1st Pelita International Conference
Volume 01 No 01 September 2023
E-ISSN: 3026-4235

https://jurnal.pelitabangsa.ac.id/index.php/pic

Comparison of K-Means and K-Medoid Algorithms in Classifying Village Status (Case Study: Gorontalo Province)

Aswan Supriyadi Sunge, Nanang Tedi Kurniadi, Edy Widodo
Informatics Engineering, Universitas Pelita Bangsa, Indonesia
Email: aswan.sunge@pelitabangsa.ac.id

Abstract
In the national development process, the village occupies a very important position. This is because it is the smallest government structure and has direct contact with the community. Seeing the importance of its role in national development, one of which is Gorontalo Province, based on directions from the central government, is trying to implement the Village Fund Allocation (ADD) policy for all villages in Gorontalo Province. In distributing the allocation of funds, it is necessary to map the status of the Village to find out the amount that must be given. This test uses the average execution time and the Davies Bouldin Index (DBI). After testing it is known that the K-Medoid Algorithm has a better DBI value than the K-Means Algorithm with the DBI value of the K-Medoid Algorithm being 0.050. On the other hand, the K-Means Algorithm has a better average execution time than the K-Medoid Algorithm, where the average execution time is 1 second.

Keywords: Village Status, Data Mining, Clustering, K-Means, K-Medoid

INTRODUCTION

The village is a legal community unit and has a role, function, and contribution to occupy a strategic position as well as the authority to regulate the community [1] It also has authority, namely in village development, and financial management to organize good governance [2]. Therefore, to see the right size in assessing whether a nation is prosperous, just, and dignified or then the village is the district that becomes the benchmark. The Unitary State of the Republic of Indonesia consists of 34 Provinces divided into regencies, cities, and villages [3] and one of them is Gorontalo Province consisting of 5 (five) regencies and 1 (one) city, namely Boalemo Regency, Gorontalo Regency, Pohuwato Regency, Bone Regency Bolango, North Gorontalo Regency, and Gorontalo City. Pohuwato Regency, which borders Central Sulawesi Province, is the largest area in Gorontalo province, while Gorontalo City, which is the provincial capital, has the smallest area [4]. The total area of Gorontalo Province is 12,435.00 km2. Pohuwato Regency, which is the westernmost region of Gorontalo province, is the largest area with an area of 4,455.60 km2. The area that has the smallest area is Gorontalo City with an area of 65.96 km2 or only 0.53% of the total area percentage of Gorontalo province [5]. One of the indicators in advancing an area is that village funds are needed and improving the welfare of the community, especially since the area of Gorontalo province is so large. Since its launch in 2015 until 2021, in 5 (five) regencies throughout Gorontalo Province, more than IDR 3.5 trillion has been collected. The village funds have been used for various activities to improve the quality of life and the economy of rural communities, both through infrastructure development programs and empowering village communities, and village funds have been able to increase the Village Development Index (IDM) in Gorontalo. This year there are six independent villages, 172 villages with advanced status, and 417 developing villages. The remaining 61 villages are underdeveloped, and one village is very underdeveloped [6].

One of them is the development of the village and improving the economy, one of which is the provision of village funds, with that the village becomes innovative [7], however, the

provision of village funds requires regional mapping at the geographical level [8] which is included in the village category which will have an impact on differentiating the number of funds provided. It is hoped that the village map will not only be village funds but can be used as a digital mapping for the potential development of cities and regions [9]. One of them is in selecting village categories with Data Mining which is an effective method used to view large data by looking at different perspectives [10], also extracting previously unknown predictive information and hidden patterns from data available in the database [11]. The increasing rate of heterogeneous data requires intelligent techniques and tools so that we extract useful knowledge from heterogeneous data [10]. However, effective data mining in data management requires an algorithm that can apply predictive results from business, trade, and security from various fields [11].

LITERATURE REVIEW

Data mining is past data or hidden data stored in data [12] then sorting out very large data than looking at patterns and relationships in solving problems or making predictions in making decisions. [13] from data processing is divided into supervised and unsupervised data, this is related to the method used in classification, clustering, or regression [14]. However, in the use of data mining techniques used, one of them is labeling or class to facilitate prediction results or accuracy in each case [15][16]. One of them is in the data mining method, namely clustering, which is the division into the same group that is more identical to one another than in other groups [17]. On the other hand, the main aim of clustering is to divide the items into uniform and distinct groups for output, and mostly, such methods are designed to handle only numeric data and display results such as hierarchical, center-based partitioning, density, and graph-based clustering [18]. The algorithm is one part of clustering which combines some data into several clusters, many studies look at the cluster section from looking at the cluster section, grouping to comparison is also used to look at clusters from existing groups and excel in large data [19][20] [21]. In addition, K-Means is faster and more accurate than other algorithms [22][23].

K-Medoids is an update of the K-Means Outliner strength in clustering and represents a cluster based on proximity and non-distance [24]. K-Medoids can be summarized in the flow using pseudocode [25] in which the essence of each cluster is represented using the average value of each cluster, as below.

Algorithm *K*-medoids clustering algorithm

Require: K, number of clusters; D, a data set of N points
Ensure: A set of K clusters
1: Arbitarily choose K points in D as initial representative points.
2: **repeat**
3: **for** each non-representative point p in D **do**
4: find the nearest representative point and assign p to the corresponding cluster.
5: **end for**
6: randomly select a non-representative point p_{rand};
7: compute the overall cost C of swapping a representative point p_i with p_{rand};
8: **if** $C < 0$ **then**
9: swap p_j with p_{rand} to form a new set of K representative points.
10: **end if**
11: **until** stop-iteration criteria satisfied
12: **return** clustering result.

Figure 1. K-Medoids Step

In K-means there is a collection of points, therefore a Euclidean Distance algorithm method is needed which is to calculate the squared distance matrix between points, the essence of which is to design an algorithm to complete and delete distance data [26] which is useful for calculating distances in one dimension giving results equal to the Pythagorean formula [27] with the formula below [19]:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk}2)}$$

Where:
dij: Distance between Object i and Object j
xik: Object value i in variable to k
Xjk: Object value j in variable to k
p: Many j variables are observed.

In K-means each separate cluster, a validation method is needed, so the Davies Bouldin Index is needed which functions to separate clusters and is a clustering result method that aims to minimize intra-cluster distances and maximize between clusters [28] and use the formula below [29]:

$$R_i = \underset{j=1\ldots k.i\neq}{max} R_{ij}$$

$$var(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$\underset{i\neq j}{R_{ij}} = \frac{var(C_i)+var(C_j)}{\|c_i - c_j\|}$$

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i$$

Where:
R : Spacing between clusters
Var : variance from data
x : data to i
$\bar{x}$: Average of each cluster
DB : Validasi Davies Bouldin

RESEARCH METHOD

The research is based on quantitative data, where the data used is secondary data obtained from the Central Bureau of Statistics of the Republic of Indonesia in the form of Village Potential Data (Podes) in 2014. The test uses Rapidminer with the Clustering model grouping and the data is in the form of Unsupervised which consists of 12 variables and out of 12 these variables are divided into 42 dependent indicators/attributes without village status labels. Indicators are 0 to 5, where the value 0 is the lowest while the value 5 is the highest. As for the simplicity of all the research steps in this study can be seen in Figure 2.

1st Pelita International Conference
Volume 01 No 01 September 2023
E-ISSN: 3026-4235
https://jurnal.pelitabangsa.ac.id/index.php/pic

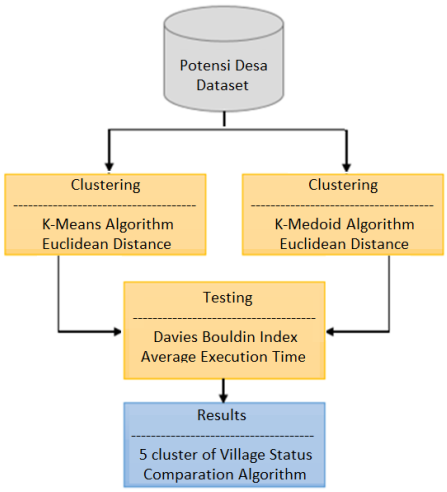

Figure 2. Research Methods

Validation will be carried out to test which algorithm is the most effective for classifying village status in Gorontalo Province. Testing will be carried out to obtain the execution time and Bouldin Index value of each algorithm used. The algorithm that has the best time efficiency is the one that gets the minimum average execution time. While the algorithm will be considered to have an optimal clustering scheme if it obtains a smaller Davies Bouldin Index value. The determination of village status will be carried out using an algorithm that obtains the smallest Davies Bouldin Index value.

RESULTS AND DISCUSSIONS

The Resulting Centroids

The tests that have been carried out, the centroid values and the number of clusters are different for each algorithm used. As mentioned in chapter 3 that in the 2014 Podes 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, so in this study to determine the status of a village in the province Gorontalo is done by calculating the number of centroids for each cluster, which is written by the equation:

$$Village\ Status = \sum CI_1, CI_2, \ldots\ldots, CI_{42}$$

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 Disadvantaged Village status and the highest sum value will be initialized as Independent Village status. The sequence of naming clusters from the lowest to the highest scores are Very Disadvantaged Villages, Disadvantaged Villages, Developing Villages, Advanced Villages, and Independent Villages. The centroid values and the number of clusters from the k-means and k-medoid algorithm tests that have been carried out are as follows:

K-Means Algorithm

Based on the use of the K-Means Algorithm with the Euclidean distance calculation method to classify 2014 Podes data in Gorontalo Province, totaling 657 villages, the number of

villages from each cluster is obtained as follows: Cluster 0 is 136 villages, Cluster 1 is 180 villages, Cluster 2 is 130 villages, Cluster 3 of 85 villages, and Cluster 4 of 126 villages. When viewed from the number of centroids calculated by equation 6 and the number of villages in each cluster, and village status can be obtained from the k-means grouping using the Euclidean distance calculation method shown in table 1.

Table 1. Status and Number of Villages with the K-Means Algorithm

Cluster	Number of Centroids	Village Status	Number of Villages
Cluster 0	101.65	Very Disadvantaged Villages	136
Cluster 1	131.41	Disadvantaged Villages	180
Cluster 2	120.08	Developing Villages	130
Cluster 3	118.60	Advanced Villages	85
Cluster 4	138.07	Independent Villages	125

K-Medoid Algorithm

Based on the use of the K-Medoid Algorithm with the Euclidean distance calculation method to classify 2014 Podes data in Gorontalo Province, totaling 657 villages, the number of villages from each cluster is obtained as follows: Cluster 0 is 315 villages, Cluster 1 is 141 villages, Cluster 2 is 34 villages, Cluster 3 of 90 villages, and Cluster 4 of 77 villages. When viewed from the number of centroids calculated using equation 6 and the number of villages in each cluster, village status can be obtained from the K-Medoid grouping using the Euclidean distance calculation method shown in table 2.

Table 2. Status and Number of Villages with the Euclidean Algorithm

Cluster	Number of Centroids	Village Status	Number of Villages
Cluster 0	134	Very Disadvantaged Villages	315
Cluster 1	125	Disadvantaged Villages	141
Cluster 2	96	Developing Villages	34
Cluster 3	94	Advanced Villages	90
Cluster 4	104	Independent Villages	77

Execution Time Testing

Time accumulation is done by executing 5 times for each algorithm used. Based on the 5 executions, will then be averaged to obtain the most efficient execution time for each algorithm. From the tests that have been carried out, different execution times have been obtained, while the execution time of the k-means and k-medoid algorithm tests that have been carried out can be seen in figure 3.

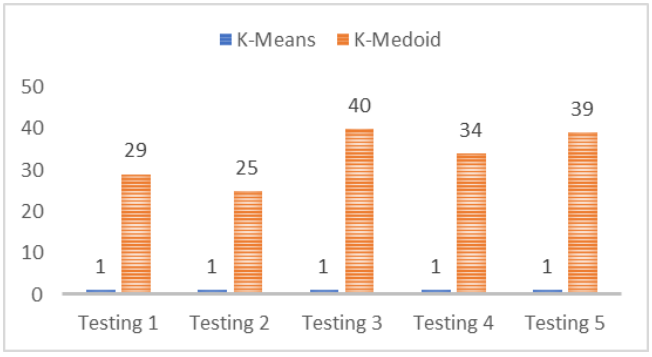

Figure 3. Length of Execution Time

In Figure 3, the execution time of the K-Means Algorithm for the first test to the 5th test in a row is only 1 second, so if the average execution time of the K-Means Algorithm is taken, it is 1 second. While the execution time of K-Medoid for the first test to test 5 in a row is 29 seconds, 25 seconds, 40 seconds, 34 seconds, and 39 seconds, so if taken the average execution time of the K-Medoid Algorithm is 33.4 seconds. The fastest execution time of the K-Means and K-Medoid Algorithms is using the K-Means Algorithm with an average execution time of 1 second.

Execution Time Testing

In this study, the Davies Bouldin Index (DBI) was used to validate data for each cluster. Measurements using DBI aim to maximize the inter-cluster distance, using DBI, a cluster will be considered to have an optimal clustering scheme if it has a minimal Davies Index. As for the tests that have been carried out, the Index Davies value from the K-Means Algorithm is 0.067, while the Index Davies value from the K-Medoid Algorithm is 0.05. Based on the results obtained, the most optimal Davies index value from the K-Medoid Algorithm is the Davies Index value of 0.05.

Analysis of Test Results

Based on the Podes data grouping test in 2014 in Gorontalo Province using the K-Means and K-Medoid Algorithms that have been carried out, the results are as follows:

a. The testing model used works well and shows the results in the form of centroid values for each cluster from the K-Means and K-Medoid Algorithms so that the status of villages in Gorontalo Province can be determined from the number of centroids in each cluster.
b. The use of the K-Means and K-Medoid algorithms affects the amount of data in each cluster.
c. The average time obtained from the tests that have been carried out shows that the K-Means Algorithm has the most efficient execution time with an average execution time of 1 second.
d. Using the Davies Bouldin Index test shows that the K-Medoid Algorithm has a more optimal clustering scheme with a DBI value of 0.05.

CONCLUSION

Based on the discussion and evaluation in previous chapters, the grouping of Podes in 2014 in Gorontalo Province into 5 groups using the K-Means and K-Medoid Algorithms showed that grouping into 5 village statuses using the K-Means Algorithm had a more efficient execution time than K-Medoid Algorithm. However, even though the execution time is longer, the K-Medoid Algorithm has a more optimal clustering scheme than the K-Means Algorithm. After grouping them into 5 village statuses, a mapping of village status was obtained, namely 90 very underdeveloped villages, 34 underdeveloped villages, 77 developing villages, 141 developed villages, and 315 independent villages. It is hoped that the use of the k-means and k-medoid algorithms can be used in future research to classify village status in Gorontalo Province, as well as a model for evaluating village fund allocations in the Province of the Republic of Indonesia.

References

Law of the Republic of Indonesia No. 32 of 2004 concerning Regional Government Article 1 paragraph 12.
Law of the Republic of Indonesia Number 6 of 2014 concerning Villages Article 18.
The 1945 Constitution of the Republic of Indonesia Article 18B Paragraph 1.
Central Bureau of Statistics for the Province of Gorontalo, Republic of Indonesia, 2018.
Central Bureau of Statistics for the Province of Gorontalo, Republic of Indonesia, 2015.
Village Fund Contributes to Human Resources Development in Gorontalo Province"

https://gorontaloprov.go.id/dana-desa-beri-sumbangsih-peningkatan-idm-gorontalo, july 26, 2021. [Online]

Wahudin, Kessa. Village Development Planning. Ministry of Villages, Development of Disadvantaged Regions, and Transmigration of the Republic of Indonesia. 2015.

H. Setiono, E. Mulyanto, and S. M. S. Nugroho, "Village Classification based on Geographic Difficulties using Backpropagation Neural Network Algorithm (Case Study: Village Potential of Sumenep Regency)," 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2019, pp. 399-403, doi: 10.1109/ISITIA.2019.8937082.

G. Yang, F. Duan, W. Zhao, W. Zhao, and L. Zhang, "Building extraction in towns and villages based on Digital Aerial Image by texture enhancing," 2010 18th International Conference on Geoinformatics, 2010, pp. 1-6, doi: 10.1109/GEOINFORMATICS.2010.5567636.

S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," 2013 International Conference on Machine Intelligence and Research Advancement, 2013, pp. 203-207, doi: 10.1109/ICMIRA.2013.45.

H. A. Madni, Z. Anwar, and M. A. Shah, "Data mining techniques and applications — A decade review," 2017 23rd International Conference on Automation and Computing (ICAC), 2017, pp. 1-7, doi: 10.23919/IConAC.2017.8082090.

B. N. Lakshmi and G. H. Raghunandhan, "A conceptual overview of data mining," 2011 National Conference on Innovations in Emerging Technology, 2011, pp. 27-32, doi: 10.1109/NCOIET.2011.5738828.

S. P. Latha and N. Ramaraj, "Algorithm for Efficient Data Mining," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), 2007, pp. 66-70, doi: 10.1109/ICCIMA.2007.150.

Li, R. Xiao, J. Feng, and L. Zhao, "A semi-supervised learning approach for detection ofphishing webpages," Elsevier, vol. 124, no. 23, pp. 6027– 6033, 2013.

D. T. Larose, Discovering Knowledge in Data: An Introduction toData Mining. John Willey & Sons, Inc, 2015.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Berry, M., Mohamed, A., Yap, B. (eds) Supervised and Unsupervised Learning for Data Science. Unsupervised and Semi-Supervised Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-22475-2_1.

W. Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection," IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, pp. 72–78, 2017.

J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," 2019 19th International Conference on Computational Science and Its Applications (ICCSA), 2019, pp. 71-81, doi: 10.1109/ICCSA.2019.000-1.

Y. Religia and A. S. Sunge, "Comparison of Distance Methods in K-Means Algorithm for Determining Village Status in Bekasi District," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 270-276, doi: 10.1109/ICAIIT.2019.8834604.

C. Shah and A. Jivani, "Comparison of Data Mining Clustering Algorithms," in Nirma University International Conference on Engineering, 2013.

B. Chaudhari and M. Parikh, "A Comparative Study of clustering algorithms Using weka tools," International Journal of Application or Innovation in Engineering & Management, vol. 1, no. 2, pp. 154-158, 2012. C. Shah and A. Jivani, "Comparison of Data Mining Clustering Algorithms," in Nirma University International Conference on Engineering, 2013.

M. Verma, M. Srivastava, N. Chack, A. K. Diswar and N. Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering

Research and Applications (IJERA), vol. 2, no. 3, pp. 1379-1384, 2012.

S. Ghosh and S. K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," International Journal of Advanced Computer Science and Applications, vol. 4, no. 4, pp. 35-39, 2013.

T. Velmurugan and T. Santhanam, "Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points," J. Comput. Sci., vol. 6, no. 3, pp. 363–368, 2010.

J. E. Gentle, L. Kaufman, and P. J. Rousseuw, "Finding Groups in Data: An Introduction to Cluster Analysis.," Biometrics, vol. 47, no. 2, p. 788, 1991.

I. Dokmanic, R. Parhizkar, J. Ranieri and M. Vetterli, "Euclidean Distance Matrices: Essential theory, algorithms, and applications," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 12-30, Nov. 2015, doi: 10.1109/MSP.2015.2398954.

H. K. Sagar and V. Sharma, "Error Evaluation on K- Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset," International Journal of Computer Applications, vol. 86, no. 18, pp. 1-5, 2014. Science and Information Technologies, vol. 5, no. 2, pp. 2501-2506, 2014.

Thomas, J.C., Cofre, M.M., & Santos, M.J. (2014). New Version of Davies-Bouldin index for clustering validation based on hyper rectangles.

A. S. Sunge, Y. Heryadi, Y. Religia and Lukas, "Comparison of Distance Function to Performance of K-Medoids Algorithm for Clustering," 2020 International Conference on Smart Technology and Applications (ICoSTA), 2020, pp. 1-6, doi: 10.1109/ICoSTA48221.2020.1570615793