Jurnal Mantik, 5 (2) (2021) 1341-1347

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

K-MEANS CLUSTERING SYMPTOMS OF COVID-19 POSITIVE IN INDONESIA

Wahyuddin S[1] , Mahir Pradana[2]

[1]Dept. of Informatics Management, AMIK Lamappapoleonro Soppeng, Indonesia
[2]Dept. of Business Administration. Telkom University, Bandung, Indonesia

E-mail: wahyu@amiklps.ac.id, mahirpradana@telkomuniversity.ac.id

ARTICLE INFO

Article history:
Received: 30 July 2021
Revised: 12 August 2021
Accepted:30 August, 2021

Keywords:
Covid-19, Pandemic, K-Means, Clustering, R Studio.

ABSTRACT

This study aims to group the positive symptoms of covid-19 in Indonesia based on coronavirus symptom data. Some symptoms are characteristic of a person contracting positive coronavirus, usually causing respiratory infections, ranging from the common cold to serious diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Like other respiratory diseases, COVID-19 can cause mild symptoms including colds, sore throats, coughs, and fevers. There are 14 main symptoms obtained from Indonesia's COVID-19 detention task force. The data was taken in August 2021. For grouping cases, this study used the K-Means Clustering method by using R studio software to analyze the data. After grouping, there are 3 groups of positive symptoms of covid-19 in Indonesia. Positive symptom grouping results are expected to provide feedback to policymakers to see which symptom groups need to take precedence.

1. Introduction

Coronavirus is a large family of viruses that cause diseases in humans and animals. In humans, it usually causes respiratory infections, ranging from the common cold to serious diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). A new type of coronavirus found in humans since an extraordinary event emerged in Wuhan China, in December 2019, then named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2) and caused Coronavirus Disease-2019 (COVID-19)[1].

Like other respiratory diseases, COVID-19 can cause mild symptoms including colds, sore throats, coughs, and fevers. About 80% of cases can be recovered without the need for special treatment. About 1 in every 6 people may suffer from severe pain, such as accompanied by pneumonia or difficulty breathing, which usually appears gradually[2]. Although the death rate of the disease is still low (about 3%), but older people, and people with pre-existing medical conditions (such as diabetes, high blood pressure, and heart disease), are usually more susceptible to becoming seriously ill[3]. Looking at the developments to date, more than 50% of confirmed cases have been declared improved, and the recovery rate will continue to increase.

A person can be infected from covid-19 sufferers. The disease can be spread through droplets from the nose or mouth during coughing or sneezing. The droplet then falls on the surrounding objects. If someone else touches an object that has been contaminated with the droplet, then the person touches the eye, nose, or mouth (face triangle), then that person can be infected with COVID-19. Or it could be someone infected with COVID-19 when accidentally inhaling droplets from a sufferer. This is why it is important to keep a distance of approximately one meter from the sick person[4].

To date, experts are still conducting investigations to determine the source of the virus, the type of exposure, and the way it is transmitted.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2. Related Work

2.1 Data Mining

Data mining activities constitute an iterative process aimed at the analysis of large databases, to extract information and knowledge that may prove accurate and potentially useful for knowledge workers engaged in decision making and problem-solving.

The data mining process is based on inductive learning methods, whose main purpose is to derive general rules starting from a set of available examples, consisting of past observations recorded in one or more databases. In other words, the purpose of a data mining analysis is to draw some conclusions starting from a sample of past observations and to generalize these conclusions concerning the entire population, in such a way that they are as accurate as possible. The models and patterns identified in this way may take on different forms, which will be described in the following chapters, such as linear equations, sets of rules in *if–then–else* form, clusters, charts, and trees[5].

2.2 K-Means Clustering

The K-means algorithm receives as input a dataset *D*, a number *K* of clusters to be generated, and a function dist(x_i, x_k) that expresses the inhomogeneity between each pair of observations, or equivalently the matrix D of distances between observations[6]. Given a cluster C_h, h = 1, 2, . . . K, the centroid of the cluster is defined as the point having coordinates equal to the mean value of each attribute for the observations belonging to that cluster[7], that is,

$$Zhj = \frac{\sum_{x_i \in C_h} x_{ij}}{card\{C_h\}}$$

2.3 K-Means Algorithm

a) During the initialization phase, *K* observations are arbitrarily chosen in *D* as the centroids of the clusters.
b) Each observation is iteratively assigned to the cluster whose centroid is the most similar to the observation, in the sense that it minimizes the distance from the record.
c) If no observation is assigned to a different cluster concerning the previous iteration, the algorithm stops.
d) For each cluster, the new centroid is computed as the mean of the values of the observations belonging to the cluster, and then the algorithm return to step 2.

3. Methods

The technique used to perform grouping is the K-Means Clustering method. K-Means is one of the algorithms in data mining that can be used to group/cluster data. In conducting data analysis this research uses RStudio software.There are several approaches to creating clusters, one of which is to create rules that dictate membership in the same group based on the level of equality between its members. Another approach is to create a set of functions that measure some of the properties of that grouping as a function of multiple parameters of a clustering[8][9].

3.1 Data Collections

We use data from the website (https://covid19.go.id/peta-sebaran) The data was taken in August 2021. The data used consists of 5 variables as in table1.

TABLE 1
DESCIPTIVE STATISTIC OF COVID-19 SYMPTOMS IN INDONESIA

Symptoms	Positive	Self-isolation	Recovered	Deaths
Cough	63,8 %	0,8 %	58 %	5 %
History of fever	43,7 %	0,4 %	39,5 %	3,8 %
Fever	38,5 %	0,5 %	34,2 %	3,8 %

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Cold	36,4 %	0,6 %	34,3 %	1,6 %
Limp	26,2 %	0,4 %	22,5 %	3,3 %
Shortness of breath	23,3 %	0,2 %	18,1 %	5 %
Headache	23,1 %	0,4 %	21 %	1,8 %
Sore throat	23,1 %	0,3 %	20,9 %	1,8 %
Muscle cramps	15 %	0,3 %	13,4 %	1,2 %
Nauseous	12,3 %	0,3 %	10,8 %	1,4 %
Stomach ache	5,5 %	0,1 %	4,6 %	0,8 %
Diarrhea	5,2 %	0,1 %	4,6 %	0,6 %
Chills	1,7 %	0 %	1,4 %	0,3 %
Others	0 %	0 %	0 %	0 %

Source: https://covid19.go.id/peta-sebaran

3.2 Positive Symptom Data of Covid-19

We take covid-19 positive symptom data and display it like in figure 1.

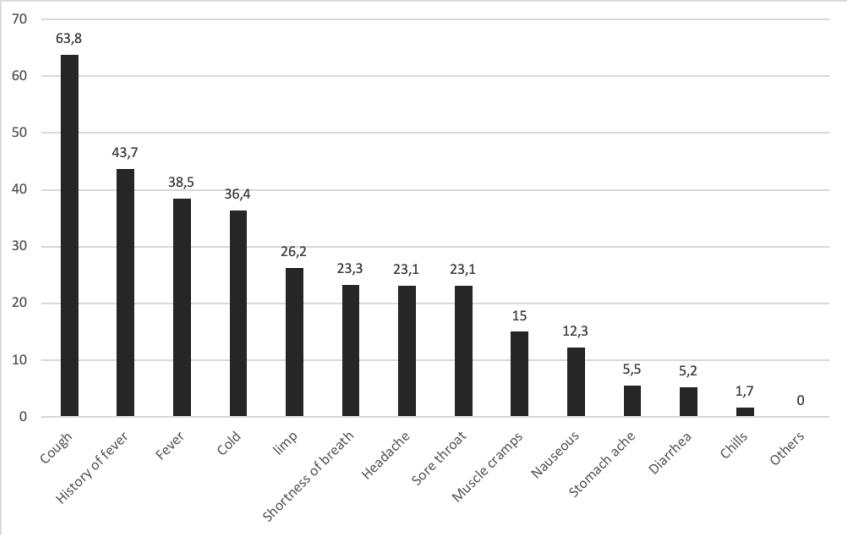

Fig 1. Plot symptoms of covid-19 positive

3.3 Elbow Method

This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach[10]. Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

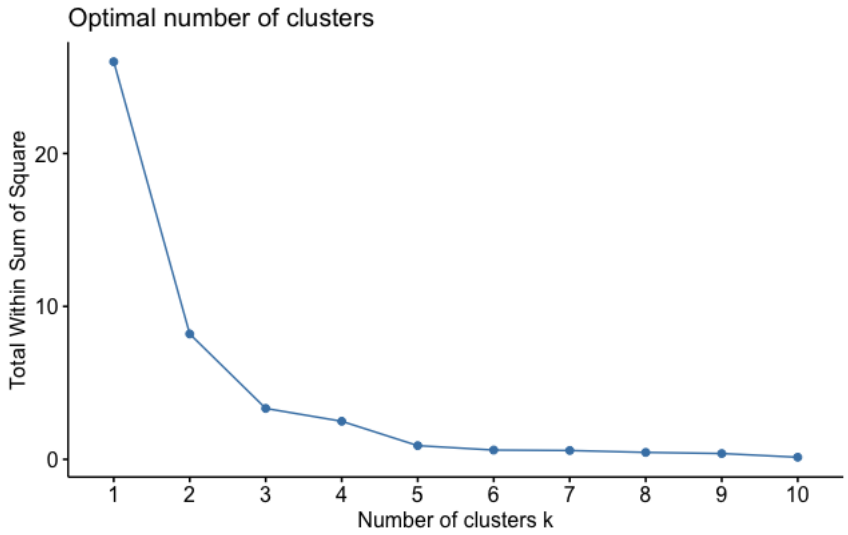

Fig 2. Plot optimal number of clusters "elbow method"

3.4 Average Silhouette Method

The silhouette value measures how similar a point is to its cluster (cohesion) compared to other clusters (separation)[11].

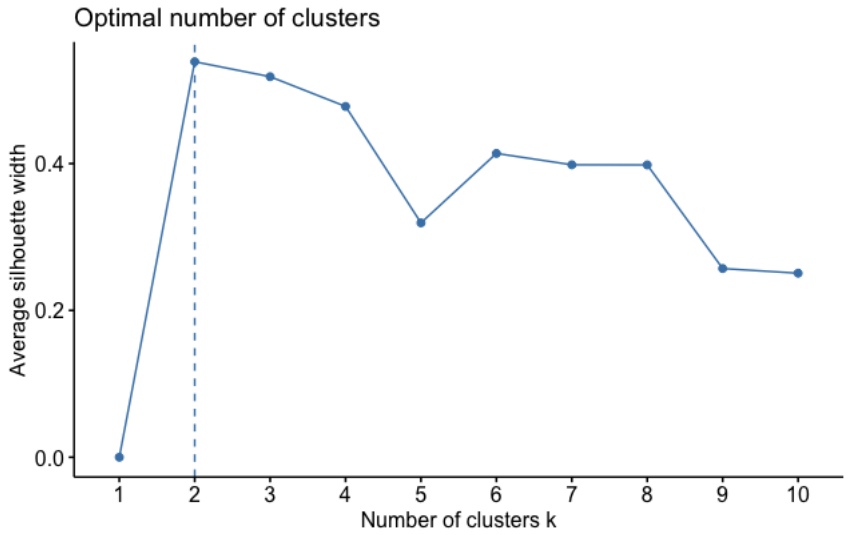

Fig 3. Plot optimal number of cluster "average silhoutte method"

3.5 Gap Statistic Method

The approach can be applied to any clustering method (i.e. K-means clustering, hierarchical clustering). The gap statistic compares the total intracluster variation for different values of *k* with their expected values under the null reference distribution of the data (i.e. a distribution with no obvious clustering). The reference dataset is generated using Monte Carlo simulations of the sampling process. That is, for each variable (x_i) in the data set we compute its range [$min(x_i)$, $max(x_j)$] [$min(x_i)$, $max(x_j)$] and generate values for the n points uniformly from the interval min to max[12].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

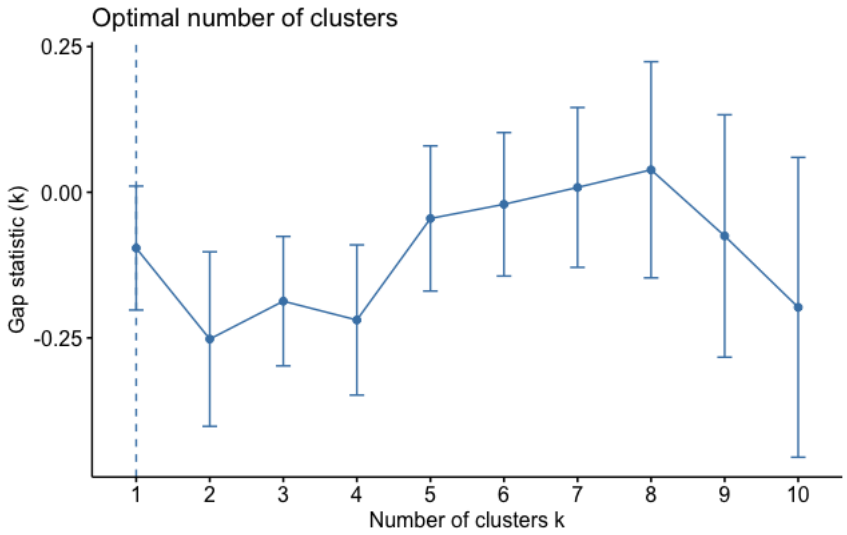

Fig 4. Plot symptoms of covid-19 positive "gap statistic method"

4. Result and Discussion

Based on descriptive statistical analysis (Table 1) there are 14 symptoms of covid-19 disease in Indonesia. Most cough symptoms were affected by positive cases with 63.8% and 8% self-isolation for cough symptoms. Cough symptoms are also symptoms that have the highest cure rate of about 58%. As for the death rate of symptoms of fever and fever history most who died about 3.8% history of fever and 3.8% fever. Cough accompanied by fever and shortness of breath are corona symptoms found in most cases of Covid-19. According to the U.S. Centers for Disease Control and Prevention (CDC)[13], several other symptom signals appear within two to 14 days after a person is exposed to the coronavirus.

In addition, the optimal number of groups k is determined using three (3) most commonly used approaches: Elbow, Silhouette, and Gap Statistics. The results can be seen in figures 2,3 and 4. (Figure 2), the Elbow method obtains the optimal k at k = 2, (Figure 3) the Silhouette method obtains many optimal clusters at k = 2, and (Figure 3) gap statistics obtain the optimal k value to form clusters at k = 2. Therefore, based on the results of this method, it can be concluded that the optimal k value for forming a cluster is 2. Furthermore, the results of clustering analysis using K-means with k = 2 are presented in Table 2. Table 2 is the result of grouping cases of posistif symptoms of covid-19, As shown in the table, cough symptoms are most and is a grouping of 1 about 63.8% followed by symptoms of History of fever, Fever, Cold and Limp. Corona cough is very similar to a regular cough. That's why people who are sick with cough are recommended to wear a mask to prevent the risk of transmission if they are positive for Covid-19[14]. To get the right diagnosis, you can undergo tests at government-appointed hospitals or private hospitals that have complete facilities for corona cases. Therefore, it can be concluded that by doing a cluster of symptoms of positive cases of covid-19, a person is given an overview of patterns and solutions for the spread of corona disease symptoms related to this distribution pattern.

In fig.5. is the result of k-means clustering, wherein the grouping there are 3 clusters of symptoms of covid-19. The results of the grouping are then displayed in table II.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

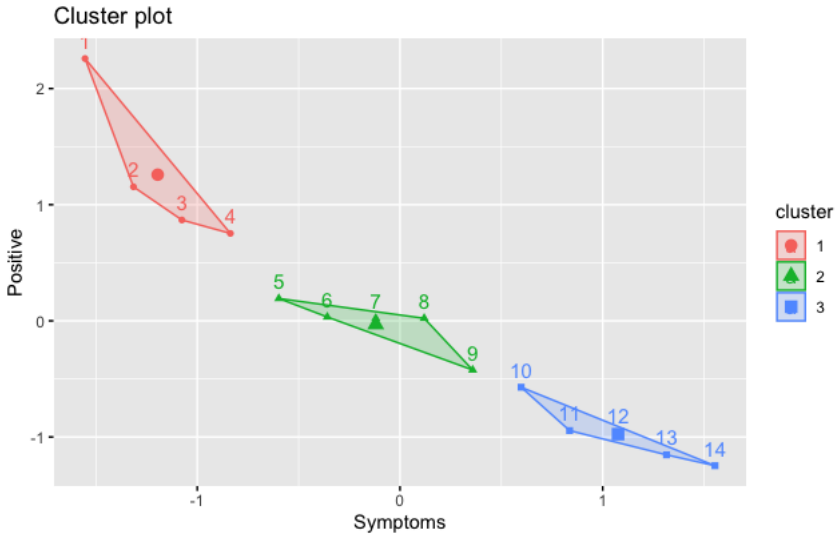

Fig 5. Result of symptoms of covid-19 positive with K-Means clustering.

TABLE 2
RESULT K-MEANS CLUSTERING SYMPTOMS OF COVID-19 POSITIVE

No	Symptoms	Positive	Cluster
1	*Cough*	63,8 %	1
2	*History of fever*	43,7 %	1
3	*Fever*	38,5 %	1
4	*Cold*	36,4 %	1
5	*Limp*	26,2 %	1
6	*Shortness of breath*	23,3 %	2
7	*Headache*	23,1 %	2
8	*Sore throat*	23,1 %	2
9	*Muscle cramps*	15 %	2
10	*Nauseous*	12,3 %	3
11	*Stomach ache*	5,5 %	3
12	*Diarrhea*	5,2 %	3
13	*Chills*	1,7 %	3
14	*Others*	0 %	3

5. Conclusion

Based on the results of this study, there are 3 clusters of symptoms of covid-19 positive, each consisting: Clusters 1 (cough, history of fever, fever, cold, limp); Clusters 2 (shortness of breath, headache, sore throat, muscle cramps); Cluster 3 (nauseous, stomach ache, diarrhea, chills, others). Positive symptom grouping results are expected to provide feedback to policymakers to see which symptom groups need to take precedence.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

References

[1] C. H. Sudre *et al.*, "Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID symptom study app," *Sci. Adv.*, vol. 7, no. 12, pp. 1–8, 2021.

[2] R. Kurniawan, S. N. H. S. Abdullah, F. Lestari, M. Z. A. Nazri, A. Mujahidin, and N. Adnan, "Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries," *2020 8th Int. Conf. Cyber IT Serv. Manag. CITSM 2020*, 2020.

[3] M. Pradana, S. Syahputra, A. Wardhana, B. R. Kartawinata, and C. Wijayangka, "The Effects of Incriminating COVID-19 News on the Returning Indonesians' Anxiety," *J. Loss Trauma*, pp. 656–661, 2020.

[4] M. Pradana, N. Rubiyanti, W. S, I. Hasbi, and D. G. Utami, "Indonesia's fight against COVID-19: the roles of local government units and community organisations," *Local Environ.*, vol. 25, no. 9, pp. 741–743, 2020.

[5] W. S. Fauzi Insan Estiko, "Analysis of Indonesia's Inflation Using ARIMA and Artificial Neural Network," *Econ. Dev. Anal. J.*, vol. 8, no. 2, pp. 151–162, 2019.

[6] Y. Qiao, Y. Li, and X. Lv, "The application of big data mining prediction based on improved K-means algorithm," *Proc. - 2019 34rd Youth Acad. Annu. Conf. Chinese Assoc. Autom. YAC 2019*, pp. 348–351, 2019.

[7] 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.

[8] S. K. Dini and A. Fauzan, "Clustering Provinces in Indonesia based on Community Welfare Indicators," *EKSAKTA J. Sci. Data Anal.*, vol. 1, no. 1, pp. 56–63, 2020.

[9] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, no. 0123456789, 2021.

[10] H. B. Tambunan, D. H. Barus, J. Hartono, A. S. Alam, D. A. Nugraha, and H. H. H. Usman, "Electrical peak load clustering analysis using K-means algorithm and silhouette coefficient," *Proceeding - 2nd Int. Conf. Technol. Policy Electr. Power Energy, ICT-PEP 2020*, pp. 258–262, 2020.

[11] A. Doroshenko, "Analysis of the distribution of COVID-19 in Italy using clustering algorithms," *Proc. 2020 IEEE 3rd Int. Conf. Data Stream Min. Process. DSMP 2020*, pp. 325–328, 2020.

[12] P. Dzimińska, S. Drzewiecki, M. Ruman, K. Kosek, K. Mikołajewski, and P. Licznar, "The use of cluster analysis to evaluate the impact of covid-19 pandemic on daily water demand patterns," *Sustain.*, vol. 13, no. 11, 2021.

[13] Centers for Disease Control and Prevention, "Symptoms of COVID-19 | CDC." [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. [Accessed: 15-Jan-2021].

[14] Healthline, "Different Symptoms for COVID, Flu, Allergies, and Cold." [Online]. Available: https://www.healthline.com/health-news/flu-allergies-coronavirus-different-symptoms. [Accessed: 15-Jan-2021].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.