Analyze K-Value Selected Method of K-Means Clustering Algorithm to Clustering Province Based on Disease Case

Septian Wulandari

***Abstract:** Disease cases throughout Indonesia has increased as seen from the Indeks Pembangunan Masyarakat (IPKM). Globalization has the effect of increasing human mobility across provinces, thus accelerating the process of spreading epidemics that could pose a threat for Indonesia. The speed of action from government is needed to reducing the level if outbreaks of the disease. For this reason, accuracy from the government is needed to solving this problem. The data were taken from data disease cases in 2015 which consisted of 34 provinces in Indonesia based on the Central Statistics Agency in Indonesian. In K-Means clustering, determining of K-value is needed because it affects in convergence results. To solve this problem, this research analyzes three methods of K-Value, there are Silhouette, Elbow, and Gap Statistics Methods.The result of testing three methods of determining K-value obtained execution times on Silhouette 13.09s, Elbow 14.76s, and Gap Statistics 20.28s. So, choosing Silhouette method produces 2 optimal clusters, there are low cluster level (C1) and high cluster level (C2). The correlation matrix to understand the relationship between each disease is performed and a value of 0.88 is obtained there is the strong linear correlation between Pneumonia and Pulmonary TB. Then, modeling the relationship between these two variables by fitting linear equations. The results of C1 cluster based on disease cases were obtained 32 provinces and for C2 cluster were 2 provinces there areWest Java and East Java. Based on the results of the clustering can be input to the Indonesian government to tackle disease cases in all provinces in Indonesia.*

***Keywords:** Data Mining; Disease; K-Value; K-Means; Clustering.*

I. INTRODUCTION

The number of disease cases in Indonesia from year to year has increased seen from the Public Health Index(IPKM). This is caused the globalization of human mobility across provinces in Indonesia, thereby accelerating the spread of disease outbreaks to all provinces in Indonesia. Therefore, the actions of the government needed to reducing the spread of disease outbreaks in Indonesia. To get a information of disease cases in Indonesia, it is necessary to have actions from the government. So, to get an idea, provincial clustering analysis will be carried out using the K-Means algorithm. This method was first proposed by James B Macqueen in 1967. K-Means clustering algorithm is a partitioning method that aims to grouping objects into k clusters with$(k < n)$and $k$is predetermined value[1].Finding objects based on the average of the nearest cluster is the basis for grouping the k-means method in this research.

The most important initial step in the K-Means algorithm is determining the value of the number of clusters or k values. The problem in this algorithm is usually the number of clusters does not have the right way to choose the value of k [2]. If the determination of $k$value is not good, it will affect the results of clustering. For this reason, the determination of k value in this study will be tested using different k value in determination methods. There are many methods of determining k value, namely the Silhouette method, the Elbouw method, and the Gap statistics method.

The results of the three methods can be a reference in determining k value in the K-Means algorithm.There are several similar studies regarding clustering analysis using the K-Means algorithm in infectious diseases in majalengka district [3]. Whereas in this study a provincial clustering based on disease case was obtained from the Central Statistic Agency Indonesian in 2015 with cases of disease taken were Pneumonia, Kusta, Tetanus Neonatorum, Campak, Diare, DBD, Pulmonary TB, and Malaria[4].

II. MATERIALS AND METHODS

A. K-Means

K-Means clustering algorithm is the partitioning methods that partition data into one or more clusters. KK-Means clustering algorithm is a method by inputting $k$as many as $n$objects so that the cluster is formed, which is a cluster that has a closeness between each member in one cluster while for members who are not close to another cluster.[5]. The cluster are formed with the calculation of the distance between data input with each centroid cluster. K-means clustering algorithm calculates different centroid centers in the input data with the same k value, so that data that has values close to centroid will become one cluster [6]. To calculate the distance between each data with each centroid cluster, it is using Euclidean distance, i.e.

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where:

$d(x,y)$ = the distance of data x to the centroid of the cluster y

x_i = data i in n data

y_i = data j in n data

n =dimension data

Revised Manuscript Received on January 5, 2020.

* Correspondence Author

Septian Wulandari*, Program StudiInformatika, UniversitasIndraprasta PGRI Jakarta, Indonesia. E-mail: septian.pmb09@rocktmail.com

Input data that has been calculated by Euclidean distance will be one cluster k if the distance between the input data with centroid has the smallest value compared to the distance between the input data and centroid on the other cluster. The new centroid of cluster is calculated using the formula:

$$y_i = \frac{\sum_{i=1}^{p} x_i}{p} \tag{2}$$

where: $x_i \in$ cluster k and $p =$ the number of cluster$k$

The results of clustering change depending on the choice of the centroid of the cluster. This results in unstable clustering results. Therefore, it is necessary to determine centroid of cluster that depends on the choice of k value.

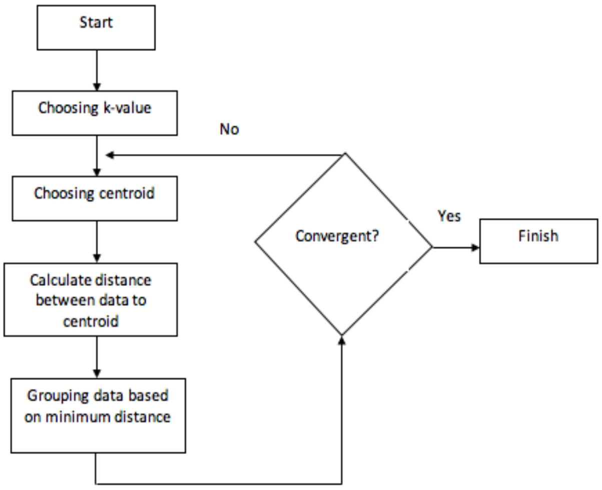

Start
Choosing k-value
No
Choosing centroid
Convergent?
Yes
Finish
Calculate distance between data to centroid
Grouping data based on minimum distance

Fig. 1. Diagram of K-Means [7]

B. K-Value

The number of clusters in the k-means clustering algorithm is determined by finding the best k value. Determination of the value of $k$value produces a cluster of data that has been inputted [2]. The identification of determining the number of $k$clusters is the most important at the first step in the K-Means algorithm. If the determination of the number of clusters $k$is not good, it will affect the results of clustering. So, we need an algorithm to determine the $k$value. There are many methods for finding the best k value. These methods include the Silhouette method, the Elbow method, and the Gap Statistics method. These three methods have been widely used in determining the $k$value. All three display a graph that will show the best k value that is used as a basis in the K-Means algorithm process. Comparison of the three methods can be a reference in selecting the best k value in this research.

- **Silhouette Method**

  The Silhouette Method is the method first introduced by Peter J Rousseeuw. The Silhouette method is widely used in determining k value or the number of clusters.k value is determined by the average of the Silhouette coefficients. For points i in cluster A, Silhouettes of i, the formula of $s\ (i)$ for i in cluster A from Silhouette of i[7]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

$a(i)$ value is the mean dissimilarity between point i and the other point in cluster A and $b(i)$ is the mean dissimilarity between point i and the points in the closest cluster to cluster A. The mean valueof silhouettevalue in the input data S' is called the width average silhouette for all points in the input data. The formula of Silhouette coefficient (SC) is:

$$SC = \max\{S'(k)\} \tag{4}$$

with $k = 2,3, \dots, n-1$.

- **Elbouw Method**

  The Elbow method is a k-value method that determines the best k value by looking at the number of places that will form Elbouw at a point [8]. Performance indicators use number of squared errors (SSE).Clusters are said to be convergent when obtaining a smaller value compared to others [2]. SSE formula in K-Means [8]:

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} \|x_i - c_k\|_2^2 \tag{5}$$

The Elbow algorithm method in K-Means as follows[9]:

1. Initialize $k = 1$
2. start
3. increment the value of k
4. measure the cost of the optimal quality solution
5. if at the same cost of the solution drops dramatically
6. that's the true k
7. end

- **Gap Statistics**

  The Gap Statistics method is a method first introduced by Tibshirani. This method is used in data mining algorithms that aim to increase the efficient estimation of the best number of clusters. In K-Means clustering algorithm to determine the number of clusters is done by calculating the total distance of all datasets from the mean cluster known as dispersion. Gap values can be maximized by minimizing the value of k. This is done by standardizing the$\log(W_k)$comparison with the zero reference data distribution[10].

  Thus, the optimum number of clusters k obtained by compilation of log values $\log(W_k)$ as the value farthest below the curve with formula [2]:

$$Gap_n(k) = E_n^*(\log(W_k) - logW_k\, E_n^*(\log(W_k)) \tag{6}$$

$$= \left(\frac{1}{p}\right) \sum_{b=1}^{p} \log(W_{kb}^*) \approx \left(\frac{1}{p}\right) \sum_{b=1}^{p} \log(W_{kb}^*) s(k)$$

$$= \sqrt{\frac{1+p}{p}}\, s(k)$$

where the expected value $E_n^*(\log(W_k))$ is calculated by Monte Carlo sampling for reference distribution. And then, the optimum number of clusters is chosen as the smallest k such that $Gap(k) \geq Gap(k+1)$.

III. RESULT AND DISCUSSION

The data in this research are cases of disease in 34 provinces in Indonesia taken from the data of CentralStatistic Agency Indonesianin 2015. Cases of disease taken are Pneumonia, Kusta, Tetanus Neonatorum, Campak, Diare, DBD, Pulmonary TB, and Malaria. The data obtained has a value that is too small and too large a value. Therefore, the data normalization process is carried out using min-max normalization [1]. This is done so that the data is between the range of 0 to 1.

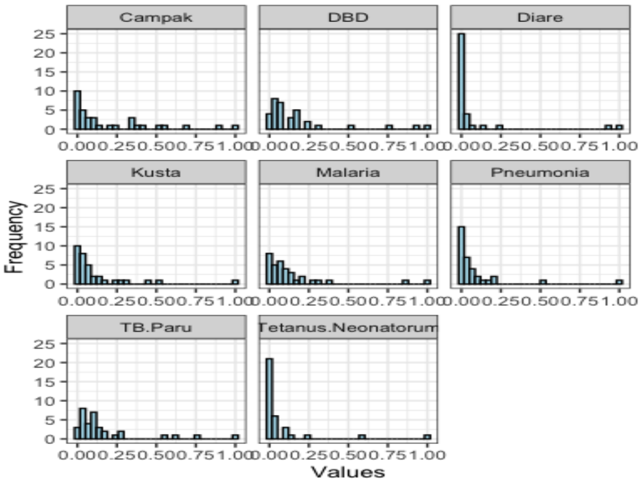


Fig. 2. The normalization of data set

Then a correlation matrix to understand the relationship between each disease is performed and the results obtained in Fig.3. From Fig.3 shows that the strong linear correlation is between Pneumonia and Pulmonary TB with the value of linear correlation is 0.88.

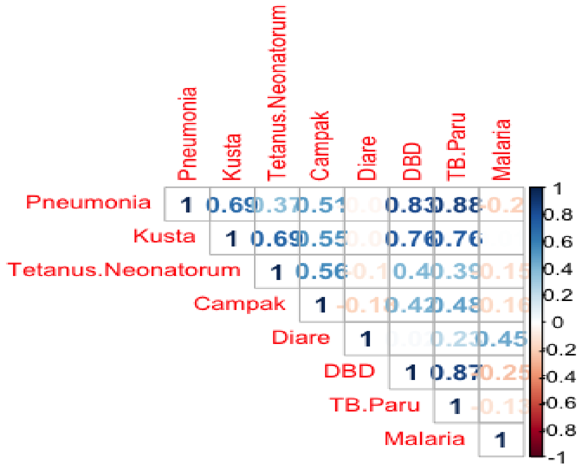
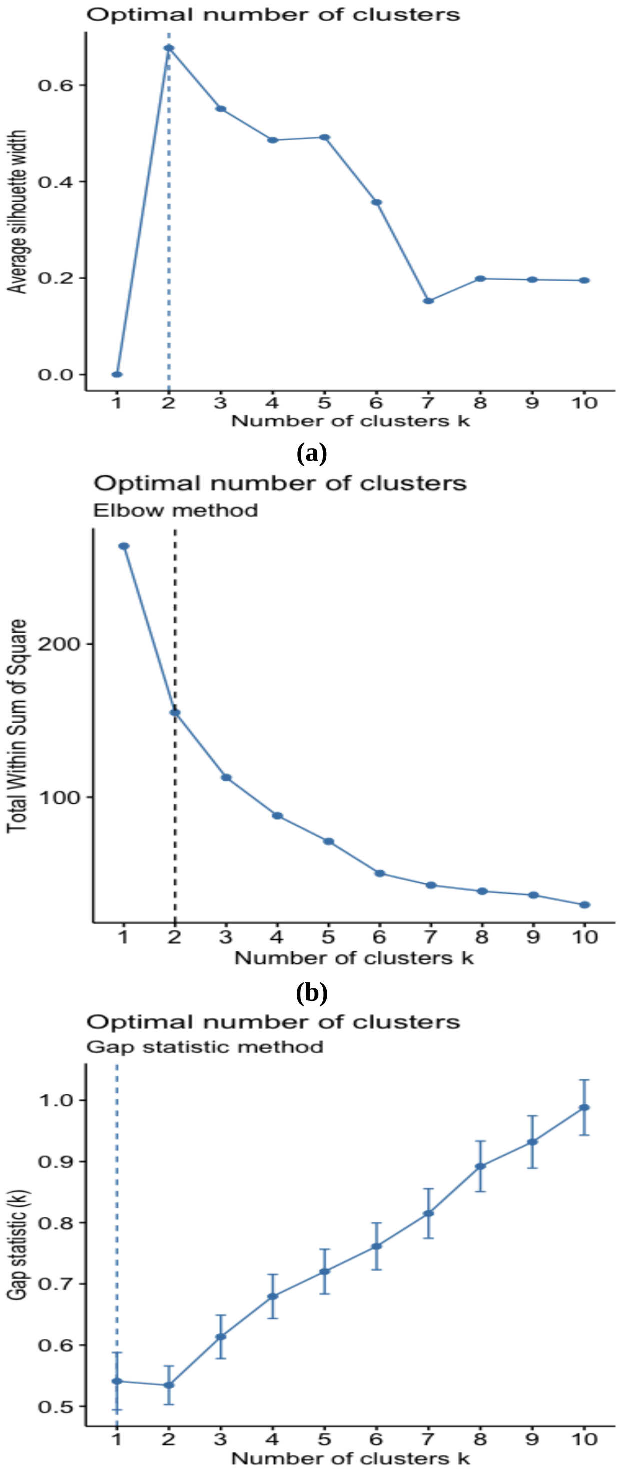


Fig.3. Correlation matrix ofdata set

Then, the k value selection process is performed to finding the best number of clusters using K-Means clustering algorithm. The first test is done using the Silhouette method, then the second is the Elbouw method, and the last is the Gap statistics method. The results of the three methods for determining k values can be seen in Fig.4 and Table- I.

Table- I: Experiment result of three algorithm

No	Name	k value	Execution Time
1	Silhouette Method	2	13.09 s
2	Elbouw Method	2	14.76 s
3	Gap statistics Method	1	20.28 s



(a)

(b)

(c)

Fig.4. Determining of$k$ value with (a) Silhouette method, (b) Elbouw method, and (c) Gap statistics method

In Table- I it can be seen that the Gap Statistics method has the longest execution time of 20.28 seconds, while the execution time for the Silhouette method has the fastest execution time of 13.09 second. *k*value for the Silhouette method produces 2 number of clusters, while for Elbouw method produces 2 number of clusters, and for Gap Statistics method produces one number of clusters. Therefore, because there are two methods that have the same*k* k value, the Silhouette method and the Elbouw method produce*k* = 2, then the *k*value used in this study is *k* = 2. After stabilizing the *k*value with*k* = 2, the next step is clustering process using K-Means clustering algorithm, the results in Fig. 5 are obtained.

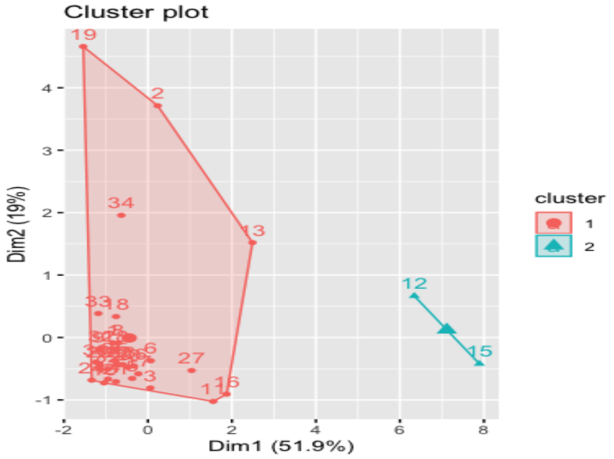


Fig.5. Clustering result

Table- II: Clustering Result

No	Cluster	Province
1	1	Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung Islands,Riau islands, DKI Jakarta, DI Yogyakarta, Central Java, Banten, Bali, NTB, NTT, West Kalimantan, CentralKalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Papua
2	2	West Java, East Java

Table 2 shows the clustering result of 34 provinces in Indonesia based on disease cases. Clustering results show that there are 2 clusters, namely low cluster level (C1) and high cluster level (C2). Cluster C1 consists of 32 provinces and Cluster C2 consists of 2 provinces namely West Java and East Java. Based on the results of the clustering can be input to the Indonesian government to tackle disease cases in all provinces in Indonesia.

IV. CONCLUSION

Based on the results of the k value analysis and testing of provincial clustering based on disease cases using the K-Means algorithm, the k value obtained using the Silhouette method produces 2 clusters with an execution time of 13.09 s, Elbouw method produces 2 clusters with an execution time of 14.76s, and Gap Statistics 20.28s. Thus, the Silhouette method provides the most optimal *k* value by dividing data into 2 clusters because of the faster time compared to other methods. For the clustering results there are 2 clusters namely low cluster level (C1) consisting of 32 provinces which are low disease spread and high cluster level (C2) consisting of 2 provinces namely West Java and East Java which are high disease spreads. Based on the results of the clustering can be input to the Indonesian government to tackle disease cases in all provinces in Indonesia.

REFERENCES

1. K. R. Adzima, A. Bustamam, and D. Aldila, “The implementation of k-means partitioning algorithm in HOPACH clustering method,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 243, no. 1, 2019.
2. C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J*, vol. 2, no. 2, 2019, pp. 226–235.
3. N. Suherni and M. Maduratna, "Analysis of Subdistrict Classification in the City of Surabaya Based on Factors Causing the Occurrence of Tuberculosis," *Sains dan Seni ITS*, vol. 2, no. 1, 2013, pp. 2337–3520.
N. Suherni and M. Maduratna, “Analisis Pengelompokan Kecamatan di Kota Surabaya Berdasarkan Faktor Penyebab Terjadinya Penyakit Tuberkulosis,” *J. Sains dan Seni ITS*, vol. 2, no. 1, 2013, pp. 2337–3520.. (in Indonesian)
4. BPS, “Total Population and Population Growth Rate in City of Bandung 2011 Regency / City,” 2018.
BPS, “Jumlah Penduduk dan Laju Pertumbuhan Penduduk di Kota Bandung 2011 Kabupaten / Kota Program dan,” 2018. (in Indonesian)
5. A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, “Using K-Means Clustering to Cluster Provinces in Indonesia,” *J. Phys. Conf. Ser.*, vol. 1028, no. 1, 2018, pp. 0–6.
6. K. Sirait, Tulus, and E. B. Nababan, “K-Means Algorithm Performance Analysis with Determining the Value of Starting Centroid with Random and KD-Tree Method,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017, pp. 0–6.
7. M. B. A.-Z. and M. al Rawi, “An Efficient Approach for Computing Silhouette Coefficients,” *J. Comput. Sci.*, vol. 4, no. 3, 2008, pp. 252–255.
8. N. Putu, E. Merliana, and A. J. Santoso, “*Analysis of Determination of the Best Number of Clusters on the K-Means Method, "in Proceedings of the Multi-Disciplinary National Seminar & Call For Papers Unisbank (Sendi_U)*, pp. 978–979.
N. Putu, E. Merliana, and A. J. Santoso, “Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means,” in *Prosiding Seminar Nasional Multi Disiplin Ilmu & Call For Papers Unisbank (Sendi_U)*, pp. 978–979. (in Indonesian)
9. P. Bholowalia, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” *Int. J. Comput. Appl.*, vol. 105, no. 9, 2014, pp. 17–24.
10. A. M. El-Mandouh, H. A. Mahmoud, L. A. Abd-Elmegid, and M. H. Haggag, “Optimized K-means clustering model based on gap statistic,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, 2019, pp. 183–188.

AUTHOR PROFILE

SeptianWulandarigraduated with a degree in Mathematics Education from Universitas Negeri Jakarta in 2013 and continued his Masters in Mathematics at the Universitas Indonesia in 2016 with a special scope is Bioinformatics. He is currently a lecturer at the UniversitasIndraprasta PGRI in Faculty of Engineering and Information. Actively writing scientific articles on statistics, mathematics education, and bioinformatics. Has appeared as a speaker at 3 International conferences. The article about Collaboration and implementation of self-organizing maps (SOM) partitioning algorithm in HOPACH clustering method was published in AIP Conference Proceedings in 2018.