Clustering Villages Based on Distance and Accessibility to Health Facilities Using the K-Means Method

Noviandi¹, Stefanny Amalia Noviantika², Bambang Irawan³
¹,²,³Department of Informatics Engineering, Esa Unggul University, Indonesia

ABSTRACT

There are 47 very underdeveloped and 63 underdeveloped villages in Melawi regency. More than 50% of the villages have no health facilities, and the percentage of road lengths with good condition is only 20.53% in Melawi County. One of the most important factors influencing health problems is the physical aspect such as the availability of health facilities. In addition, the distance and easy access to health facilities also influence how quickly people are treated and vaccinated during the Covid 19 pandemic. The objective of this study is to determine the degree of accessibility of health facilities in villages by forming village clusters that are likely to be important to the government in ensuring treatment and distribution of Covid 19 vaccine. The clustering method used is the K-Means method with Euclidean spacing to calculate the spacing of the data and the Elbow method to determine the optimal number of clusters on the data, and the Silhouette coefficient evaluation method to test the degree of accuracy of the model created with K-Means. The results of the Elbow method showed the optimal number of clusters to be 2 clusters. Based on the results of the K-Means algorithm process, the clusters that have a larger average distance and access is rated as difficult are cluster 1 with 92 villages in it, and cluster 1 has a smaller average distance and access is relatively easy with 77 villages in it. The result of the evaluation with the silhouette coefficient is 0.299.

Keywords:
Clustering
K-Means
Euclidean Distance
Elbow Method
Silhouette Coefficient

This is an open access article under the CC BY-SA license.

Corresponding Author:
Noviandi
Department of Informatics Engineering
Esa Unggul University
Jakarta, Indonesia
Email: noviandi@esaunggul.ac.id
© The Author(s) 2021

1. Introduction

Social health problems, especially in developing countries such as Indonesia, are influenced by two factors, namely physical factors and non-physical factors. Physical aspects such as health facilities and disease treatment, the second is non-physical aspects related to health problems[1]. Distance and ease of access to health facilities are also important things that must be considered, especially during the Covid-19 pandemic, because it affects how quickly people get treatment and vaccinations[2]. Melawi Regency, West Kalimantan has 73 health facilities from 169 villages, and has good road conditions of 20.53%[3]. Therefore, it is important to know the level of coverage of village health facilities by forming clusters. There are some
studies that use cluster algorithms, such as Fuzzy C-Means [4], and K-Means [5]. Fuzzy C-Means algorithm has a faster and easier process time to interpret [6], however, it has weaknesses in the calculation process and fuzzy iterations that use longer time than the K-Means algorithm [7]. The K-Means algorithm is widely applied to research because it is more efficient in categorizing data with very large amounts, but this algorithm is not quite right in random selection of centroid starting points and determining the initial number of clusters[8].

The K-Means algorithm has a higher consistent rate and stands out than fuzzy C-Means, but when executed with different iterations Fuzzy C-Means stands out more than the K-Means algorithm. Based on these problems, researchers tested the level of accuracy of the model produced by the K-Means algorithm using the Silhouette Coefficient method, applied the Elbow method to determine the best number of clusters, and the Euclidean Distance method to determine the distance of the data to the initial centroid point.

2. Research Methodology

Cross Standard Industry Processing for Data Mining (CRISP-DM) data mining methodology used in this study [9]. CRISP-DM has data mining standards as a commonly used solution in research and business [10]. This methodology consists of six steps, namely; Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

1. Data Understanding
   At this step, village development data was collected from the Central Statistics Agency (BPS) of Melawi Regency. Village development data consists of 169 data records, with 46 variables. The criteria for the data variables used are the name of the village, the name of health facilities, the number of health facilities, the distance to reach the nearest health facility, and the ease of access to reach the nearest health facility.

2. Data Preparation
   This step is necessary to build the raw data into the final dataset used at the model creation stage. Data preprocessing is carried out by three methods, namely:
   a. Data Cleaning
      Data cleaning is applied to remove noise, inconsistent, outliers, and missing values [11]. Data on village development in Melawi Regency is still missing, so this study uses the average method by calculating the average amount of data in the variable.
   b. Data Transformation
      Datayang digunakan di menjadi formattransformasi .csv.
   c. Data Reduction
      Data reduction is applied to reduce the volume of the dataset while maintaining data integrity [12]. The method used in this study is Feature Selection [13] to select the variables used in the modelling step. The variables used are the name of the village, the name of health facilities, the number of health facilities, the distance to reach the nearest health facility, and the ease of access to reach the nearest health facility.

3. Modelling
   The techniques applied at this step have special conditions on the form of data, making it possible to return to the data preparation step. The tool used in this study is Jupyter Notebook with Python program language. Libraries used Scikit-lean and Matplotlib. The step of creating a cluster model with the K-Means algorithm is[14]:
   a. Determining the optimal number of k with the Elbow method [15]. The sum of k is determined based on the sum of square error values using equation (1) [16]. The number of k is selected by the largest margin of descent and forms an elbow on the chart, and then it is determined the initial centroid of each k.

   \[
   SSE = \sum_{k=1}^{\hat{k}} \sum_{x_i \in s_k} \|x_i - C_j\|_2^2
   \] (1)

   b. Calculating the distance from each data to the centroid cluster using the Euclidean Distance method [17] with equation 2.

   \[
   d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}
   \] (2)
c. The data with the shortest Euclidean distance will be grouped into one cluster.

d. Calculations are performed to obtain a new centroid value in the next iteration, by calculating the average distance of each data in the cluster.

e. The 2nd to 4th stages will be repeated in each iteration until the centroid value no longer changes.

4. Evaluation
At this step the model is evaluated to ascertain whether it meets the objectives. The method used is Silhouette Coefficient [18] to evaluate the cluster so that it can be known how well the cluster is formed. The resulting value can determine how good the cluster structure is, where if the value <= 0.25 is no-structure, value > 0.25 and <= 0.50 is weak structure, value > 0.50 and <= 0.7 is medium structure, value > 0.7 and <=1 is strong structure [19].

3. Result and Discussion
3.1. Data Preprocessing
Village development data of Melawi Regency consists of 169 records with 46 variables. The variables used are the name of the village, the name of health facilities, the number of health facilities, the distance to reach the nearest health facility, and the ease of access to reach the nearest health facility. The Average approach is taken to clean up the missing values. Variable selection using feature selection. The result of selecting variables using the tools jupyter notebook in Figure 4.1

![Figure 4.1 Variable Selection with a Jupyter Notebook](image)

Selection of variables by conducting an interview with the Melawi District Health Office. The next step is to change the 11 selected variable names with the aim of making it easier to identify variable names.

![Figure 4.2 Variable Name Initiation](image)
Keterangan

distance_A : Distance to reach the nearest hospital
access_A : Easy access to the nearest hospital

distance_B : Distance to reach the health center with the nearest hospitalization
access_B : Easy access to reach the health center with the nearest hospitalization

distance_C : The distance to reach the nearest health center without hospitalization
access_C : Easy access to reach the nearest health center without hospitalization

distance_D : Distance to reach the nearest auxiliary health center
access_D : Easy access to the nearest auxiliary health center

distance_E : Distance to the nearest pharmacy
access_E : Easy access to the nearest pharmacy

3.2. Determination of the Optimal Number of k with the Elbow Method

The optimal number of k in this study used the Elbow method, because k-Means has a weakness in determining the number of initial clusters determined randomly [8]. The best number of k for clusters 1 to 10 using the Elbow Method is k=2. The highest Niali Sum Square Error (SSE) between values is used as the number of clusters (Table 4.1).

| Cluster | SSE          | Difference |
|---------|--------------|------------|
| 1       | 554115.394   | 554115.3938 |
| 2       | 390819.122   | 163296.2721 |
| 3       | 296004.316   | 94814.80568 |
| 4       | 236341.186   | 59663.13106 |
| 5       | 205719.449   | 30621.73703 |
| 6       | 179512.049   | 26207.40034 |
| 7       | 160301.045   | 19211.00314 |
| 8       | 152049.829   | 8251.216232 |
| 9       | 138682.915   | 13366.91455 |
| 10      | 128902.45    | 9780.464125 |

Figure 4.3 shows the optimal jumlak, where k=2 experienced a decrease of at most 163296.2721, and was used in determining the number of clusters in 11 variables of village development data of Melawi Regency.

![Figure 4.3 Elbow Method Graph in Determining the Number of k](image-url)
The calculation of the SSE value (equation 1) for the resulting model with the value of k=2 in the k-means algorithm is the i-th total data of the formed cluster. The calculation process is carried out up to the 169th data and then the total results of calculating the distance from each data to the centroid.

Example of calculating the value of SSE, k=1 in the 1st data

$\sum (99.9 - 57.6355)^2 + (4 - 2.863905)^2 + (21.5 - 48.225)^2 + (3 - 2.378698)^2 + (21.5 - 28.6213)^2 + (3 - 2.656805)^2 + (21.5 - 16.29527)^2 + (4 - 1.840237)^2 + (63 - 36.94024)^2 + (3 - 2.846154)^2 = 2550.648735$

| Data Distance to Centroid | R104N |
|---------------------------|-------|
| Nanga Tangkit             | 2550.648735 |
| Landau Kabu               | 6894.31909  |
| Penyengkuang              | 9956.138202  |
| Sungai Sampuk             | 7021.37264  |
| Melona                    | 1966.053705  |
| Batu Begigi               | 2061.271693  |
| **Sum of Square Error (SSE)** | **554115.3938** |

A cluster model with the k-Means algorithm is used to determine centroid points. The tools used are jupyter notebooks with Scikit-lean. The iteration process will stop if the centroid does not undergo displacement or change in value.

| Variable | Centroid 1 | Centroid 2 |
|----------|------------|------------|
| Distance_A | 57.9738 | 57.3012 |
| Access_A   | 2.92     | 2.8       |
| Distance_B | 21.3952 | 23.5565 |
| Access_B   | 2.47619 | 2.28355  |
| Distance_C | 29.5976 | 27.6565 |
| Access_C   | 2.70238 | 2.61176  |
| Distance_D | 16.1286 | 16.46     |
| Access_D   | 1.80952 | 1.87059  |
| Distance_E | 36.6369 | 37.24     |
| Access_E   | 2.83333 | 2.85882  |

The cluster results using 10 variables, and 169 record data are distance and ease of access for hospital health facilities, puskesmas with inpatient, puskesmas without hospitalization, auxiliary health centers, and pharmacies. The next step taken after determining the centroid point is to determine the distance of each data to centroid 1 or to centroid 2. The distance determination process is carried out using the Euclidean Distance method (Equation 2).

For the 1st data distance (Nanga Tangkit) to Centroid 1:

$$D(1,1) = \sqrt{\sum (X_i - C_i)^2} = \sqrt{(99.9 - 57.6355)^2 + (4 - 2.863905)^2 + (21.5 - 48.225)^2 + (3 - 2.378698)^2 + (21.5 - 28.6213)^2 + (3 - 2.656805)^2 + (21.5 - 16.29527)^2 + (4 - 1.840237)^2 + (63 - 36.94024)^2 + (3 - 2.846154)^2} = 50.5331$$
1st data distance (Nanga Tangkit) to Centroid 2.

$$D(1,1) = \sqrt{\sum (Xi - Ci)^2}$$

$$= \sqrt{(99.9 - 57.3012)^2 + (4 - 2.8)^2 + (21.5 - 23.5565)^2 + (3 - 2.28235)^2 + (21.5 - 27.6565)^2 + (3 - 2.61176)^2 + (21.5 - 16.46)^2 + (4 - 1.87059)^2 + (63 - 37.24)^2 + (3 - 2.85882)^2}$$

$$= 50.5216$$

The results of village clusters with low and high health facility coverage based on distance and ease of access with analysis of the resulting cluster model. Villages that have a greater average distance to the nearest hospital are in cluster 1 with a value of 83.42 km. The average accessibility is 3, which means that it is classified as difficult. The villages in cluster 2 have an average distance closer to the value of 26.81 km. The average access density is a value of 3, which means that it is classified as difficult.

Villages that have a greater average distance to the nearest health center with hospitalization are in cluster 2 with a value of 23.19 km, and the average access density is a value of 2, which means that it is relatively easy. On the other hand, the villages in cluster 1 have an average distance closer to the value of 21.89 km, with an average accessibility value of 3, which means that it is classified as difficult.

However, the villages that have a greater average distance to the nearest puskesmas without hospitalization are in cluster 1 with a value of 34.15 km, and the average accessibility is a value of 3, which means that it is classified as difficult. Villages in cluster 2 have an average distance of 22.02 km, and the average accessibility is a value of 2, which means that it is relatively easy. Villages that have a greater average distance to the nearest health center are in Cluster 1 with a value of 19.09 km, and the average accessibility is a value of 2, which means that it is relatively easy. Villages in cluster 2 are closer to the value of 12.96 km on average, and the average accessibility is a value of 2, which means that it is relatively easy. Villages that have a greater average distance to the nearest pharmacy are in cluster 1 with a value of 47.15 km.
km, and the average accessibility is a value of 3, which means that it is relatively difficult. On the other hand, the villages in cluster 2 are closer to the value of 24.74 km on average, and the average accessibility is a value of 3, which means that it is classified as difficult.

Based on the description of the resulting pattern, the results of the analysis are that Cluster 1 has a low coverage of health facilities with 92 villages (see Table 4.5), while Cluster 2 has a high coverage of health facilities with 77 villages. It can be concluded that the 92 villages in Cluster 1 listed in Table 4.6 have an average closer distance of 41.14 km, and the average access with a value of 3 which is classified as difficult, while villages in Cluster 2 have an average closer distance of 21.94 km with an average access value of 2 which is relatively easy.

4. Conclusion

In this study, grouping of villages was done using K-Means algorithm based on distance and ease of access to health facilities, which is expected to form village clusters based on distance and ease of access to find out which village clusters have low and high coverage of health facilities, to confirm the quality of the resulting cluster model, the author uses Elbow method to determine the optimal number of clusters. The quality of the resulting model is tested using the silhouette coefficient. It can be concluded that:

1. Based on the pattern picture of cluster 1 and cluster 2, it can be concluded that the level of distance does not affect the ease of access to reach health facilities even after experiencing the clustering process.
2. Based on the average distance and overall access of health facilities in cluster 1 and cluster 2. It can be concluded that in villages that have a low range of health facilities are in cluster 1 with an average longer distance of 41.14 km, and the average access with a value of 3 which is classified as difficult, while villages in cluster 2 have an average closer distance of 21.94 km with an average access value of 2 which is relatively easy.

References

[1] D. R. S. Mayangsari, S. Solikhun, and I. Irawan, “Pengelompokkan Jumlah Desa/Kelurahan Yang Memiliki Sarana Kesehatan Menurut Provinsi Dengan Menggunakan Metode K-Means Cluster,” KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer), vol. 3, no. 1, pp. 370–377, 2019, doi: 10.30865/komik.v3i1.1615.
[2] P. Geldsetzer et al., “Mapping physical access to health care for older adults in sub-Saharan Africa and implications for the COVID-19 response: a cross-sectional analysis,” Lancet Heal. Longev., vol. 1, no. 1, pp. e32–e42, 2020, doi: 10.1016/S2666-7568(20)30010-6.
[3] B. P. S. Kapupaten Melawi, “Kapupaten Melawi Dalam Angka 2021,” BPS Kapupaten Melawi, Melawi, pp. 1–287, 2021.
[4] H. Murfi, N. Rosaline, and N. Hariadi, “Deep autoencoder-based fuzzy c-means for topic detection,” Array, vol. 13, no. August 2021, p. 100124, 2022, doi: 10.1016/j.array.2021.100124.
[5] C. Virmani, A. Pillai, and D. Juneea, “Clustering in aggregated user profiles across multiple social networks,” Int. J. Electr. Comput. Eng., vol. 7, no. 6, pp. 3692–3699, 2017, doi: 10.11591/ijece.v7i6.pp3692-3699.
[6] A. N. Ulfah and S. ‘Uyun, “Analisis Kinerja Algoritma Fuzzy C-Means Dan K-Means Pada Data Kemiskinan,” J. Jatisi, vol. 1, no. 2, pp. 1–15, 2015.
[7] A. K. Dubey, U. Gupta, and S. Jain, “Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 1, pp. 18–29, 2018, doi: 10.18517/ijaset.8.1.3490.
[8] K. R. Nirmal and K. V. V. Satyanarayana, “Issues of K means clustering while migrating to map reduce paradigm with big data: A survey,” Int. J. Electr. Comput. Eng., vol. 6, no. 6, pp. 3047–3051, 2016, doi: 10.11591/ijece.v6i6.11207.
[9] C. Pete et al., “Crisp-Dm 1.0,” Cris. Consort., p. 76, 2000.
[10] K. Rahayu, L. Noviantika, and M. Kusnandar, “Implementation Data Mining with K-Means Algorithm for Clustering Distribution Rabies Case Area in Palembang City,” J. Phys. Conf. Ser., vol. 1500, no. 1, 2020, doi: 10.1088/1742-6596/1500/1/012121.
[11] F. Schäfer, C. Zeiselmair, and J. Becker, “2020 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2020,” 2020 IEEE Int. Conf. Technol. Manag. Oper. Decis. ICTMOD 2020, pp. 190–195, 2020.
[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.
[13] V. Kotu and B. Deshpande, Data Science: Concept and Practice, Second Edi. 50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States: Jonathan Simpson, 2019.
[14] F. Mar’i and A. A. Supianto, “Clustering Credit Card Holder Berdasarkan Pembaruan Tagihan Menggunakan Improved K-Means dengan Particle Swarm Optimization,” J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 6, p. 737, 2018, doi: 10.25126/itik.20185685.
[15] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster,” IOP Conf. Ser. Mater. Sci. Eng., vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.
[16] E. Patel and D. S. Kushwaha, “Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model,” Procedia
[17] M. Nishom, “Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square,” *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019. doi: 10.30591/jpit.v4i1.1253.

[18] Y. A. Auliya, W. F. Mahmudy, and Sudarto, “Land Clustering for Potato Plants Using Hybrid Particle Swarm Optimization and K-Means Improved by Random Injectio,” *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 42–56, 2019. doi: 10.1109/COMITEE.2019.8921207.

[19] S. Monalisa, “Klusterisasi Customer Lifetime Value dengan Model LRFM menggunakan Algoritma K-Means,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 2, p. 247, 2018, doi: 10.25126/jtiik.201852690.