Hierarchical Clustering) process begins by measuring the distance between each pair of data and then combining the two clusters that have the closest distance, forming a new cluster. This process continues by combining larger clusters until all data is combined into one main cluster. In the last step, it is explained that all data will be combined into one main cluster, but in this study, it was formed into three clusters; this is what causes the high difference.

Table 4. Results of AHC Method Clustering

Cluster 1	Cluster 2	Cluster 3
Nias, Mandailing Natal, Tapanuli Selatan, Tapanuli Tengah, Tapanuli Utara, Toba, Labuhan Batu, Asahan, Simalungun, Dairi, Karo, Deli Serdang, Nias Selatan, Humbang Hasundutan, Pakpak Barat, Samosir, Serdang Bedagai, Batu Bara, Padang Lawas Utara, Padang Lawas, Labuhanbatu Selatan, Labuhanbatu Utara, Nias Utara, Nias Barat, Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Binjai, Padang Sidimpuan, Gunung Sitoli	Langkat	Medan

K-Means Clustering

The k-means method is used to group the impoverished in North Sumatra province. The first step is to determine the number of clusters to build. The number of clusters to be formed is three. Next, the center value (centroid) was determined randomly, along with the initial centroid selected based on the data used in the study.

Table 5. Centroid Early K-Means Method

Centroid	TPT	JPM	PPM	P_1	P_2
C_1	1.3	14.94	8.04	1.02	0.18
C_2	3.49	20.09	7.01	0.92	0.18
C_3	7.81	47.09	11.5	1.58	0.3

Then the third is to calculate the closest distance to the centroid using the Euclidean distance.

$d(x_1,c_1) = \sqrt{((2.31-3.1)^2 + (21.99-14.94)^2 +(15.1-8.04)^2 +1.95-1.02)^2 + (0.4-0.18)^2)} = 10.07$

$d(x_2,c_1) = \sqrt{((7.54-3.1)^2 + (41.04-14.94)^2 +(8.86-8.04)^2 +1.48-1.02)^2 + (0.44-0.18)^2)} = 26.83$

Description: $x_{1,2,3,...}$ is the district/city, and $c_{1,2,3}$ is the centroid.

The calculation is done up to $d(x_{33},c_3)$. The results of the Euclidean distance calculation can determine the clustering by looking at the nearest distance, which is the next step in the k-means method.

Table 6. Results of 1st Iteration Clustering

C_1	C_2	C_3	Closest Distance	Cluster
25.95	8.46	10.07	8.46	2
6.61	21.41	26.83	6.61	3
...	...	...	...	...
25.62	8.10	10.12	8.10	2

The next step is to define a new centroid for the calculation of the next iteration.

$c_{11} = 7.45 + 7.81 + 5.99 + 6.12 + 5.35 + 2.63 + 8.62 + 6.33 + 3.48 + 4.97 + 5.88 + 4.84 +8.67)/13 = 6.01$

$c_{12} = 41.04 + 47.09 + 42.58 + 61.69 + 69.21 + 35.65 + 82.7 + 98.16 + 54.29 + 45.88 + 49.18 + 34.13 +187.28)/13 = 65.30$. And so on until C_{35}.

Table 7. Centroid's New K-Means Method

Centroid	TPT	JPM	PPM	P_1	P_2
C_1	2.47	13.38	10.71	1.20	0.23
C_2	4.22	22.86	10.40	1.35	0.29
C_3	6.01	65.30	9.03	1.35	0.34

Table 8. Final Results of the 10th Iteration of K-Means Method Clustering

C_1	C_2	C_3	Closest Distance	Cluster
165.57	37.86	5.10	5.10	1
146.25	18.19	20.52	18.19	2
...	...	...	...	...
165.47	37.66	4.66	4.66	1

Because the grouping process stopped at the 10th iteration, the results of the clustering were obtained with cluster 1 with as many as 1 district/city, cluster 2

with as many as 10 districts/cities, and cluster 3 with as many as 22 districts/cities.

Table 9. Results of the K-Means Method Clustering

Cluster 1	Cluster 2	Cluster 3
Nias, Tapanuli Selatan, Tapanuli Utara, Toba, Dairi, Karo, Humbang Hasundutan, Pakpak Barat, Samosir, Padang Lawas Utara, Padang Lawas, Labuhanbatu Selatan, Labuhanbatu Utara, Nias Utara, Nias Barat, Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Binjai, Padang Sidimpuan, dan Gunung Sitoli	Mandailing Natal, Tapanuli Tengah, Labuhanbatu, Asahan, Simalungun, Deli Serdang, Langkat, Nias Selatan, Serdang Bedagai, dan Batu Bara	Medan

Validasi Agglomerative Hierarchical Clustering (AHC)

First, the cluster data distance will be determined using the clustering results obtained in the previous clustering process as well as the centroid, which is the result of clustering. In the calculation of distance, the Euclidean distance formula is used.

Table 10. Centroid AHC

Centroid	TPT	JPM	PPM	P_1	P_2
C_1	0	16.74	104.63	0	0
C_2	16.74	0	89.17	0	0
C_3	104.63	89.17	0	0	0

$d(x_1,c_1) = \sqrt{((0-2.31)^2 + (16.74-21.99)^2 + (104.63-15.1)^2 + (0-1.95)^2 + (0-0.4)^2} = 89.74$

$d(x_1,c_1) = \sqrt{((0-7.54)^2 + (16.74-41.04)^2 + (104.63-8.86)^2 + (0-1.48)^2 + (0-0.44)^2} = 99.10$

After the calculation of the distance between clusters is carried out to $d(x_3, c_1)$, then the next step is to calculate the SSW value.

$SSW_1 = (89.74 + 99.10 + 97.74 + 98.27 + 96.58 + 96.62 + 100.22 + 106.56 + 110.2 +\ldots+ 98.08 + 90.11)/31 = 97.34$

$SSW_2 = 127.04/1 = 127.04$

$SSW_3 = 137.47/1 = 137.47/1$

The third step is to calculate the SSB value.

$SSB_{1,C1} = \sqrt{((0-0)^2 + (16.74-16.74)^2 + (104.63-104.63)^2 + (0-0)^2 + (0-0)^2)} = 0$

$SSB_{1,C2} = \sqrt{((0-16.74)^2 + (16.74-0)^2 + (104.63-89.17)^2 + (0-0)^2 + (0-0)^2)} = 28.27.$

Table 11. SSB Matrix AHC Method

	Centroid		
SSB	**1**	**2**	**3**
1	0	28.27	164.75
2	28.27	0	153.71
3	164.75	153.71	0

The fourth step in finding the DBI is to find the value of the comparison ratio between the ith cluster and the j cluster.

$R_{1,2} = \frac{97.34+127.04}{28.27} = 7.94$

$R_{1,3} = \frac{97.34+137.47}{164.75} = 1.43$

$R_{2,3} = \frac{127.04+137.47}{153.71} = 1.72$

If transformed into a matrix, it will be like a table with $R_{max.}$

Table 12. R_{max} AHC Method

R	1	2	3	R_{max}
1	0	7.94	1.43	7.94
2	7.94	0	1.72	7.94
3	1.43	1.72	0	1.72

For the last step of DBI calculation.

$DBI = (7.94 + 7.94 + 1.72)/3 = 5.86$

Based on the description above, the DBI value for the AHC method is 5.86.

Validasi K-Means Clustering

First, it will be determined that the cluster data distance uses the clustering results obtained in the previous clustering

process as well as the centroids taken from the centroids in the last iteration.

Table 13. Centroid K-Means

Centroid	TPT	JPM	PPM	P_1	P_2
C_1	3.62	20.94	10.33	1.29	0.27
C_2	6.2	59.19	9.23	1.42	0.37
C_3	8.67	187.28	8	0.92	0.19

$d(x_1,c_1) = \sqrt{((3.62-2.31)^2 + (20.94-21.99)^2 + (10.33-15.1)^2 + (1.29-1.95)^2 + (0.27-0.4)^2} = 5.10$

$d(x_3,c_1) = \sqrt{((3.62-3.49)^2 + (20.94-20.09)^2 + (10.33-7.01)^2 + (1.29-0.92)^2 + (0.27-0.18)^2} = 3.45$. And so on until $d(x_n, c_3)$.

After the calculation of the distance between clusters is carried out, the next step is to calculate the SSW value using the formula in the previous equation in the AHC validation.

$SSW_1 = (5.10 + 3.45 + 6.31 + 6.83 + 3.77 + 14.93 + \ldots + 7.13 + 4.66)/22 = 8.06$

$SSW_2 = (18.19 + 12.41 + 16.65 + 2.79 + 10.16 + \ldots + 13.49 + 10.24)/10 = 15.66$

$SSW_3 = 0$

The third step is to calculate the SSB value.

$SSB_{1,C1} = \sqrt{((3.62-3.62)^2 + (20.94-20.94)^2 + (10.33-10.33)^2 + (1.29-1.29)^2 + (0.27-0.27)^2)} = 0$

$SSB_{1,C2} = \sqrt{((3.62-6.2)^2 + (20.94-59.19)^2 + (10.33-9.23)^2 + (1.29-1.42)^2 + (0.27-0.37)^2)} = 38.35$

And so on, such as the calculations carried out on the validation of AHC.

Table 14. SSB K-Means Matrix

	Centroid		
SSB	**1**	**2**	**3**
1	0	38.35	166.44
2	38.35	0	128.12
3	166.44	128.12	0

Finding the value of the comparison ratio between the i-th and j-th clusters is the fourth step in determining DBI.

$$R_{1,2} = \frac{8.06+15.66}{38.35} = 0.62$$

$$R_{1,3} = \frac{8.06+0}{166.44} = 0.05$$

$$R_{2,3} = \frac{15.66+0}{128.12} = 0.12$$

If transformed into a matrix, it will be like a table with R_{max}.

Table 15. R_{max} K-Means

R	1	2	3	R_{max}
1	0	0.62	0.05	0.62
2	0.62	0	0.12	0.62
3	0.05	0.12	0	0.12

The last step of calculating the DBI value is as follows:

$$DBI = (0.62 + 0.62 + 0.12)/3 = 0.45$$

Based on the description above, the DBI value for the k-means clustering method is 0.45. The DBI value shows how well clustering has been carried out by calculating the quantities and criteria derived from the dataset. The smaller the DBI value, the better the clustering that is formed. So in this study, it was found that the validation process using the k-means method is the best method for grouping districts/cities in North Sumatra. When compared with previous studies, the results obtained showed that using the AHC method is better with a value of 0.57 compared to the K-Means method with a DBI value of 2.02 (Zuhal, 2022).

CONCLUSIONS AND SUGGESTIONS

Two algorithms are used to group districts/cities based on low, medium, and high levels of poverty: the agglomerative hierarchical clustering (AHC) algorithm and the k-means clustering algorithm. Following the clustering procedure, each approach produced three clusters, each of which contained a different district or city. Three districts/cities comprise Cluster 1 (AHC technique), Cluster 2 (one district/city), and Cluster 3 (one district/city). Meanwhile, 22 districts/cities comprise Cluster 1 in the k-means approach, 10 districts/cities comprise Cluster 2, and 1 district/city

comprises Cluster 3. The ultimate goal of this study is to compare the two approaches taken to obtain the best clustering results. Clustering validation techniques, especially Davies Bouldin Index (DBI) validation, are used to compare the two approaches. As determined by previous findings and discussions, the k-means clustering algorithm method produces a DBI value of 0.45 while the AHC method produces a DBI value of 5.86. Based on the results of the DBI value above, the smallest DBI value is obtained in the K-Means method, so the K-Means method is the best method used in this study. The first cluster consists of 20 districts/cities (Nias, South Tapanuli, North Tapanuli, Toba, Dairi, Karo, Humbang Hasundutan, West Pakpak, Samosir, North Padang Lawas, South Labuhanbatu, North Labuhanbatu, North Nias, West Nias, Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Binjai, Padang Sidimpuan, and Gunung Sitoli. Mandailing. A total of ten districts/cities, including Mandailing Natal, Labuhanbatu, Asahan, Simalungun, Deli Serdang, Langkat, South Nias, Serdang Bedagai, and Batu Bara, are included in Cluster 2. The last cluster, namely Medan City, consists of one district/city and is the cluster with the highest poverty rate..

Researchers can further develop their suggestions for grouping poverty in an area using other clustering methods, both hierarchical and non-hierarchical. Researchers can also add more variables to further optimize the clustering results, and researchers can also use software to make the clustering results more accurate.

REFERENCES

Amna, S, W., Sudipa, I. G. I., Putra, T. A. E., Wahidin, A. J., Syukrilla, W. A., … Santoso, L. W. (2023). *Data mining*. PT Global Eksekutif Teknologi.

Aprilia, R., Afsari, K., Rahma, R., Nasution, N., Ouri, S., & Putri, D. (2022). Analisis cluster dengan metode k-means cluster pada jenis data surat di bpprd sumatera utara. *Jurnal Pengabdian Kepada Masyarakat*, *6*(2).

Asyfani, Y., Nur, I. M., Amri, I. F., Yunanita, N., Lestari, F. A., Hisani, Z. A., & Rohim, F. H. N. (2024). Pengelompokan kabupaten/kota di jawa tengah berdasarkan kepadatan penduduk menggunakan metode hierarchical clustering. *Journal of Data Insights*, *2*(1), 1–8.

Badan Pusat Statistik. (2021). Profil kemiskinan provinsi sumatera utara. In *Badan Pusat Statistik Provinsi Sumatera Utara*.

Faran, J., & Aldisa, R. T. (2023). Analisis data mining dalam komparasi average linkage ahc dan k-means clustering untuk dataset facebook live sellers. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(4), 2041. https://doi.org/10.30865/mib.v7i4.6892

Fathurrahman, F., Harini, S., & Kusumawati, R. (2023). Evaluasi clustering k-means dan k-medoid pada persebaran covid-19 di indonesia dengan metode davies-bouldin index (dbi). *Jurnal Mnemonic*, *6*(2), 117–128. https://doi.org/10.36040/mnemonic.v6i2.6642

Fikri, R., Mushardiyanto, A., Laudza'Banin, M. N., Maureen, K., & Patria, H. (2021). Pengelompokan kabupaten/kota di indonesia berdasarkan informasi kemiskinan tahun 2020 menggunakan metode k-means clustering analysis. *Seminar Nasional Teknik Dan Manajemen Industri*, *1*(1), 190–199. https://doi.org/10.28932/sentekmi2021.v1i1.76

Hidayat, F. P., Putra, R. P., Alfitrah, M. D., & Widodo, E. (2023). Implementasi clustering k-medoids dalam pengelompokan kabupaten di provinsi aceh berdasarkan faktor

yang mempengaruhi kemiskinan. *Indonesian Journal of Applied Statistics*, *5*(2), 121. https://doi.org/10.13057/ijas.v5i2.55080

Latupeirissa, S. J., Lewaherilla, N., & Hiariey, A. (2022). Pengelompokkan kabupaten/kota di provinsi maluku berdasarkan data kemiskinan tahun 2021 mengunakan metode k- means cluster. *Variance*, *4*, 15–22.

Luchia, N. T., Handayani, H., Hamdi, F. S., Erlangga, D., & Octavia, S. F. (2022). Perbandingan k-means dan k-medoids pada pengelompokan data miskin di indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2). https://doi.org/10.57152/malcom.v2i2.422

Luthfi, E., & Wijayanto, A. W. (2021). Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia. *INOVASI*, *17*(4), 761–773. https://doi.org/10.30872/jinv.v17i4.10106

Manurung, J., Sari Ramadhan, P., & Suryanata, M. (2020). Perbandingan algoritma k-means dan k-medoids untuk pengelompokkan data masyarakat miskin pada kantor camat hatonduhan stmik triguna dharma. *Jurnal CyberTech*, *3*(9).

R, N. N. F., Anggraeni, D. S., & Enri, U. (2022). Pengelompokkan data kemiskinan provinsi jawa barat menggunakan algoritma k-means dengan silhouette coefficient. *TEMATIK*, *9*(1). https://doi.org/10.38204/tematik.v9i1.901

Riska, S. Y., & Farokhah, L. (2023). Perbandingan hasil evaluasi algoritma k-means dan k-medoid berdasarkan kunjungan wisatawan mancanegara ke indonesia. *INTEGER: Journal of Information Technology*, *8*.

Riswanda, G. P., Kusnandar, D., & Imro'ah, N. (2023). Perbandingan klaster k-means dan k-median pada data indikator kemiskinan kabupaten/kota di provinsi kalimantan barat. *BIMASTER: Buletin Ilmiah Math. Stat. Dan Terapannya*, *12*(6), 537–544.

Sachrrial, R. H., & Iskandar, A. (2023). Analisa perbandingan complate linkage ahc dan k-medoids dalam pengelompokkan data kemiskinan di indonesia. *Building of Informatics, Technology and Science (BITS)*, *5*(2). https://doi.org/10.47065/bits.v5i2.4310

Saputra, N. (2021). Metodelogi penelitan kuantitatif. In *Pascal Books* (Vol. 11).

Saputri, F. W., & Arianto, D. B. (2023). Perbandingan performa algoritma k-means, k- medoids, dan dbscan dalam penggerombolan provinsi di indonesia berdasarkan indikator kesejahteraan masyarakat. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *7*(2), 138–151. https://doi.org/10.47111/jti.v7i2.9558

Septiani, I. W., Fauzan, Abd. C., & Huda, M. M. (2022). Implementasi algoritma k-medoids dengan evaluasi davies-bouldin- index untuk klasterisasi harapan hidup pasca operasi pada pasien penderita kanker paru-paru. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *3*(4), 556. https://doi.org/10.30865/json.v3i4.4055

Sujjada, A., Insany, G. P., & Noer, S. (2024). Analisis clustering data penyandang disabilitas menggunakan metode agglomerative hierarchical clustering dan k-means. *Jurnal Teknologi Dan Manajemen Informatika*, *10*(1), 1–12. https://doi.org/10.26905/jtmi.v10i1.10654

Tempola, F., Muhammad, M., & Mubarak, A. (2020). Penggunaan internet

dikalangan siswa sd di kota ternate: Suatu survey, penerapan algoritma clustering dan validasi dbi. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(6). https://doi.org/10.25126/jtiik.2020722370

Tuhpatussania, S., Erniwati, S., & Mutaqin, Z. (2024). Perbandingan metode agglomerative hierarchical clustering dan metode k medoids dalam pengelompokkan data titik panas. *Journal Computer and Technology*, *2*(1), 21–38.

Umagapi, I. T., Umaternate, B., Komputer, S., Pasca Sarjana Universitas Handayani, P., Kepegawaian Daerah Kabupaten Pulau Morotai, B., & Riset dan Inovasi, B. (2023). Uji kinerja k-means clustering menggunakan davies-bouldin index pada pengelompokan data prestasi siswa. *Seminar Nasional Sistem Informasi Dan Teknologi (SISFOTEK)*, *7*(1).

Yusri, A. Z. (2020). Teori, metode dan praktik penelitian kualitatif. *Jurnal Ilmu Pendidikan*, *7*(2).

Zuhal, N. K. (2022). Study comparison k-means clustering dengan algoritma hierarchical clustering. *Prosiding Seminar Nasional Teknologi Dan Sains*, *1*.