Journal of Computer Science and Information Technology
https://jcsitech-upiyptk.org/ojs
2022 Volume 8 Issue 1 Hal: 22-26 e-ISSN: 2502-1486

Application of Data Mining Clustering the Development of Covid-19 Using K-Medoids Method

Novia Gusmantoni✉
Universitas Putra Indonesia "YPTK" Padang, Indonesia
gusmantoninovia@gmail.com

Abstract

At the beginning of March, Indonesia was hit by the entry of the corona virus (covid) outbreak. Every day the cases of the spread of covid-19 in Indonesia continue to increase. The public is asked to carry out social distancing in order to break the chain of the spread of Covid-19 which is spread in various regions in Indonesia. Therefore, the data that has been accommodated is certainly a lot, from the data it can be seen that the patterns of determining the grouping of the spread of Covid-19 are based on test scores. Public. K-Medoids is a partitional clustering analytical method that aims to get a set of k-clusters among the data that is closest to an object in grouping a data. The results of the study of grouping the spread of the new covid-19 show that people come from various regions in Indonesia. Characteristics with a body temperature above 36.9 C and accompanied by fever and continuous cough show one of the symptoms of Covid-19.

Keywords: K-Medoids, Clustering, Determining, Data Mining, Covid-19.

1. Introduction

Corona (Covid-19) is a family of viruses that infect humans and animals and then spread from humans to humans through droplets of liquid that come from the mouth and nose when an infected person sneezes or coughs. This virus was first discovered in an animal and seafood market in Wuhan City [1]. Covid-19 is an infectious disease characterized by the presence of symptoms in the acute respiratory tract coronavirus 2 (severe acute respiratory syndrome coronavirus 2 or SARS-CoV-2). This virus is a large family of Coronaviruses that infect animals [2].

At first, this virus was called 2019-nCoV, later, experts from ICTV: International Committee on Virus Taxonomy called it the SARS-CoV-2 virus because it is the same as the cause of the SARS virus (SARS-CoVs). CoVs have become a major pathogen of respiratory diseases. This virus can infect other species and humans, from the common cold to more severe ones such as MERS and SARS [3].

In order to prevent the spread of the Covid-19 outbreak in Indonesia, there are pros and cons regarding the policy of locking or quarantining areas that have herd immunity. Indonesia has adopted a policy to implement large-scale social restrictions (PSBB). Applied in DKI Jakarta, followed by West Java, Banten, Central Java, and East Java [4].

The problem that occurs in handling COVID-19 cases in clustering areas exposed to COVID-19 is that this virus spreads so quickly that the data entered has not been updated every day, and the algorithm used results in the data being inputted. get it takes quite a long time to process. So it is difficult to determine the cause zoning for each province in Indonesia, while we all need data changes every day.

To overcome this problem, this study uses a data mining approach to cluster provinces in Indonesia. Data mining is a stage that can find useful information on large data stores. Data mining also performs data analysis to obtain clear relationships and conclusions about previously unknown relationships so that they can be reached [5][6].

To perform clustering in data mining, a clustering approach is usually used. Clustering is the process of grouping records, observations, or grouping classes into the same object. The difference between clustering and classification is that the target variable in clustering does not exist. When performing analysis, clustering is the first step in the data mining process [7]. Clustering performs object analysis by not finding information on known class labels. Clustering can be used to generate labels [8]. Grouping is a method that has unsupervised properties [9].

The algorithm used in this clustering approach is the K-Medoids algorithm. K-Medoids is an algorithm used to obtain medoids in a cluster which is the center point of the cluster. In K-Medoids it is found that k is a representative object so that the number of inequalities can be found from the object data set [10]. Data mining applications are made using the concept of software engineering which is the use of expertise possessed by manufacturing technicians to produce software that is affordable, powerful, and efficient so that it can be applied to machines [11]. The data mining application that will be made is web-based. A website is a collection of pages that contain stored information and can be accessed via the internet [12].

Accepted: 27-01-2022 | Revision: 19-01-2022| Publication: 31-01-2022 | doi: 10.35134/jcsitech.v8i1.29

2. Research Method

This study uses research stages starting from preliminary research to testing as depicted in Figure.1.

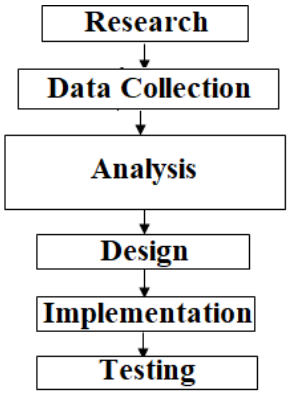

Fig. 1 Research Framework

Figure. 1 is the sequence carried out during the study. Beginning with preliminary research to identify problems in the object of research. Data collection was carried out by interviewing the relevant agencies and collecting data at the health service. The analysis was carried out on the data obtained by using the K-Medoids algorithm. Design models and interfaces of data mining applications. Implementing the model into a form program. perform tests on the functions that exist in data mining applications. The research methodology carried out on the K-Medoids process is shown in Figure. 2.

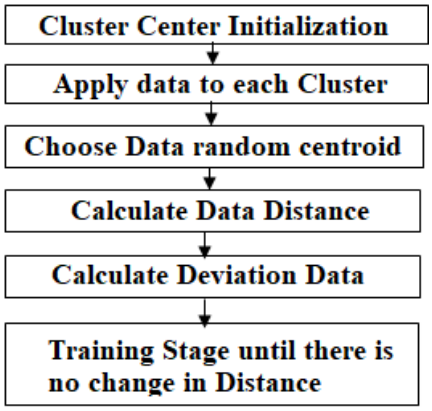

Fig. 2 Research Methodology

Figure. 2 describes the K-Medoids process starting from the initialization of the cluster center as many ask, namely the number of clusters that you want to form. This followed by allocating each data to the nearest cluster using the Euclidean Distance measure. Then randomly select one of the data in each cluster as a candidate for a new medoid. Then calculate the distance of each data in each cluster by taking a new medoid. 5. Calculate the total deviation (S) by calculating the new total distance - the old total distance. If S<0, then replace an object with a data cluster to get a new group of k objects as medoid. 3rd stage until there is no medoid change so that clusters and their respective cluster members are obtained.

3. Result and Discussion

In this study, the authors took data from the West Sumatra Provincial Health Office. Data at the West Sumatra Provincial Health Office is in the form of numbers, where the number is a value for the positive COVID-19 variable, recovered from COVID-19, and died from COVID-19 can be seen in Table. 1.

Table. 1 Cases of COVID-19 Provinces in Indonesia

NO	Nama Provinsi	Kasus Positif (X)	Kasus Sembuh(Y)	Kasus Meninggal(Z)
1	Dki Jakarta	319293	297641	5010
2	Jawa Barat	179982	150916	2132
3	Jawa Tengah	143419	91420	6016
4	Jawa Timur	123280	109978	8668
5	Sulawesi Selatan	52959	48088	809
6	Kalimantan Timur	49856	40451	1171
7	Bali	30821	27313	823
8	Riau	30277	28562	729
9	Sumbar	28236	26405	636
10	Banten	27675	19892	604
11	D.I. Yogya	25335	19096	599
12	Sumut	23108	19862	789
13	Kalsel	19978	17562	694
14	Papua	16135	8887	171
15	Sumsel	15253	13107	733
16	Sulut	14629	11057	488
17	Kalteng	12856	11516	338
18	Lampung	11603	9283	602
19	Sultenggara	9858	8892	190
20	Aceh	9403	7762	382
21	Sulteng	9346	6777	232
22	Kalut	8811	5757	133
23	Kepri	8514	7923	209
24	Ntb	8158	6585	336
25	Ntt	7517	4685	199
26	Papua Barat	7036	6392	119
27	Maluku	6769	5982	103
28	Babel	6278	5382	99
29	Jambi	5091	3849	77
30	Sulbar	4921	2941	93
31	Bengkulu	4743	4475	146
32	Gorontalo	4643	4212	124
33	Kalbar	4299	2931	32
34	Maluku Utara	3886	3111	110

Table. 1 describes the confirmation data that occurred. The data will later be processed for analysis using the clustering method with the following stages:

A. Determine the Min and Max Nilai Values

Looking for Min and Max Values on positive, recovered, ,died variables that need to be searched and can be seen in Table. 2.

Table. 2 Min dan Max Value

Himpunan	Positif Rate	Case Save	Death Case
Min	3886	2931	32
Max	319293	297641	8668

B. Data Normalization

This normalization is used as a reference for calculations in the k-medoid method, which can be seen in Formula. 1.

$$V^{I} = \frac{V - MINa}{MAXa - MINa} \quad (I)$$

Positif = (319293-3886)/(319293-3886) = 1,0000
Save = (297641-2931)/(297641-2931) = 1,0000
Death = (5010-32)/(8668-32) = 0,5764

do the calculations for all provinces, the results obtained can be seen in Table. 3.

Tabel 3. Normalization Date Result

Positif Rate	Saved Case (Y)	Death Case (Z)
1,0000	1,0000	0,5764
0,5583	0,5021	0,2432
0,4424	0,3003	0,6929
0.3785	0,3632	1,0000
0,1556	0,1532	0,0900
0,1457	0,1273	0,1319
0,0854	0,0827	0,0916
0,0837	0,0870	0,0807
0,0772	0,0797	0,0699
0,0754	0,0576	0,0662
0.0680	0.0549	0.0657
0,0609	0,0574	0,0877
0,0510	0,0496	0,0767
0,0388	0,0202	0,0161
0,0360	0,0345	0,0812
0.0341	0,0276	0,0528
0,0284	0,0291	0,0354
0,0245	0,0216	0,0660
0,0189	0,0202	0,0183
0,0175	0,0164	0,0405
0,0173	0,0131	0,0232
0,0156	0,0096	0,0117
0.0147	0.0169	0.0205
0,0135	0,0124	0,0352
0,0115	0,0060	0,0193
0,0100	0,0117	0,0101
0,0091	0,0104	0,0082
0.0076	0.0083	0.0078
0,0038	0,0031	0,0052
0,0033	0,0000	0,0071
0,0027	0,0052	0,0132
0,0024	0,0043	0,0107
0,0013	0,0000	0,0000
0,0000	0,0006	0,0090

C. Determine the desired (number of clusters)

At this stage it is necessary to randomly select a cluster which can be seen in Table. 4.

Tabel 4. Matriks Perbandingan Berpasangan

Kasus Positif (X)	Kasus Sembuh (Y)	Kasus Meninggal (Z)
0.0024	0.0043	0.0107
0.0013	0.0000	0.0000
0.0000	0.0006	0.0090

D. Calculating the distance

At this stage it is necessary to process Table 4 with table 3 normalizing the data in which there are formulas and tables that can be seen in Formula. 2.

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

COST1:

$\sqrt{(0.0024 - 1{,}0000)^2 + (0.0043 - 1{,}0000)^2 + (0.0107 - 0{,}5764)^2}$ = 1.5188

$\sqrt{(0.0024 - 0{,}5583)^2 + (0.0043 - 0{,}5021)^2 + (0.0107 - 0{,}2432)^2}$ = 0.7816

Perhitungan COST2:

$\sqrt{(0.0013 - 1{,}0000)^2 + (0.0000 - 1{,}0000)^2 + (0.0000 - 0{,}5764)^2}$ = 1.5263

$\sqrt{(0.0013 - 0{,}5583)^2 + (0.0000 - 0{,}5021)^2 + (0.0000 - 0{,}2432)^2}$ = 0.7884

Perhitungan COST3:

$\sqrt{(0.0000 - 1{,}0000)^2 + (0.0006 - 1{,}0000)^2 + (0.0090 - 0{,}5764)^2}$ = 1.5263

$\sqrt{(0.0000 - 0{,}5583)^2 + (0.0006 - 0{,}5021)^2 + (0.0090 - 0{,}2432)^2}$ = 0.78624.

In the proximity column the value is obtained from the MIN value (COST 1, COST 2, COST3). As for the Cluster column, the value is obtained from the value of the description column contained in the COST column. If the information value is in COST 1, then the Clustering value is 1. For more details, see Table. 5.

Table. 5 Cluster Result

Cost 1	Cost 2	Cost 3	Kedekatan	Cluster
1.5188	1.5263	1.5234	1.5188	1
0.7816	0.7884	0.7862	0.7816	1
0.8641	0.8746	0.8679	0.8641	1
1.1176	1.1288	1.1211	1.1176	1
0.2279	0.2353	0.2325	0.2279	1
0.2244	0.2334	0.2289	0.2244	1
0.1399	0.1493	0.1444	0.1399	1
0.1355	0.1445	0.1400	0.1355	1
0.1216	0.1304	0.1262	0.1216	1
0.1061	0.1149	0.1105	0.1061	1
0.0994	0.1085	0.1039	0.0994	1
0.1103	0.1206	0.1146	0.1103	1
0.0936	0.1040	0.0979	0.0936	1
0.0401	0.0455	0.0440	0.0401	1

0.0837	0.0948	0.0875	0.0837	1
0.0576	0.0680	0.0617	0.0576	1
0.0436	0.0532	0.0481	0.0436	1
0.0620	0.0732	0.0655	0.0620	1
0.0241	0.0324	0.0288	0.0241	1
0.0355	0.0466	0.0393	0.0355	1
0.0213	0.0311	0.0256	0.0213	1
0.0143	0.0208	0.0182	0.0143	1
0.0202	0.0298	0.0248	0.0202	1
0.0281	0.0393	0.0317	0.0281	1
0.0126	0.0226	0.0164	0.0126	1
0.0106	0.0177	0.0150	0.0106	1
0.0094	0.0154	0.0134	0.0094	1
0.0072	0.0130	0.0109	0.0072	1
0.0058	0.0065	0.0059	0.0058	1
0.0057	0.0074	0.0039	0.0039	3
0.0027	0.0143	0.0068	0.0027	1
0.0000	0.0116	0.0047	0.0000	1
0.0116	0.0000	0.0091	0.0000	2
0.0047	0.0091	0.0000	0.0000	3
	Calculation		6.0235	

In the process in Table 5, the results were obtained with the Cluster value in the Cluster column. From the results of clustering, the results obtained are 1 record in Cluster 3, 2 records in Cluster2, and 32 records in Cluster 1.

1. Provinces included in Cluster 1: DKI Jakarta, West Java, Central Java, East Java, South Sulawesi, East Kalimantan, Bali, Riau, West Sumatra, Banten, D.I. Yogya, North Sumatra, South Kalimantan, Papua, South Sumatra, North Sulawesi, Central Kalimantan, Lampung, Southeast Sulawesi, Aceh, Central Sulawesi, North Kalimantan, Riau Islands, NTB, NTT, West Papua, Maluku, Babylon, Jambi, Bengkulu, Gorontalo.
2. Provinces that are included in Cluster 2: West Kalimantan.
3. Provinces included in Cluster 3: North Maluku, Sulbar.

System testing will describe how a system runs. This testing stage contains the results of the execution program and an explanation of the program that has been made. This page contains all data on Covid 19 cases from all provinces in Indonesia, which can be seen in Figure. 3.

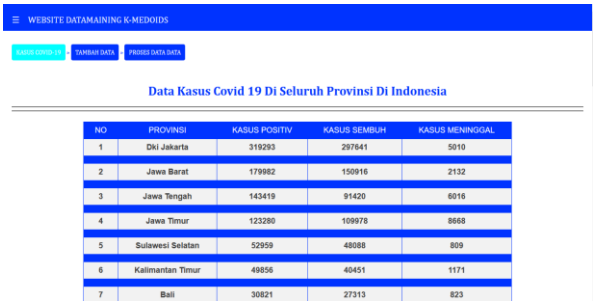

Fig. 3 Pages of all cases of Covid 19

The page that displays the results of the normalization process for the case min and max values can be seen in Figure. 4.

Fig. 4 Pages of Normalization Process Results

The result page for calculating the distance between objects and clusters of each object can be seen in Figure. 5.

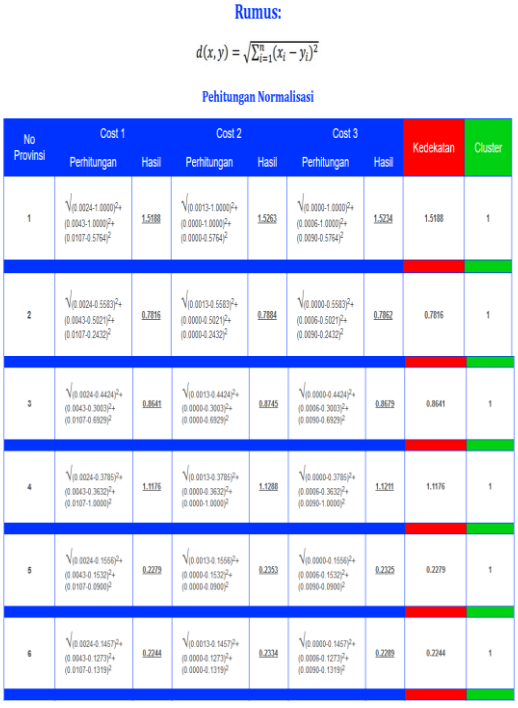

Fig. 5 Pages of Distance Calculation Results and Cluster Determination

Conclusion

From the research that has been carried out, it is obtained, namely the application of K-medoids data mining to COVID-19 cases in Indonesia can predict regional clusters affected by the COVID-19 virus. This K-medoids data mining application has been proven to be able to classify positive cases, recovered cases, and cases of death and can be used as a guide in determining dangerous areas from the number of COVID-19 cases based on the clusters that have been obtained.

References

[1] Rizkiana Prima, R., Yashintia Arien, E., & Sutikno Departemen Statistika, F. S. Analisis Cluster Virus Corona (COVID-19) di Indonesia pada 2 Maret 2020–12 April 2020 dengan Metode K-Means Clustering.

[2] Dwitri, N., Tampubolon, J. A., Prayoga, S., Zer, F. I. R., & Hartama, D. (2020). Penerapan algoritma K-Means dalam menentukan tingkat penyebaran pandemi COVID-19 di Indonesia. JurTI (Jurnal Teknologi Informasi), 4(1), 128-132.

[3] Rosyanti, L., & Hadi, I. (2020). Dampak psikologis dalam memberikan perawatan dan layanan kesehatan pasien COVID-19 pada tenaga profesional kesehatan. Health Information: Jurnal Penelitian, 12(1), 107-130.

[4] Junaedi, D., & Salistia, F. (2020). Dampak Pandemi Covid-19 Terhadap Pasar Modal Di Indonesia. Al-Kharaj: Jurnal Ekonomi, Keuangan & Bisnis Syariah, 2(2), 109-131.

[5] Sindi, S., Ningse, W. R. O., Sihombing, I. A., Zer, F. I. R., & Hartama, D. (2020). Analisis algoritma k-medoids clustering dalam pengelompokan penyebaran covid-19 di indonesia. JurTI (Jurnal Teknologi Informasi), 4(1), 166-173.

[6] Ardiansyah, A. H., Nugroho, W., Alfiyah, N. H., Handoko, R. A., & Bakhtiar, M. A. (2020, August). Penerapan Data Mining Menggunakan Metode Clustering untuk Menentukan Status Provinsi di Indonesia 2020. In Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) (Vol. 4, No. 3, pp. 329-333).

[7] Pramesti, D. F., Furqon, M. T., & Dewi, C. (2017). Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[8] Mandala, E. P. W., Yanto, M., & Putri, D. E. (2018). Aplikasi Pengelompokan Penjualan Dengan Clustering Data Mining Pada Toko Retail Kota Padang. Prosiding SISFOTEK, 2(1), 1-8.

[9] Putri, D. E. (2015). Metode Non Hierarchy Algoritma K-Means Dalam Mengelompokkan Tingkat Kelarisan Barang (Studi Kasus: Koperasi Keluarga Besar Semen Padang). Prosiding Senatkom, 1.

[10] Gunawan, I., Anggraeni, G., Rini, E. S., Putri, Y. M., & Zikri, Y. K. (2020). Klasterisasi provinsi di Indonesia berbasis perkembangan kasus Covid-19 menggunakan metode K-Medoids. SENATIK, 301-306.

[11] Kurniawan, A. (2012). Rekayasa Perangkat Lunak Aplikasi Penjualan Pada Toko Story Time Factory Outlet Menggunakan Pemrogram Java. Jurnal, Universitas Andalas, hlm, 3.

[12] Mandala, E. P. W. (2015). Web Programing, Project 1 epwm forum. Yogyakarta: Andi.