Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 3 No. 1 (2023) 36 - 39 E-ISSN: 2809-8064

Cluster Analysis of Covid-19 in Indonesia Using K-means Method

Claudia Larasvaty[1], Siti Khomsah[2], Rona Nisa SA[3]

[1,3]Department of Information System, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[2]Department of Data Science, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[1]17103005@ittelkom-pwt.ac.id, [2]siti@ittelkom-pwt.ac.id, [3]rona@ittelkom-pwt.ac.id

Abstract

These days technology is rapidly increasing and developing in various fields, especially data storage. The information that has been stored in a database is usually called a dataset. Covid-19 is a new type of respiratory disease that attacks the respiratory system with rapid transmission, followed by the increasing number of Covid-19 cases that continues to increase every day in all provinces in Indonesia. This study aims to cluster the spread of Covid-19 in every province in Indonesia by using the data that obtained from the website named kaggle with many data variables. The method used in this research is K-Means. From many variables in the data, for this study only 3 variables were taken, which are: Number of Recovery, Number of Deaths, and Number of total Cases in Covid-19 in Indonesia. These 3 variables then will be applied using the K-Means method and formed 3 provincial groups. By using the clustering method and the K-means algorithm, this research can be carried out to find the characteristics of the distribution in each province in Indonesia by looking at the best clusters.

Keywords: Covid-19, Clustering, Data Mining, K-Means, Knowledge Discovery in Data Process, Rapid Miner

1. Introduction

Covid-19 is a new type of pneumonia that attacks the respiratory system, this disease has been declared as a global pandemic since March 2020 [1]. According to the Governor of DKI Jakarta, in June 2021 Jakarta was in the highest position during the pandemic, the distribution of positive cases was based on city and district areas, East Jakarta was in the highest position with 9,394 positive cases [2]. The entire community continues to strive to overcome the ongoing pandemic by doing the smallest things such as wearing masks, washing hands, avoiding crowds and closing a number of areas and locations that can invite crowds [3]. Until 2022, Covid-19 still exists and can infect someone, marked by the presence of a new variant called Omicron which spreads faster. The number of deaths in the third wave of the Covid-19 pandemic reached 22,108 cases per day. The third wave can still be overcome by resting at home and taking medicine and vitamins. The number of additional cases of Covid-19 continues to increase every day in all provinces in Indonesia [4],[5]

Technology is currently advancing and developing in various fields such as science, government, health where these fields face a large amount of data and new technologies to store, process and analyze these data [6]. The presence of this technology will be more useful if the devices and modules that are prepared can contribute and be efficient properly and correctly in order to produce something that is effective and efficient for the community [7]. Various techniques and methods can be used to organize our group, one of which is Clustering. Clustering is one of the methods in Data Mining that is used to analyze or group large data into a cluster. Clustering has many methods in it, one of which is K-means [8].

The K-means method is a clustering method that can group large amounts of data with fast and efficient computation time. The application of K-means requires three parameters which are defined by the user. K-means is an iterative clustering algorithm where the K-means algorithm assigns a random cluster value, and that value becomes the center of the cluster or is called the mean. Calculation of the distance of each existing data to each means using a formula until the closest distance can be found from each data with means [9]. By using the clustering method and the K-means algorithm, research can be carried out to find the characteristics of each province with the most Covid-19 spread.

Received: 04-09-2022 | Revised: 01-02-2023 | Published: 02-02-2023

2. Research Methods

This research process includes steps that begin with literature study, data acquisition, selection, data mining, analysis of cluster results, conclusions, and suggestions.

2.1. *Data Acquisition*

This step is data acquisition to obtain the data needed for this research [9],[10]. The data obtained is from a website called kaggle which is a data sourced from covid.go.id, Kemendagri.go.id, bps.go.id, and bnpb-inacovid19.hub.arcgis.com.

2.2. *Selection*

After data acquisition, the next step is choosing 3 variables for the clustering. Total Healed, Total Deaths, and Total cases are the variables that will be used in this study as shown on table 1.

Table 1. Variable Selection

Variables	Name
X1	Total Healed
X2	Total Death
X3	Total Case

2.3. Mining using K-Means

Table 2. Distribution Data

Provincxe	Total Death	Total Healed	Total Case
Aceh	2066	36333	38416
Bali	4046	110003	114233
Banten	2688	129872	132693
Bengkulu	473	22612	23104
DIY	5263	150965	156769
DKI Jakarta	13596	849875	864045
Gorontalo	460	11374	11834
Jambi	789	28968	29768
Jawa Barat	14737	692101	707934

Pick centroid value randomly and three data were taken randomly which are Bali, Bengkulu, and West Java as the center of the cluster.
C1 = (4046, 110003, 114233)
C2 = (473, 22612, 23104)
C3 = (14737, 692101, 707934)
Calculation Example for searching the *centroid:*

A (2066, 36333, 38416) -> C1 (4046, 110003, 114233)

$$\sqrt{(2066-4046)^2+(36333-110003)^2+(38416-114233)^2} = 73696.09$$

Continue calculating all the cluster and will get the result of the distance of the centroid from calculating the first iteration. These are the results of the previous calculation; Colored data is the smallest data in each cluster shown in Table 3 and Table 4.

Table 3. Distribution Data

VARIABLE	Data								
	Aceh (A)	Bali (B)	Banten (C)	Bengkulu (D)	DIY (E)	DKI (F)	Gorontalo (G)	Jambi (H)	JABAR (I)
X1	2066	4046	2688	473	5263	13596	460	789	14737
X2	36333	110003	129872	22612	150965	849875	11374	28968	692101
X3	38416	114233	132693	23104	156769	864045	11834	29768	707934

Table 4. Distribution Data

Provinsi	X1	X2	X3	C1	C2	C3	Keputusan cluster
A	2066	36333	38416	73696.09	20621.85	655891	C2
B	4046	110003	114233	0	126310.9	831524.7	C1
C	2688	129872	132693	27154.98	153360.2	804455.6	C1
D	473	22612	23104	126310.9	0	957815.8	C2
E	5263	150965	156769	59065.03	185374.7	772464.1	C1
F	13596	849875	864045	1053432	1179711	221956.3	C1
G	460	11374	11834	142218.5	15915.58	973729.1	C2
H	789	28968	29768	117096.6	9214.526	948603.7	C1
I	14737	692101	707934	831524.7	957815.8	0	C3

C1= {A,C,E,F,H} →{(2066, 110003,114233), (2688,129872, 132693), (5263, 150965, 156769), (789, 28968, 29768)}
C2={D,G}→{(473,22612,23104), (460, 11374, 11834)}
C3={A,I}→{(2066, 36333, 38416)}

Every cluster already have their group, next is searching for a new centroid value for the 3 cluster that choosen because every cluster each have 3 atribute. This is the example of calculating for new centroid cluster
C1=
X1→(2066+2688+5263+13596+789)/5= 4880.4

X2→(110003+129872+150965+849875+29768)/5= 254096,6
X3→(114233+132693+156769+864045+29768)/5= 259501,6
Continue calculating until C3 and the result of the new centroid value is: [C1 = 4880.4, 254096,6, 259501,6], [C2= 466,5, 16993, 17469], [C3= 8401,5, 3642175,5, 372905]

Continue calculating new coordinat posistion with the variables until calculating the iteration are done. Calculation is done when the value of the centroid already in a stable value or the centroid didn't change

2.4. *Result*

This step will be doing the results are analyzed by looking at the lowest value using the Davies Bouldin Index to get the best cluster comparison.

Step 1 Davies Bouldin Index is to calculate the variance of each cluster with the formula:

$$Var(x) = \frac{1}{N-1}\sum_{i}^{n}(x_i - \overline{x})^2$$

$\bar{x} = RMean\ cluster\ x$

N = Count Cluster

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

With $R_i = \max R_{ij}$

and $R_{ij}\ \frac{var(C_i)+var(c_j)}{\|c_i - c_j\|}$

The next step is to calculate DBI with the following formula. If the DBI calculation has been carried out, then continue the Ratio calculation using the Var Char value that has been obtained previously. The results analyzed are the number of cities in each existing cluster.

3. Results and Discussion

The mining process, clustering, is carried out using the Rapid Miner application. It starts with inputting the data that has been previously selected, then by using the Clustering and Performance operators. After the operator is connected then start trying to input the number of clusters from 2 to 10 as can be seen in Figure 3.1

In the Davies-Bouldin Index (DBI), the lower the cluster value, the better the cluster. Based on the DBI results in table 3.2, the lowest score is in cluster 6 with a value of 0.158.

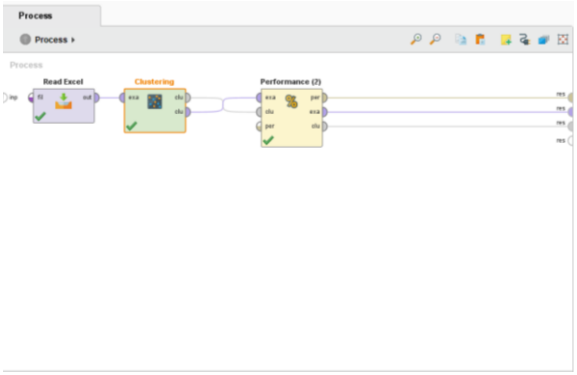

Figure 1. Process Mining

Cluster 6 there are 34 provinces and 6 Clusters of them; Cluster_0 has 8 provinces, namely Bali, Banten, DIY, East Kalimantan, Riau, South Sulawesi, West Sumatra, North Sumatra, Cluster_1 has 1 province, namely West Java, Cluster_2 has 1 province, namely East Java, Cluster_3 has 22 provinces, namely Aceh, Bengkulu, Gorontalo, Jambi, West Kalimantan, South Kalimantan, Central Kalimantan, North Kalimantan, Bangka Belitung Islands, Riau Islands, Lampung, Maluku, North Maluku, NTB, NTT, Papua, West Papua, West Sulawesi, Central Sulawesi, Southeast, North Sulawesi, South Sumatra, Cluster_4 has 1 province, namely DKI Jakarta, and Cluster_5 has 1 province, namely Central Java.

The results above show that the cluster with the highest number of cities is in Cluster_3 Aceh, North Kalimantan, Bangka Belitung Islands, Riau Islands, Lampung, Maluku, North Maluku, NTB, NTT, Papua, West Papua, West Sulawesi, Central Sulawesi, Southeast, North Sulawesi, South Sumatra. Provinces that are included in the selected cluster are large provinces in Indonesia, so that the spread of Covid-19 is more, one of which is North Kalimantan where the spread of COVID-19 cases is mostly due to local transmission which has an impact on increasing Covid-19 cases.

4. Conclusion

The conclusions obtained based on the results of research and cluster analysis of the spread of Covid-19 in Indonesia using the K-Means method are as follows: Meanwhile, the highest number is located in cluster_1 with a value of 0.387. Suggestions for developing cluster analysis research on the spread of Covid-19 in Indonesia using the K-Means method are: This research can be developed using other methods and adding variables for Covid-19 clustering.

References

[1] D. Cucinotta dan M. Vanelli, "WHO declares COVID-19 a pandemic," *Acta Biomed.*, vol. 91, no. 1, hal. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.

[2] "Anies: Jakarta Sedang Hadapi Gelombang Pasien COVID Tertinggi Selama Pandemi." [Daring]. Tersedia pada: https://news.detik.com/berita/d-5622367/anies-jakarta-sedang-hadapi-gelombang-pasien-covid-tertinggi-selama-pandemi.

[3] "Puncak Gelombang 3 Sudah Nampak, Omicron Terus Bermutasi." [Daring]. Tersedia pada: https://www.cnbcindonesia.com/news/20220220134029-4-316789/puncak-gelombang-3-sudah-nampak-omicron-terus-bermutasi.

[4] S. Tentang, W. E. B. E. Di, dan K. Kota, "e-journal 'Acta Diurna' Volume VI. No. 3. Tahun 2017," vol. VI, no. 3, 2017.

[5] Fairuz, Ardianne Luthfika, Rima Dias Ramadhani, and Nia Annisa Ferani Tanjung. "Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial Twitter." *Journal of Dinda: Data Science, Information Technology, and Data Analytics* 1.1 (2021): 42-51.

[6] R. A. Indraputra dan R. Fitriana, "K-Means Clustering Data COVID-19," *J. Tek. Ind.*, vol. 10, no. 3, hal. 275–282, 2020, doi: 10.25105/jti.v10i3.8428.

[7] M. S. Retno Tri Wulandari, S.Si., *DATA MINING*. Yogyakarta: Penerbit Gava Media, 2017.

[8] K. Velusamy, "Data Clustering Using Data Mining Techniques."

[9] Awaliyah, Cahyani Ainun, Agi Prasetyadi, and Apri Junaidi. "Sistem Rekomendasi Desain Website Berdasarkan Tingkat Kemiripan Menggunakan Euclidean Distance." *Journal of Dinda: Data Science, Information Technology, and Data Analytics* 2.2, pp. 75-81, 2022.

[10] Hilmawan, Muhammad David. "Deteksi Sarkasme Pada Judul Berita Berbahasa Inggris Menggunakan Algoritme Bidirectional LSTM." *Journal of Dinda: Data Science, Information Technology, and Data Analytics* 2.1, pp. 46-51, 2022