

Manhattan, Euclidean And Chebyshev Methods In K-Means Algorithm For Village Status Grouping In Aceh Province

Amali*¹, Gatot Tri Pranoto²

^{1,2}Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University
email: amali@pelitabangsa.ac.id^{*1}, gatot.pranoto@pelitabangsa.ac.id²

*Corresponding author

Abstract - The Ministry of Villages, Development of Disadvantaged Regions and Transmigration (Ministry of Villages PDTT) is a ministry within the Government of Indonesia in charge of developing villages and rural areas, empowering rural communities, accelerating the development of disadvantaged areas, and transmigration. The 2014 Village Potential Data (Podes 2014) is data released by the Central Statistics Agency in collaboration with the Ministry of Villages PDTT in unsupervised form and consists of 6474 villages in the province of Aceh. Podes 2014 data is based on the level of village development (village specific) in Indonesia by using the village as the unit of analysis. Data mining is a method that can be used to group objects in a data into classes that have the same criteria (clustering). One of the algorithms that can be used for the clustering process is the k-means algorithm. Grouping data using k-means is done by calculating the shortest distance from a data point to a centroid point. In this study, a comparison of the distance calculation method on k-means between Manhattan, Euclidean and Chebyshev will be carried out. Tests will be performed using the execution time and the Davies boulder index. From the tests that have been carried out, it is found that the number of villages in each cluster is 2,639 developing villages, 1,188 independent villages, 1,182 very underdeveloped villages, 1,266 developed villages and 199 disadvantaged clusters. The Chebyshev distance calculation method has the most efficient accumulation of time compared to Manhattan and Euclidean, while the Euclidean method has the most optimal Davies Index.

Keywords – Davies Bouldin index, rural development, k-means, Manhattan, Euclidean, Chebyshev.

1. INTRODUCTION

The State of Indonesia is a unitary state in the form of a republic and a legal state, where sovereignty is in the hands of the people which is implemented according to the constitution. The Unitary State of the Republic of Indonesia is divided into provinces, wherein the provinces are divided into districts and cities [1]. Each district in Aceh Province consists of villages that have origin rights, traditional rights in regulating and managing the interests of the local community and play a role in realizing the ideals of independence. In the course of the state administration of the Republic of Indonesia, villages have developed in various forms so that they need to be protected and empowered to become strong, advanced, independent and democratic. To protect and empower the village, village autonomy was formed [2]. With village autonomy and Presidential Regulation Number 165 of 2014 concerning the Arrangement of

Duties and Functions of the Working Cabinet, the Ministry of Villages, Development of Disadvantaged Regions, and Transmigration of the Republic of Indonesia was formed.

The Ministry of Villages, Development of Disadvantaged Regions, and Transmigration (Kementerian Desa PDTT) is a ministry within the Indonesian Government led by the Minister and responsible to the President. The Ministry of Villages PDTT is in charge of village and rural area development, empowerment of rural communities, acceleration of development of underdeveloped areas, and transmigration [3]. The National Medium-Term Development Plan (RPJMN) of the Ministry of Villages PDTT 2015-2019 is a strategic document of development plans that must be carried out by the government for the next five years. This RPJMN document contains village development reducing targets that must be achieved, namely the number of Disadvantaged Villages to 5,000 villages and increasing the number of Independent Villages to at least 2,000 villages in 2019 [4]. In the context of village development, the Government and Regional Governments are required to develop a village information system and rural area development [5].

In developing a village information system, it is necessary to provide data about the village. The Ministry of Villages PDTT in collaboration with the National Development Planning Agency and the Central Statistics Agency issued data on Village Potential in 2014 (Podes 2014) consisting of 6474 villages in Aceh Province. and has 42 indicators/attributes dependent without a village status label. The 2014 Podes data is a measurement method that is compiled based on the level of village development in Aceh Province which makes the village a unit of analysis by referring to Law Number 6 of 2014 concerning villages, which is intended to capture the level of village development in Aceh Province and can be used as a reference for the preparation of policy planning and monitoring of village development [6]. With the 2014 Podes data, it can be used as a reference for the preparation of policy planning. Based on the PDTT Ministry of Village Regulation number 2 of 2016, the status of the village is grouped into 5 statuses, namely Independent Village, Advanced Village, Developing Village, Disadvantaged Village and Very Disadvantaged Village [7]. There is no grouping of Village Potential data in 2014 into 5 village statuses in Aceh Province and the current grouping into 5 village statuses is using the Developing Village Index data which is grouped by regions in Aceh Province. Making the grouping of village status in Aceh Province using 2014 Village Potential data is needed.

In information technology, data is an important part that cannot be separated from information retrieval. Information related to the status of the village as mentioned above can be obtained using a data mining process on the 2014 Podes data. Data mining is a series of activities used to find new, hidden or unexpected patterns contained in the data. The term data mining is often considered as a synonym for knowledge discovery from data (KDD), namely the discovery of knowledge from data that focuses on the purpose of the mining process [8]. Data mining can be used to perform clustering, classification and association. Clustering_ namely the process of grouping data which is done by finding the similarity of characteristics between the data according to certain class groups [9]. In simple terms, clustering can be used to analyze a set of data and generate a set of clustering rules that can be used to group future data.

In the real world, sometimes data is not only grouped into binary status (binary) class, but it also needs to be grouped into multi-status (multi-class). In the case of multi-class data-sets, grouping will be more difficult than in the case of binary. There are several algorithms that can be used for multi-class. A study has been conducted on the use of the k-means algorithm for multi-class which shows that the k-means algorithm provides effective results for grouping multi-class data-sets [10]. The k-means algorithm is an interactive clustering algorithm that partitions the data-set number of K clusters a predetermined In another study, we compared hierarchy partition-based clustering, clustering -based clustering and revealed density-based which that the k-means algorithm is a partition-based algorithm that provides better

performance, is able to divide clusters well and is superior to large/lots of data compared to other algorithms. clustering based on hierarchy and density [11] [23]. In addition, several other studies also mention that clustering using the k-means algorithm is faster than clustering with other algorithms and also produces clusters when using data-sets [13] [14] [15] [16]. In performing grouping, the k-means algorithm requires a distance calculation method to calculate the closest distance between an instance data centroid point.

The calculation of the distance in the k-means algorithm can use Manhattan, Euclidean and Chebychev. Research has been done on the comparison of Manhattan and Euclidean distance calculation methods on the k-means algorithm to determine the number of squared errors, the data used in this study is a data-set bank that was tested using the WEKA tool [17]. The test results show that the Manhattan distance calculation method is better than the Euclidean method [17]. In another study, a comparison of 3 methods of calculating distances in the k-means algorithm was carried out, namely Manhattan, Euclidean and Minkowski to find the best distance calculation method, the study was carried out by comparing the results of previous studies which concluded that the Euclidean distance calculation method was better than the Euclidean distance calculation method. Manhattan and Minkowski method [18]. There have also been other studies on the comparison of the Manhattan, Euclidean and Chebyshev distance calculation methods on the k-means algorithm to determine the accuracy and mean absolute error [19]. From the tests conducted using the flower data-set, it was found that the Chebyshev distance calculation method is better than the Manhattan and Euclidean methods [19]. From previous studies, it is known that the Manhattan, Euclidean and Chebyshev distance calculation methods are superior to each other depending on the data-set used.

Based on the considerations mentioned above, this study will group village status in Aceh Province using the k-means algorithm into 5 village statuses, and compare which distance calculation method is the most effective for the 2014 Podes data grouping.

2. RESEARCH METHOD

This research is a quantitative method where there is data on Village Potential in 2014 (Podes 2014) in Aceh Province which will be processed. In this study, will be carried out clustering village status in Aceh Province, using the k-mean algorithm with the Manhattan, Euclidean and Chebyshev distance calculation methods into 5 village statuses, as well as comparing which distance calculation method has the most efficient accumulation of time and which has the most efficient time accumulation. The value Davies Index is most optimal.

2.1 Data Collection Methods The data

Used in this study is the potential of villages in Aceh Province in 2014 (Podes 2014). Podes 2014 data is secondary data issued by the Central Statistics Agency based on Law Number 6 of 2014 concerning villages. The 2014 Podes data consists of 6,474 instances and has 42 dependent attributes without labels. The description of the 2014 Podes data owned can be seen in Table 1.

Table 1. Image data sets Podes 2014

No	Village Name	I1	I2	I3	I4	I5	I6	I7	I8	I40	I41	I42
1	KEUDE BAKONGAN	3	3	4	4	3	3	4	4	5	5	5
2	UJUNG MANGKI	3	3	4	4	3	3	4	4	3	4	2
3	UJUNG PADANG	3	3	4	4	3	3	4	4	3	4	5
4	KAMPONG DRIEN	3	3	4	3	3	3	3	4	3	4	3
5	DARUL IKHSAN	4	3	4	4	3	3	4	4	3	4	3

No	Village Name	I1	I2	I3	I4	I5	I6	I7	I8	I40	I41	I42
6	PADANG BERAHAN	3	3	4	4	3	3	3	4	3	4	3
7	GAMPONG BARO	3	3	4	4	3	3	3	4	3	5	3
8	FAJAR HARAPAN	3	3	4	4	3	3	3	4	3	5	2
..
..
6473	BANGUN SARI	3	4	4	4	3	3	4	4	4	4	3
6474	DARUSSALAM	3	4	4	4	3	3	3	3	4	4	3

In Table 1 there are attributes I1, I2, I3 to I42 where the initial I stands for "Indicator". The value of each attribute is 0 to 5, where the value 0 is the lowest value while the value 5 is the highest value.

Data on the potential of villages in Aceh Province in 2014 is the result of measurements made based on the level of village development (village specific) in Aceh Province by using the village as the unit of analysis. Podes 2014 in Aceh Province is used as a reference for the main indicators that make up the index, and data on government administration areas (MDNR Indonesia, 2015) which is used as a reference standard for the number of integrated villages in Aceh Province. Podes 2014 in Aceh Province is a complex multidimensional concept consisting of dimensions, variables and indicators that are used as measuring tools for village development. In the 2014 Podes data in Aceh Province, there are 5 dimensions, 12 variables and 42 indicators. Completely from the dimensions, variables and indicators can be seen in Table 2.

Table 2. Table of Indicators for Compiling Podes 2014

code	Description Indicator	Variable Indicators	Dimension Indicator
I1	Availability and access to TK/RA/BA	Educational	Services Basic Services
I2	Availability and access to SD equivalent		
I3	Availability and access to SMP equal		
I4	Availability and access to SMA equal		
I5	Availability and easy access to hospitals	Health Services	
I6	Availability and ease of access to maternity hospitals		
I7	Availability and ease of access to puskesmas		
I8	Availability and ease of access to polyclinics/medical centers		
I9	Availability and ease of access to doctors		
I10	Availability and ease of access to midwives' practices		
I11	Availability and convenience access to poskesdes or polindes		
I12	Availability and ease of access to pharmacies		
I13	Availability of shops, minimarkets, or grocery stores	Economics	Infrastructure Condition of Infrastructure
I14	Availability of markets		
I15	Availability of restaurants, restaurants or food stalls/shops		
I16	Availability of hotel accommodation or lodging		
I17	Bank availability aan	of Energy Infrastructure	
I18	Electrification		
I19	Lighting conditions on main roads		

code	Description Indicator	Variable Indicators	Dimension Indicator	
I20	Fuel for cooking	Health and Sanitation Infrastructure		
I21	Sources of drinking water			
I22	Sources of water for bathing/washing			
I23	Defecation facilities			
I24	Availability and quality of cellular communication facilities	Communication and Information Infrastructure		
I25	Availability of internet facilities and postal or goods delivery			
I26	Traffic and road quality	Means of Transportation	Accessibility / Transportation	
I27	Accessibility			
I28	Availability of public transportation			
I29	Transport operations			
I30	Travel time per kilometer of transportation to the sub-district office	Transportation Accessibility		
I31	Cost per kilometer of transportation to the sub-district office			
I32	Travel time per kilometer of transportation to the regent			
I33	Cost per kilometer of transportation to the regent			
I34	Handling of extraordinary events	Public Health Public		Services
I35	Handling of malnutrition			
I36	Availability of sports facilities	Sports		
I37	Existence of k groups Sports			
I38	Completeness of village government	Independence		of Government Administration
I39	Village autonomy			
I40	Assets/wealth			
I41	Quality of human resources of the village head	Quality of Human Resources		
I42	Quality of human resources of village secretary			

2.2 Sample Selection Methods

Data samples taken randomly from the original data is data 2014 Village Potential in Aceh province as many as 15 villages and 42 indicators initialized as I1 to I42 to be grouped using the k-means algorithm, the data obtained are shown in Table 3.

Table 3. Table of Indicators for Compiling Podes 2014

x	Village Name	I1	I2	I3	I4	I5	I41	I42
1	KEUDE BAKONGAN	3.00	3.00	4.00	4.00	3.00	4.00	3.00
2	UJUNG MANGKI	3.00	3.00	4.00	4.00	3.00	5.00	5.00
3	UJUNG PADANG	2.00	4.00	4.00	4.00	3.00	4.00	3.00
4	KAMPONG DRIEN	2.00	3.00	4.00	4.00	3.00	4.00	3.00
5	DARUL IKHSAN	3.00	4.00	5.00	5.00	0.00	4.00	5.00
6	PADANG BERAHAN	4.00	4.00	4.00	4.00	3.00	5.00	3.00
7	GAMPONG BARO	4.00	3.00	4.00	3.00	3.00	4.00	3.00
8	FAJAR HARAPAN	3.00	4.00	4.00	2.00	3.00	0.00	3.00

x	Village Name	I1	I2	I3	I4	I5	I41	I42
9	KRUENG BATEE	2.00	3.00	3.00	4.00	1.00	4.00	3.00
10	PASI KUALA ASAHAN	3.00	3.00	3.00	3.00	3.00	4.00	3.00
11	GUNUNG PULO	3.00	3.00	4.00	3.00	3.00	4.00	5.00
12	PULO IE I	3.00	4.00	4.00	3.00	3.00	4.00	0.00
13	JAMBO MANYANG	4.00	4.00	5.00	4.00	3.00	4.00	5.00
14	SIMPANG EMPAT	4.00	4.00	5.00	5.00	3.00	4.00	5.00
15	LIMAU PURUT	4.00	3.00	2.00	3.00	3.00	4.00	3.00

From table 3 will be used for grouping into 5 clusters initialized as C1 SD C5. The grouping will be done using the k-means algorithm with three different distance calculation methods, namely Manhattan, Euclidean and Chebyshev.

2.3 Evaluation

In this study, the evaluation will be carried out by grouping the village potential data. The grouping will use the k-mean algorithm where the distance calculation will use Manhattan, Euclidean and Chebyshev. The results of the grouping obtained are the grouping of village potential data into 5 clusters, namely cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4. Until this stage it is not a known cluster which can be called a cluster independent village, developed villages, developing villages, underdeveloped villages and very underdeveloped villages.

In the village potential data, each attribute/indicator has a value of 0 to 5, where a value of 0 is the lowest value while a value of 5 is the highest value. From each cluster obtained has a centroid value, where the centroid is the "midpoint" of the cluster. So to determine the status of the village, we can calculate the number of centroids for each cluster, which can be written with the equation:

$$\text{Village status} = \sum [CI]_{-1}, [CI]_{-2}, \dots, [CI]_{-42} \dots \dots \dots 1$$

From equation 1, CI is the centroid of each indicator and each cluster has 42 indicators. Determination of village status will be sorted based on the sum of the centroid values of each indicator in each cluster, where the lowest sum value will be initialized as very underdeveloped village status and the highest sum value will be initialized as independent village status.

2.4 Validation

In this study, validation will be carried out to test the distance calculation method on which k-mean algorithm is most effectively used for grouping village potential data. The test will be carried out using the Rapid Miner tool to obtain the accumulated time and the value of the Bouldin Index for each distance calculation method used. The best time efficiency is the one that has the minimum length of time. Meanwhile, by using the Davies Bouldin Index, a cluster will be considered to have an optimal clustering scheme which has a minimal Davies Bouldin Index.

3. RESULTS AND DISCUSSION

3.1. Test the proposed

Method Testing the Manhattan, Euclidean and Chebyshev distance calculation methods on the k-means algorithm used to group the 2014 Village Potential data in Aceh Province will be carried out using the latest model clustering that is validated using execution time and the Davies Bouldin Index.

3.1.1. Manhattan Distance

From the use of the k-means algorithm with the Manhattan Calculation method the distance to group the 2014 Podes data in Aceh Province which amounted to 6,474 villages, the number of villages from each obtained cluster was as follows:

- Cluster 0: 643 villages
- Cluster 1: 996 villages
- Cluster 2: 960 villages
- Cluster 3: 1055 villages
- Cluster 4: 2820 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-means grouping using the Manhattan Calculation method distance as shown in Table 4.

Table 4. Status and Number of Villages Using Manhattan

Clusters	Number of Centroids	Village Status	Number of Villages
Clusters 0	3.42	Independent	643 villages
Clusters 1	2.92	Up	996 villages
Clusters 2	2.22	Very Underdeveloped	960 villages
Clusters 3	2.87	Develop	1,055 villages
Clusters 4	2.66	Left behind	2,820 villages

3.1.2. Euclidean Distance

From the use of the k-means algorithm with the Euclidean Calculation method the distance to group the 2014 Podes data in Aceh Province which amounted to 6,474 villages, the number of villages from each obtained cluster was as follows:

- Cluster 0: 1,145 villages
- Cluster 1: 1.8 56 villages
- Cluster 2: 1,275 villages
- Cluster 3: 781 villages
- Cluster 4: 1,417 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-means grouping using the Euclidean Calculation method distance as shown in Table 5.

Table 5. Status and Number of Villages Using Euclidean

Clusters	Number of Centroids	Village Status	Number of Villages
Clusters 0	3.28	Independent	1,145 villages
Clusters 1	2.77	Develop	1,856 villages
Clusters 2	2.27	Very Underdeveloped	1,275 villages
Clusters 3	2.86	Up	781 villages
Clusters 4	2.64	Left behind	1,417 villages

3.1.3. Chebyshev Distance

From the use of the k-means algorithm with the Chebyshev Calculation method the distance to group the 2014 Podes data in Aceh Province, amounting to 6,474 villages, the number of villages from each obtained cluster is as follows:

- Cluster 0: 2,639 villages
- Cluster 1: 1,188 villages
- Cluster 2: 1,182 villages
- Cluster 3: 1,266 villages
- Cluster 4: 199 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-means grouping using the Chebyshev Calculation method distance as shown in Table 6.

Table 6. Status and Number of Villages Using Chebyshev

Clusters	Number of Centroids	Village Status	Number of Villages
Clusters 0	2.73	Develop	2,639 villages
Clusters 1	3.17	Independent	1,188 villages
Clusters 2	2.41	Very Underdeveloped	1,182 villages
Clusters 3	2.74	Up	1,266 villages
Clusters 4	2.49	Left behind	199 villages

3.2. Test the proposed

Accumulation of time is carried out by executing 5 times for each distance calculation method used. The 5 executions will then be averaged to obtain the most efficient execution time for each distance calculation method. From the tests that have been carried out, it is obtained that the length of execution time is different, as for the length of execution time from the Manhattan, Euclidean and Chebyshev distance calculation methods that have been carried out, it can be seen in Figure 1.

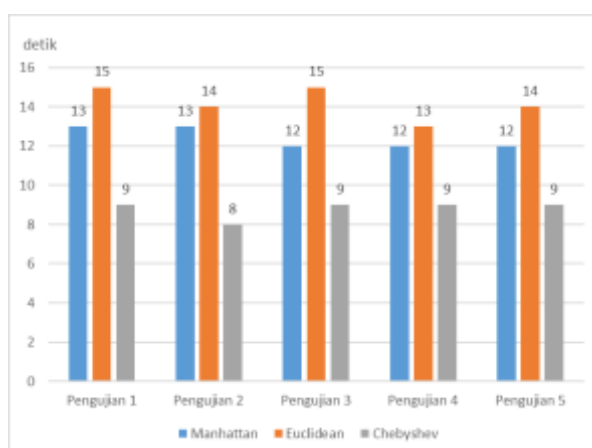


Figure 1. Execution Time

In Figure 1. It can be seen that the execution time of the Manhattan method distance to test 1 to test 5 in a row is 13 seconds, 13 seconds, 12 seconds, 12 seconds and 12 seconds, so when taken on average execution time of a Manhattan distance is 12.4 seconds. Meanwhile, the execution time of the Euclidean Method for testing 1 to 5, respectively, is 15 seconds, 14 seconds, 15 seconds, 13 seconds and 14 seconds, so that the average execution time of the Euclidean distance is 14.2 seconds. Then the execution time of the Chebyshev Method for testing 1 to 5, respectively, namely 9 seconds, 8 seconds, 9 seconds, 9 seconds and 9 seconds, so that when taken the average execution time of Chebyshev distance is 8.8 seconds. The more easily the execution time required for the Manhattan, Euclidean and Chebyshev methods can be seen in Table 7.

Table 7. Old Time Execution

testing	Execution Time		
	Manhattan	Euclidean	Chebyshev
1	13 seconds	15 seconds	9 seconds
2	13 seconds	14 seconds	8 seconds
3	12 seconds	15 seconds	9 seconds
4	12 seconds	13 seconds	9 seconds
5	12 seconds	14 seconds	9 seconds
Average	12.4 seconds	14.2 seconds	8.8 seconds

3.3. Testing the Davies Bouldin Index

In this study, the Davies Bouldin Index (DBI) was used to validate the data in each cluster. Measurement using DBI aims to maximize the distance between clusters. By using DBI A cluster will be considered to have an optimal clustering scheme if it has a minimum Davies Index. As for the tests that have been carried out, the values obtained Davies Index from the Manhattan, Euclidean and Chebyshev methods are shown in Figure 2.

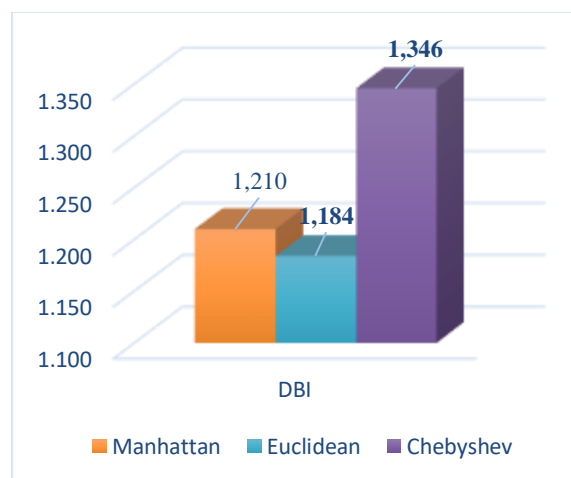


Figure 2. Davies Index of Manhattan, Euclidean and Chebyshev methods

From Figure 2, it can be seen that the value The *Davies Index* from the Manhattan method is 1.210, the value The *Davies Index* from the Euclidean method is 1,184 and the value The *Davies Index* from the Chebyshev method is 1,346. As for the easier value of the *Davies Index* from the Manhattan, Euclidean and Chebyshev methods, it can be seen in Table 8.

Table 8. Status and Number of Villages Using Chebyshev

Davies Bouldin Index		
Manhattan	Euclidean	Chebyshev
1,210	1,184	1,346

From Table 8 it can be seen that the most optimal value of the Manhattan, Euclidean and Chebyshev methods is the Euclidean Method distance with the Davies Index value of 1.184.

3.4. Analysis of Test Results

From testing the 2014 Village Potential data grouping method in Aceh Province using the k-means algorithm with the distance calculation methods Manhattan, Euclidean and Chebyshev that have been carried out, the results are:

1. The test model used can run well and show the results in the form of centroid values for each cluster from the methods Manhattan, Euclidean and Chebyshev, so that the status of the village can be determined from the number of centroids in each clusters.
2. The use of the distance calculation method used affects the amount of data in each cluster.
3. The lengthier time obtained from the tests that have been carried out shows that the distance calculation method of Chebyshev has the most efficient execution time with an average accumulation time of 8.8 seconds.
4. By using the test, the Davies Bouldin Index shows that the distance calculation method Euclidean has the Davies Index most optimal value with a value of 1.184.

From the tests that have been carried out, it can be seen that the 2014 Village Potential data grouping in Aceh Province using the k-means algorithm with the distance calculation method Chebyshev has the most efficient accumulation of time compared to Manhattan and Euclidean, while the The Euclidean method has the most optimal Davies Index value compared to the method. Manhattan and Chebyshev. So when viewed from the quality of the cluster based on the Davies Index, the cluster status of the village is obtained from the k-means algorithm with the distance calculation method Euclidean as follows:

- Cluster Very Disadvantaged Village As many as 1,275 villages
- Disadvantaged Village Clusters As many as 1,417 villages
- Cluster Developing Village Of 1,856 villages
- Advanced Village Cluster as many as 781 villages
- Independent Village cluster of 1,145 villages

4. CONCLUSION

From the discussion and evaluation in the previous chapters, the 2014 Village Potential data grouped into 5 groups using the k-means algorithm with the Manhattan, Euclidean and Chebyshev distance calculation methods, the conclusions are:

1. The 2014 Village Potential data in Aceh Province has been grouped into 5 village statuses by obtaining the number of villages for each cluster, namely cluster as many as 1,275 villages,

- cluster as many as 1,417 villages, cluster as many as 1,856 villages, cluster as many as 781 villages and clusters as many as 1,145 villages.
2. The 2014 Village Potential data grouping into 5 village statuses using the k-means algorithm with the Chebyshev distance calculation method has the most efficient accumulation of time compared to Manhattan and Euclidean, while the Euclidean method has the most optimal Davies Index

REFERENCES

- [1] Al-Roby, MF, & El-Halees, AM (2013). Classifying Multi-Class Imbalance Data. 37(5), 74–81.
- [2] Amandeep Kaur Mann, NK (2013). Review Paper on Clustering Techniques. Global Journal of Computer Science and Technology.
- [3] Awasthi, R., Tiwari, AK, & Pathak, S. (2013). Empirical Evaluation On K Means Clustering With Effect Of Distance Functions For Bank Dataset. International Journal of Innovative Technology and Research, 1(3), 233–235.
- [4] Mishra, BK, Rath, A., Nayak, NR, & Swain, S. (2012). Far efficient K-means clustering algorithm. ACM International Conference Proceeding Series. <https://doi.org/10.1145/2345396.2345414>
- [5] Chakraborty, S., Nagwani, NK, & Dey, L. (2011). Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms. International Journal of Computer Applications. <https://doi.org/10.5120/3346-4611>
- [6] Chaudhari, B., & Parikh, M. (2012). A Comparative Study of Clustering Algorithms using Weka Tools. International Journal of Application or Innovation in Engineering and Management (IJAIEM).
- [7] Claypo, N., & Jaiyen, S. (2015). Opinion mining for Thai restaurant reviews using K-Means clustering and MRF feature selection. 2015 7th International Conference on Knowledge and Smart Technology (KST), 105–108. <https://doi.org/10.1109/KST.2015.7051469>
- [8] Deepa, VK, Remy, J., & Geetha, R. (2013). Rapid development of applications in data mining. 2013 International Conference on Green High Performance Computing, ICGHPC 2013. <https://doi.org/10.1109/ICGHPC.2013.6533916>
- [9] Ding, S., Wu, F., Qian, J., Jia, H., & Jin, F. (2015). Research on data stream clustering algorithms. Artificial Intelligence Reviews. <https://doi.org/10.1007/s10462-013-9398-7>
- [10] Directorate General of Public Administration, KDN (nd). <https://www.bps.go.id/statictable/2014/09/05/1366/wide-area-dan-number-islands-menurut-provinsi-2002-2016.html>.
- [11] Gandhi, G., & Srivastava, R. (2014). Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms. International Journal of Computer Applications, 87(9), 10–13. <https://doi.org/10.5120/15235-3770>
- [12] Ghosh, S., & Kumar, S. (2013). Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. International Journal of Advanced Computer Science and Applications. <https://doi.org/10.14569/ijacsa.2013.040406>
- [13] Grabusts, P. (2011). The choice of metrics for clustering algorithms. Video. Technology. Resurrection - Environment, Technology, Resources. <https://doi.org/10.17770/etr2011vol2.973>

- [14] Hope, FR (2013). The Impact of Urbanization on City Development in Indonesia. *Society*, 1(1), 35–45. <https://doi.org/10.33019/society.v1i1.40>
- [15] Kouser, K., & Sunita, S. (2013). A comparative study of K Means Algorithm by Different Distance Measures. *International Journal of Innovative Research in Computer and Communication Engineering*.
- [16] KumarSagar, H., & Sharma, V. (2014). Error Evaluation on K-Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset. *International Journal of Computer Applications*. <https://doi.org/10.5120/15066-3429>
- [17] Pratap, S., Kushwah, S., Rawat, K., & Gupta, P. (2012). Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining. 3, 109–113.
- [18] Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*. <https://doi.org/10.5120/11430-6785>
- [19] Soleh, A. (2017). Village Potential Development Strategy. *Sungkai Journal*, 5(1), 35–52.
- [20] Verma, M., Srivastava, M., Chack, N., Diswar, AK, & Gupta, N. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Engineering Research and Applications* www.ljera.com.
- [21] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2014.2362522>
- [22] Yadav, J., & Sharma, M. (2013). A Review of K-mean Algorithms. *International Journal of Engineering Trends and Technology*.
- [23] GT Pranoto, W Hadikristanto, Y Religia (2022). "Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm" *JISA (Journal of Informatics and Science)* 5 (1): 28-34.