International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Analysis of Covid-19 Case Data in Classification of the Distribution in DKI Jakarta Using Algorithm Method K-Means Clustering

Ahmad Makih[1], Dr. Avinanta Tarigan[2]
[1, 2]Information Systems Management, Postgraduate Program, University of Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, East Java
Email address: [1]makihahmad @ gmail.com, [2]avinanta @ staff.gunadarma.ac.id

Abstract— *The application of data mining is now widely used in various aspects of life. Data mining is the collection of important information from a large data source. During the pandemic, the spread that occurred in the Covid-19 case had a pattern in groups at the same place and time and with relatively similar symptoms. In the spread of the Covid-19 case, data analysis applies an approach with object clusters that is in accordance with the Covid-19 spread case and in this study using the K-Means Clustering Algorithm method. The K-Means Clustering algorithm is basically applied to identify and summarize objects from a larger number so that it is easier to represent the characteristics of each group. The K-Means Clustering algorithm will classify objects so that the object will look for other objects that have a high level of closeness with similar characteristics or properties to the center point (centroid). This research will also analyze the data on the highest increase of an area that has been exposed to Covid-19 so that special handling can be carried out for the affected areas such as socialization, health assistance and spraying of disinfectants, etc. The results of the research using the K-Means Clustering Algorithm can classify areas that have low-high intensity with precise and accurate. This research is expected to be able to collect and map data carefully and to provide good, appropriate and periodic reports. So that it can be done for decision making related to handling Covid-19.*

Keywords— *Analysis, Data Mining, Covid-19, Algorithm, K-Means, Clustering.*

I. Introduction

The case of the spread of Covid-19 in Indonesia was officially announced by the government of the Republic of Indonesia in early March 2020. The first case was found in the Depok area, as many as 2 people who were exposed to Covid-19. Although some health observers and experts say that the spread of the Covid-19 virus may have entered Indonesia by the end of January 2020. Over time, the spread of Covid-19 continues to spread to the Greater Jakarta area including the entire DKI Jakarta area. Until now, the spread of Covid-19 in Indonesia has spread to 34 provinces. Based on a report in July 2020 from the DKI Jakarta Provincial Government regarding the Covid-19 Handling and Prevention Task Force, there were around 13,598 cases of people exposed to Covid-19 and 3,613 died. This has increased significantly compared to the June 2020 report of 11,276 related to the latest cases for people who tested positive for Covid-19.

Based on the results of the mapping of the spread of Covid-19 in Indonesia, the province of DKI Jakarta is the area exposed to the most Covid-19 out of 34 provinces. The Central Government and the President have taken many further actions related to the spread of the Covid-19 virus, such as Government Regulation Number 21 of 2020 concerning Large-Scale Social Restrictions in the Context of Accelerating Handling of Corona Virus Disease 2019 (Covid-19). Then several Presidential Decrees regarding the handling of Covid-19 in Indonesia, namely Presidential Decree Number 7 of 2020 concerning the Task Force for the Acceleration of Handling of Corona Virus Disease 2019 (Covid-19), Presidential Decree Number 11 of 2020 concerning the Determination of Public Health Emergencies for Corona Virus Disease 2019 (Covid-19), Presidential Decree Number 12 of 2020 concerning the Designation of Non-Natural Disaster for the Spread of Corona Virus Disease 2019 (Covid-19) as a National Disaster.

The policy of the Government of the Republic of Indonesia regarding the Covid-19 outbreak began to be significant when the issuance of Presidential Decree (Keppres). The total budget for this is IDR 405.1 trillion. On April 3, 2020, the President issued Presidential Regulation (Perpres) No. 54 of 2020 concerning Posture Changes in Details and the 2020 State Budget. This Presidential Regulation is a follow-up to Perppu No. 1 of 2020. The budget of several ministries is cut by Rp. 97.42 trillion. However, several Ministries experienced an increase in their budgets, such as the Ministry of Education and Culture from Rp. 36 trillion to Rp. 70 trillion; and the Ministry of Health from IDR 57 trillion to 76 trillion (https://kemlu.go.id). One way to find out the spread of Covid-19, especially in the DKI Jakarta area, is by implementing data grouping that has the same characteristics using the clustering technique.

Clustering algorithm works by grouping data objects (patterns, entities, events, units, results of observations) into a certain number of clusters. In other words, the Clustering Algorithm performs the separation / splitting / segmentation of data into a number of groups (clusters) according to certain characteristics.

In this study, researchers will take data from official information from the DKI Jakarta Provincial Government through the Jakarta Open Data regarding Covid-19 case data. The data will later be managed through the data pre-processing stage so that the data can be used as input for calculations. The output results can be a solution to help monitor the reporting system for the spread of Covid-19 cases in DKI Jakarta province.

The results of the research using the K-Means Clustering Algorithm can group areas that have low-high intensity with precise and accurate information accompanied by visual output in the form of clusters of distribution of areas affected by Covid-19. This research is expected to be able to collect and map data carefully and provide good, appropriate and periodic reports. So that it can be done for decision making related to handling Covid-19.

II. THEOROTICAL BASIC

Cluster Analysis is an unsupervised analysis of object mining methods (unsupervised analysis), while K-Means Cluster Analysis is a non-hierarchical cluster analysis method that seeks to partition existing objects into one or more clusters or groups of objects based on their characteristics, so that Objects that have the same characteristics are grouped into the same cluster and objects that have different characteristics are grouped into other clusters. The purpose of grouping is to minimize the objective function set in the clustering process, which basically tries to minimize variations within one cluster and maximize variation between clusters.

This cluster method includes sequential threshold, parallel threshold and optimizing threshold. Sequential threshold conducts grouping by first selecting one basic object that will be used as the initial cluster value, then all existing clusters in the closest distance to this cluster will join, then the second cluster is selected and all objects that have similarities to this cluster will be combined, and so on. so that several clusters are formed with all the objects contained therein. If given a set of objects, the K-Means Cluster Analysis algorithm will partition X into k clusters, each cluster has a centroid of the objects in the cluster. In the early stages of the K-Means Cluster Analysis algorithm, randomly selected k objects as centroids, then the distance between the object and the centroid is calculated using the euclidean distance, the object is placed in the closest cluster calculated from the center point of the cluster. Centroid is only defined when all objects are placed in the nearest cluster. The process of determining the centroid and placing the objects in the cluster is repeated until the value of the centroid converges (the centroid of all clusters does not change anymore).

III. RESEARCH METHOD

The stages in this research are calculating the K-Means Clustering. In general, the method in this study consists of several stages of the process, namely:

1. *Raw data*

The raw data used for the study is data on the Covid-19 DKI Jakarta cases per district on the official website of Open Data Jakarta as many as 44 data according to the number of sub-districts in DKI. Jakarta. The data used is data for June 2020 with the extension .xlsx.

2. *Data Pre-Processing*

Before processing data, it is necessary to carry out a data preprocessing process to facilitate extracting information from the results of data mining.

3. *Data Cleaning*

In this stage, what is done is to remove data outside the DKI Jakarta area in the Province Name or City Name column and to remove the Remarks column that has no value.

4. *The Clustering Process*

At this stage the main process will be carried out, namely the segmentation or grouping of data on the spread of Covid-19. The following is an application of the k-means algorithm with the assumption that the input parameter is the number of datasets as many as n data and the number of initialized k = 3 centroid according to the study. The number of data taken for the research is 44 to be used as an example of the application of the k-means algorithm. The experiment was carried out using the following parameters:

Number of clusters	: 3
Number of data	: 44
Number of attributes	: 5

5. K-Means Method

The application of the K-means method to the Clusterization of the Covid-19 Virus Distribution in DKI Jakarta was carried out in several stages, namely:

a. *Data retrieval*

The research data was obtained from the Open Data Jakarta website. The data obtained is adjusted to the database specifications.

b. *Dataset Selection*

Specifies the data to be processed based on the data date. The dataset used is cumulative, so the data selected is the total of all cases up to the selected date.

c. *Determining the Number of Clusters*

Users can determine how many clusters they want to form in each process. In this study, the minimum number of clusters is limited to 2 and the maximum number of clusters is 10. In this study, the number of clusters is 3.

d. *Determining the starting point of the cluster*

The center of the initial cluster, or also known as the initial centroid, is determined randomly based on the number of clusters and the amount of data to be processed.

e. *Distance of Each Data to the Center of the Cluster*

The distance between each data and each cluster is calculated using the Euclidean Distance (D) formula as presented in the Equation

Description:

D	= cluster distance
Xik	= data value (i, k)
Cjk	= centroid value (j, k)
n	= number of clusters

$$D_{(i,j)} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - C_{jk}\right)^2}$$

f. *Data Grouping Based on the Nearest Cluster*

Note which cluster has the closest distance to the data, then group the data into these clusters.

$$C_i = \frac{\Sigma d_i}{n_k}$$

g. *Calculating the Center for the New Cluster*

After all data has been grouped into clusters, calculate the new cluster center point by calculating the average distance between the data and the cluster center using the equation.

IV. Results and Discussion

Based on the results obtained from the iteration process, we can see that Cluster 1 is occupied by the Grogol Petamburan, Kali Deres, Kebon Jeruk, Palmerah, Senen, Kebayoran Lama, Pesanggrahan, Kelapa Gading, Pademangan, Penjaringan, Tanjung Priok districts. Cluster 1 is a cluster with a high rate of spread or being affected by Covid-19.

Then Cluster 2 is occupied by the districts of Kembangan, Taman Sari, Gambir, Johar Baru, Menteng, Sawah Besar, Kebayoran Baru, Pancoran, Cipayung, Jatinegara, Makassar, Pasar Rebo, Kep. Seribu Selatan, and Kep. Thousand North. Cluster 2 is a cluster with a moderate level of Covid-19 spread or impact. Then the last one is Cluster 3 which is occupied by the districts of Cengkareng, Tambora, Cempaka Putih, Kemayoran, Tanah Abang, Cilandak, Jagakarsa, Mampang Prapatan, Pasar Minggu, Setia Budi, Tebet, Cakung, Ciracas, Duren Sawit, Kramat Jati, Matraman, Pulo Gadung, Cilincing, and Koja. Cluster 3 is a cluster with a low Covid-19 spread.

From each cluster that has been described per sub-district, researchers are also able to identify the types of people affected by Covid-19 in the region.

TABLE I. Types of People Affected by Covid-19 Cluster 1

District	ODP	PDP	Positive	Recovered	Death
Grogol	1810	184	182	118	12
Kali Deres	2221	245	251	168	16
Kebon Jeruk	2500	279	148	95	12
Palmerah	2511	275	301	168	17
Senen	970	181	192	124	8
Kebayoran Lama	2404	354	330	145	20
Pesanggrahan	1002	114	125	82	6
Kelapa Gading	1396	228	256	130	14
Pademangan	1832	247	142	101	8
Penjaringan	891	117	104	61	7
Tanjung Priok	1180	116	189	77	18

TABLE II. Types of People Affected by Covid-19 Cluster 2

District	ODP	PDP	Positive	Recovered	Death
Kembangan	1465	262	237	109	18
Taman Sari	902	82	198	91	10
Gambir	1045	219	135	77	10
Johar Baru	2000	133	330	110	14
Menteng	1599	251	375	278	26
Sawah Besar	1604	247	125	69	15
Kebayoran Baru	1371	298	133	71	7
Pancoran	841	177	96	67	3
Cipayung	2044	304	221	111	21
Jatinegara	1638	132	78	44	7
Makasar	885	131	84	58	5
Pasar Rebo	1327	389	163	89	15
Kep. Seribu Selatan	1953	210	125	67	16
Kep. Seribu Utara	1435	130	112	65	10

TABLE III. Types of People Affected by Covid-19 Cluster 3

District	ODP	PDP	Positive	Recovered	Death
Cengkareng	1738	198	187	106	14
Tambora	1661	176	178	97	11
Cempaka Putih	1213	229	105	61	2
Kemayoran	1916	176	124	51	9
Tanah Abang	1735	244	341	209	28
Cilandak	859	125	207	105	13
Jagakarsa	1283	215	194	110	14
Mampang Prapatan	907	86	98	50	7
Pasar Minggu	1336	257	178	106	13
Setia Budi	789	131	109	61	5
Tebet	1698	182	173	105	13
Cakung	1549	221	221	106	11
Ciracas	1964	454	139	89	16
Duren Sawit	1351	299	202	130	11
Kramat Jati	2305	158	243	129	14
Matraman	2179	164	270	156	16
Pulo Gadung	2580	416	444	247	31
Cilincing	572	5	11	10	0
Koja	329	16	1	1	0

Based on the results of the analysis, it can be seen that the most sub-districts in DKI Jakarta are in Cluster 3, which is 19 Districts, followed by Cluster, namely 14 Districts and Cluster 1, which is 11 Districts.

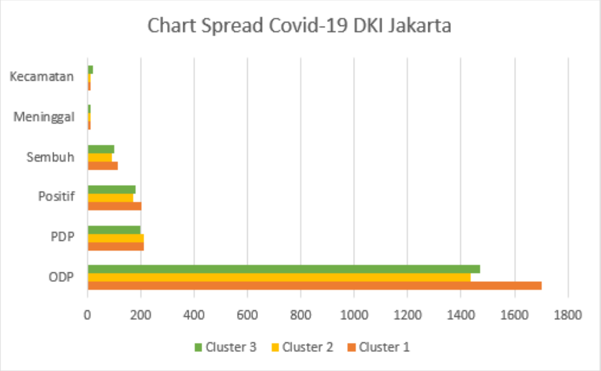

Fig. 1. The result of Spread Covid-19 DKI Jakarta

TABLE IV. Result Analysis Clustering

Result Cluster 1				
ODP	PDP	Positive	Recovered	Death
18717	2340	2220	1269	138
11	11	11	11	11
1701.545	212.7273	201.8182	115.3636	12.54545
Result Cluster 2				
ODP	PDP	Positive	Recovered	Death
20109	2965	2412	1306	177
14	14	14	14	14
1436.357	211.7857	172.2857	93.28571	12.64286
Result Cluster 3				
ODP	PDP	Positive	Recovered	Death
27964	3752	3425	1929	228
19	19	19	19	17
1471.789	197.4737	180.2632	101.5263	13.41176

Cluster 1 has ODP of 18,717 people, PDP numbering 2,340 people, positively infected with Covid-19 as many as 2,220 people, Cured as many as 1,269 people and died as many as 138 people. Cluster 1 is spread over 11 districts, namely Grogol Petamburan, Kali Deres, Kebon Jeruk, Palmerah, Senen, Kebayoran Lama, Pesanggrahan, Kelapa Gading, Pademangan, Penjaringan, Tanjung Priok. Then the average ODP Cluster is 1,701 people per district, PDP with an average of 212 people per district, a positive average of 201 people per district, a cure rate of 115 people per district and 12 people per district.

Cluster 2 has ODP of 20,109 people, PDP of 2,965 people, Positive infected with Covid-19 as many as 2,412 people, Healed as many as 1,306 people and died as many as 177 people. Cluster 2 is spread over 14 Districts, namely Kembangan, Taman Sari, Gambir, Johar Baru, Menteng, Sawah Besar, Kebayoran Baru, Pancoran, Cipayung, Jatinegara, Makasar, Pasar Rebo, Kep. Seribu Selatan, and Kep. Thousand North. Then the average ODP Cluster is 1,436 people per district, PDP with an average of 211 people per district, a positive average of 172 people per district, a cure rate of 93 people per district and 12 people per district.

Cluster 3 has an ODP of 27,964 people, PDP totaling 3,752 people, positively infected with Covid-19 as many as 3,425 people, Healed 1,929 people and died as many as 228 people. Cluster 3 is spread across 19 Districts, namely Cengkareng, Tambora, Cempaka Putih, Kemayoran, Tanah Abang, Cilandak, Jagakarsa, Mampang Prapatan, Pasar Minggu, Setia Budi, Tebet, Cakung, Ciracas, Duren Sawit, Kramat Jati, Matraman, Pulo Gadung areas. , Cilincing, and Koja. Then the average ODP Cluster is 1,471 people per sub-district, PDP with an average of 197 people per sub-district, a positive average of 180 people per district, a cure rate of 101 people per district and 13 people per district.

V. Conclusion

A. *Conclusion*

Based on the results of research conducted by researchers, the following conclusions can be drawn:

1. The application of the K-means clustering algorithm results in the analysis into three categories, namely low, medium and high according to the number of odp, pdp, positive, recovered and died.
2. Cluster 1 is spread over 11 districts, namely Grogol Petamburan, Kali Deres, Kebon Jeruk, Palmerah, Senen, Kebayoran Lama, Pesanggrahan, Kelapa Gading, Pademangan, Penjaringan, Tanjung Priok. Then Cluster 2 is spread across 14 Districts, namely Kembangan, Taman Sari, Gambir, Johar Baru, Menteng, Sawah Besar, Kebayoran Baru, Pancoran, Cipayung, Jatinegara, Makasar, Pasar Rebo, Kep. Seribu Selatan, and Kep. Thousand North. Then Cluster 3 is spread across 19 Districts, namely Cengkareng, Tambora, Cempaka Putih, Kemayoran, Tanah Abang, Cilandak, Jagakarsa, Mampang Prapatan, Pasar Minggu, Setia Budi, Tebet, Cakung, Ciracas, Duren Sawit, Kramat Jati, Matraman, Pulo. Gadung, Cilincing, and Koja. This can be an input or recommendation to the Covid-19 task force for sub-districts that are in certain categories. So that the deployment cluster can be minimized and even anticipated quickly and measurably.

B. *Suggestions*

Given that there are still many things that cannot be implemented from this study, the authors consider several suggestions, namely:

1. The results of clustering that are formed can be developed into a knowledge base for a provincial mapping decision support system with data adapted to each region.
2. Combining with other methods or approaches in order to get better research results like Hierarchical clustering, Partitional Clustering, Principal Component Analysis, Singular Value Decomposition, and Independent Component Analysis.
3. This research can be developed by comparing the clustering results with other research which also discusses the covid-19 case.

References

[1] Dinas Kesehatan, "Data Rekap Kasus Covid19 Per Kelurahan di Provinsi DKI Jakarta Bulan Juni 2020", Januari, 2021. [Daring]. Tersedia: https://data.jakarta.go.id/dataset/data-rekap-harian-kasus-covid19-per-kelurahan-di-provinsi-dki-jakarta-bulan-juni-2020. [diakses 30 Januari 2021].

[2] M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran", Iran, 2020.

[3] Solichin, A., and Khairunnisa, K., "Klasterisasi persebaran virus Corona (Covid-19) di DKI Jakarta menggunakan metode K-Means", Fountain of Informatics Journal, Jakarta, 2020.

[4] Zarikas, V and Zervas, E., "Clustering analysis of countries using the COVID-19 cases dataset", Data in Brief, 2020.

[5] M. Habibi dan P. W. Cahyo, "Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform", Indonesia, 2020.

[6] L. Maulida, "Penerapan Datamining dalam Mengelompokkan Kunjungan Wisatawan ke objek Wisata Unggulan di Prov. DKI Jakarta dengan K-Means", Jakarta, 2018.

[7] Nurhayati, Busman, and V. Amrizal, "Big data analysis using hadoop framework and machine learning as decision support system (DSS) (case study: knowledge of Islam mindset)", 6th International Conference on Cyber and IT Service Management (CITSM), 2018.

[8] Asroni and R. Adrian, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang", Jawa Tengah, 2015.

[9] Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method", Jakarta, 2017.

[10] Maghdi, Ghafoor, "A Smartphone enabled Approach to Manage COVID-19 Lockdown and Economic Crisis.", Iraq, 2020.

[11] Muhamadan Ali, "Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering", Jawa Tengah, 2020

[12] Gupta, Siddiqui, "Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis", India, 2020

[13] Chinchorkar Satish, "Dening Covid 19 containment zones using Kmeans dynamically", India, 2020

[14] Zhang, Lin. "Generalized K-Means in GLMs with Applications to the Outbreak of COVID-19 in the United States", China, 2020

[15] Yeasmin, Banik, "Impact of COVID-19 pandemic on the mental health of children in Bangladesh: A cross-sectional study", Bangladesh, 2020.