JurnalMantik
Volume 4, Number 4, Feb 2021, pp. 2336-2341 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Grouping of Covid-19 Affected Areas in Bogor City Using The K-Means Algorithm

Zulia Imami Alfianti[1], Sugiono[2], Mochammad Abdul Azis[3], Ahmad Fauzi[4]

[1,2,3,4]Fakultas Teknologi Informasi, Universitas Bina SaranaInformatika, Jakarta, Indonesia

E-mail: zulia.zim@bsi.ac.id, Sugiono.sgx@bsi.ac.id,mochamad.mmz@bsi.ac.id,ahmad.aau@bsi.ac.id

ARTICLEINFO

Article history:
Received: 01/12/2020
Revised: 10/12/2020
Accepted: 03/01/2021

Keywords:
Datamining, Clustering, Covid-19, Algorithm, K-Means

ABSTRACT

Clustering plays an important role in processing big data, making predictions and overcoming anomalies in data, identical characteristics in data sets are grouped using iterative techniques. Because data is always evolving from day to day, very large data sets with little can be identified into interesting patterns by grouping, special methods are needed to handle it. In December 2019 there was an outbreak of acute respiratory syndrome caused by coronavirus 2 infection that occurred in Wuhan and on February 12, 2020, the World Health Organization officially named the disease Corona Virus 2019 (Covid 19). This research will conduct clustering of areas affected by Covid 19 in the City of Bogor. The clustering was done using the K-Means method and dividing the data into 3 clusters, namely the low-impact cluster, the medium-impact cluster and the high-impact cluster. The results showed that from 68 urban villages in the city of Bogor, 45% of the area was in the low-affected category, 35.29% of the area was in the medium-affected category and 19.12% of the area was in the high-affected category.

1. Introduction

Continuous data storage can have an impact on the occurrence of data accumulation on a large scale, this data will not be useful without processing or grouping to determine a value for a document[1]. The most widely accepted definition of data mining is converting raw data into useful data or information.

The process of identifying groupings or clusters in multidimensional data based on some measure of similarity[2]. It takes an analytical method with a specific goal to be very important. The concept of data mining is one of the important tools in information management because the amount of information is getting bigger. Data mining itself is often called clustering as Knowladge Discovery in Database (KDD) is an activity that includes collecting, using historical data to find regularities, patterns of relationships in large data sets [3]. The results of data mining can be used for decision making in the future.

This study uses the K-Menas algorithm which is one of the methods that aims to simplify grouping, a grouping problem that aims to minimize multiple errors. The purpose of this study is to apply K-Means in clustering data on the impact of covid-19 in the Bogor city district on November 5, 2020. This can be input to the state government which has the highest priority in handling Covid-19 in certain areas.

2. Method

2.1 Data Mining

Data mining is a process of looking for new correlations, patterns and trends by digging a large number of data repositories using statistics and mathematical techniques. The current development of data playing is so fast because it has the ability to explore useful patterns and trends that come from existing database. Many companies have spent billions of rupiah to collect large amounts of data but have not benefited from it. Even though these data contain a number of valuable information, their existence is still hidden in the data repository [4].Data mining, also known as pattern recognition, is a method used for data processing to find hidden patterns from a set of processed data. The data processed with data mining will then produce information or new knowledge that comes from old data which will be useful in making decisions in the future. [3].

2.2 *K-Means*

The K-Means algorithm is one of the simplest unsupervised learning algorithms that solves problems, as a K-Means clustering method, groups data based on their proximity to each other (preeti), a method that

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

performs the modeling process without supervision and belongs to the non-hierarchical data collection method or a method of partitioning data into two or more groups. Where the data in one group have the same characteristics as each other and have different characteristics from the data in the group [5]. Testing the components of the data population and classifying the data into a defined cluster based on the minimum distance between population components and each cluster center [6][7]. The K-Means algorithm classifies data and information from each centroid value in each cluster [8]. The average or cluster center is formed by random selection of an object. By comparing most of the similarities, other objects assign to the cluster. For each data vector this algorithm calculates the distance between the data vector and each cluster centroid using the eucledian formula:

$$d_{ik} = \sqrt{\sum_j^m (c_{ij} - c_{kj})^2} \quad \text{.....(1)}$$

Where:
d_{ik} = distance
c_{ij} = cluster centroid
c_{kj} = data

2.3 Corona Virus *Diseaster* 2019 (COVID-19)

Coronaviruses (CoVs) are a group of viruses that can cause respiratory, enteric, hepatic, and neurological system diseases. The rate of infection with this virus varies greatly between humans and animals. symptoms of this viral infection include fever, fatigue, and cough. of several types of diseases caused by the corona virus, this type of NCoV which causes severe infection in humans.

The data used in this study came from the Covid-19 team report published on the covid19kotabogor.co.id news page, recap on November 5, 2020. The process of processing the data that has been obtained is by compiling patterns and grouping the patterns into several clusters. with the k-means method. The results of this processing obtained patterns and conclusions so that they are useful for other research.

To complement the basic knowledge in conducting this research, literature studies are used that come from books, journals and related research.

3. Results and Discussion

The framework for the K-Means method begins by determining the number of clusters desired and determining the data used as the initial centroid point of this process is carried out randomly. The next step is to use the eucledian formula. Calculate the distance of each data with the centroid, then the cluster membership will be found based on the smallest distance in each cluster.

$$d_{ik} = \sqrt{\sum_j^m (c_{ij} - c_{kj})^2} \quad \text{.....(2)}$$

d_{ik} = distance
c_{ij} = cluster centroid
c_{kj} = data

In the next step, namely by calculating the average value of themembers cluster to be used as the next cluster centroid point. If the centroid point is different from the previous centroid point, the process of determining cluster members is carried out again by recalculating the distance of each data from the centroid point. The steps mentioned above are carried out until there is no change in the centroid point. If no changes are found in the centroid, the calculation process is complete. These steps can be described as follows.

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

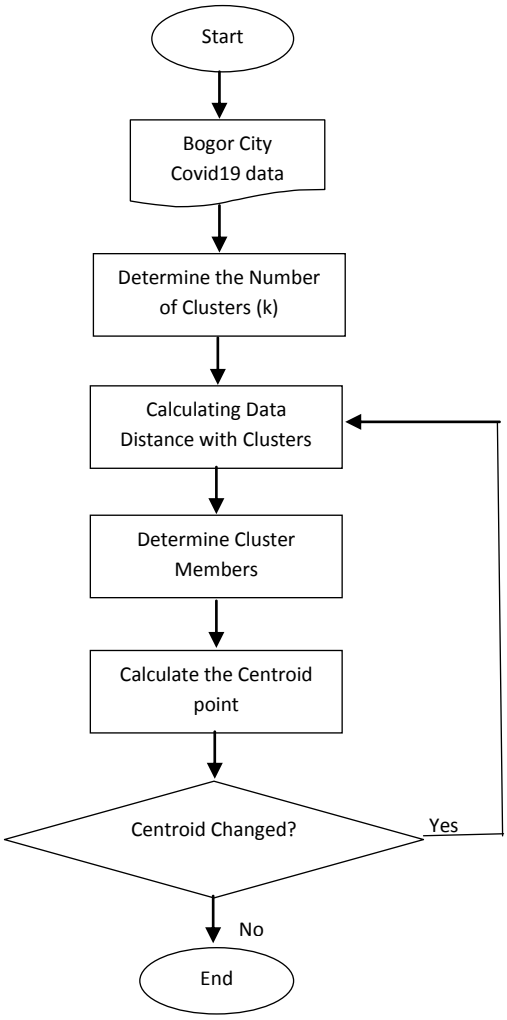

Fig 1. Clustering Steps Using K-Means Method

The clustering steps use the K-Means method: The process of grouping areas affected by covid-19 in Bogor City using the K-Means steps begins by determining the number of initial clusters and centroids. Bogor city consists of 68 sub-districts and is marked with a code of K1 to K68. Next, we determine the 3 desired clusters and use K1, K38 and K61 as the initial centroid points.

TABLE 1
DATA OF VILLAGE AFFECTED BY COVID-19, BOGOR CITY, 5 NOVEMBER 2020

Kode	Kelurahan	Kecamatan	Positif	Sembuh	Penanganan	Meninggal
K1	Balungbangjaya	Bogor Barat	11	11	0	0
K2	Bubulak	Bogor Barat	32	26	6	0
K3	Cilendek Barat	Bogor Barat	60	55	3	2
K4	Cilendek Timur	Bogor Barat	42	39	2	1
…	…	…	…	…	…	…
K38	KebonKelapa	Bogor Tengah	17	14	3	0
…	…	…	…	…	…	…
K61	Kedungbadak	Tanah Sareal	83	74	4	5
K62	Kedungjaya	Tanah Sareal	26	22	4	0
K63	Kedungwaringin	Tanah Sareal	30	27	2	1
K64	Kencana	Tanah Sareal	59	45	12	2
K65	Mekarwangi	Tanah Sareal	61	53	7	1
K66	Sukadamai	Tanah Saream	46	34	12	0
K67	Sukaresmi	Tanah Sareal	37	32	4	1
K68	Tanah sareal	Tanah Sareal	30	23	4	3

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

From the data above, it is determined that 3 clusters are looking for and use data K1, K38, and K61 as the initial centroid of the cluster, marked as C1, C2, and C3.

TABLE 2.
INITIAL CENTROID

Centroid Awal	Atribut 1	Atribut 2	Atribut 3	Atribut 4
C1	11	11	0	0
C2	17	14	3	0
C3	83	74	4	5

In the next step, calculate the distance between each data point and the centroid using the eucledian formula.

$d_{K1-C1} = \sqrt{(11-11)^2 + (11-11)^2 + (0-0)^2 + (0-0)^2}$ = 0

$d_{K1-C2} = \sqrt{(11-17)^2 + (11-14)^2 + (0-3)^2 + (0-0)^2}$ = 7,35

$d_{K1-C3} = \sqrt{(11-83)^2 + (11-74)^2 + (0-4)^2 + (0-5)^2}$ = 95,89

TABLE 3.
RESULTS OF CLUSTER DISTANCE CALCULATIONS WITH CENTROID

Kode	C1	C2	C3
K1	0.00	7.35	95.89
K2	26.50	19.44	70.24
K3	65.95	59.45	30.00
K4	41.83	35.38	54.09
…	…	…	…
K38	7.35	0.00	89.34
…	…	…	…
K61	95.89	89.34	0.00
K62	19.03	12.08	77.32
K63	24.94	18.44	70.98
K64	60.07	53.01	38.60
K65	65.68	58.94	30.82
K66	43.57	36.36	55.30
K67	33.67	26.94	62.42
K68	23.02	16.12	73.58

From the results of the above calculations, it can be determined that the cluster membership can be determined by reading the smallest distance between the data and the centroid so that a cluster membership is produced with 12 members C1, 44 C2 and 12 C3 with details as follows:

C1 = K1, K9, K10, K16, K18, K23, K25, K31, K35, K39, K41, dan K42.

C2 = K2, K4, K5, K6, K7, K8, K11, K12, K13, K15, K17, K19, K20, K21, K22, K24, K26, K27, K28, K29, K30, K32, K33, K34, K36, K37, K38, K40, K44, K46, K47, K48, K49, K51, K52, K53, K54, K55, K59, K62, K63, K66, K67, dan K68.

C3 =K3, K14, K43, K45, K50, K56, K57, K58, K60, K61, K64, dan K65.

The next step is to calculate the average value of cluster members so that the centroid value is obtained as follows:

TABLE 4.
CENTROID VALUE OF CALCULATIONS IN THE FIRST ITERATION

Kluster	Atribut 1	Atribut 2	Atribut 3	Atribut 4
C1	8.75	6.25	1.75	0.75
C2	28.25	22.57	4.82	0.86
C3	65.25	54.17	9.00	2.08

Based on the results of the calculation of the centroid value above, there is a difference between the cantroid value generated in the first iteration and the previous centorid value. So the next iteration is needed by recalculating the distance between the data and the last centroid value, then the cluster membership is determined again through the smallest distance between the data and the cluster and calculating the next cantroid value.

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

TABLE 5.
CALCULATION OF CENTROID VALUE

Kluster	Atribut 1	Atribut 2	Atribut 3	Atribut 4
CENTROID ITERASI 0				
C1	11	11	0	0
C2	17	14	3	0
C3	83	74	4	5
CENTROID ITERASI 1				
C1	8.75	6.25	1.75	0.75
C2	28.25	22.57	4.82	0.86
C3	65.25	54.17	9.00	2.08
.....				
CENTROID ITERASI 4				
C1	14.73	11.57	2.50	0.67
C2	34.04	27.32	5.72	1.00
C3	64.46	52.92	9.46	2.08
.....				
CENTROID ITERASI 6				
C1	15.10	11.71	2.74	0.65
C2	34.38	27.79	5.54	1.04
C3	64.46	52.92	9.46	2.08
CENTROID ITERASI 7				
C1	15.10	11.71	2.74	0.65
C2	34.38	27.79	5.54	1.04
C3	64.46	52.92	9.46	2.08

The calculation process is completed in the 7th iteration after there is no difference between the centroid value generated in the 7th iteration calculation and the centroid value in the 6th iteration calculation. In this last iteration, the number of cluster membership for C1 was 31, C2 was 24 and C3 was 13, with the following details.:

C1 = K1, K9, K10, K16, K17, K18, K19, K20, K21, K23, K24, K25, K26, K27, K29, K31, K32, K33, K35, K36, K37, K38, K39, K40, K41, K42, K46, K47, K49, K51, dan K59.

C2 = K2, K4, K5, K6,K7, K8, K11, K12, K13, K15, K22, K28, K30, K34, K48, K52, K54, K55, K62, K63, K66, K67, dan K68.

C3 = K3, K14, K43, K44, K45, K50, K56, K57, K58, K60, K61, K64, dan K65.

After the process of clustering the affected areas of Covid 19 in the city of Bogor using the k-means method and the final centroid and the number of cluster membership, it can be concluded that there are three clusters, namely the cluster of low affected areas with low centroid value (C1), the cluster of affected areas with moderate value. medium centroid (C2), and high impact area cluster with the highest centroid value (C3). Of the 68 sub-districts affected by Covid 19 in the city of Bogor, 31 sub-districts are in the low category, 24 urban villages are included in the medium category and 13 urban villages are included in the high category.

4. Conclusion

After the clustering process of areas affected by Covid19 in the city of Bogor using the K Means method, it can be concluded that of the 68 sub-districts in the city of Bogor 45% of the area is in the low affected category, 35.29% of the area is in the medium affected category and 19.12% area is included in the high affected category. In low-affected areas, it was found that the average number of positive cases was 15.10, the average number of cured was 11.71, the average treatment was 2.47 and an average death was 0.65. In moderately affected areas, an average number of positive cases was found as much as 34.38, the average number of cured was 27.79, in the average treatment was 5.54 and died on average 1.04. Whereas in the highly affected areas, an average number of positive cases was found as much as 64.46, the average number of cured was 52.92, in treatment an average of 9.46 and an average death of 2.08. Given the Covid-19pandemic that is still ongoing, further studies are needed regarding the areas affected by Covid19 in the city of Bogor using the latest data.

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

5. References

[1] D. T. Kusuma, N. Agani, J. Ticom, and V. No, "Prototipe Komparasi Model Clustering Menggunakan Metode K-Means Dan FCM untuk Menentukan Strategi Promosi : Study Kasus Sekolah Tinggi Teknik-PLN Jakarta," vol. 3, no. 3, pp. 1–10, 2015.

[2] S. Clustering and D. A. N. K. Neightbor, "Klasifikasi data multidimensi menggunakan subtractive clustering dan k-nearest neightbor (," vol. 10, no. 1, pp. 11–19, 2012.

[3] A. K. Wardhani, "Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan," *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.

[4] F. Nasari and C. J. M. Sianturi, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat," *CogITo Smart J.*, vol. 2, no. 2, p. 108, 2016, doi: 10.31154/cogito.v2i2.19.108-119.

[5] D. E. Putri, S. Kom, and M. Kom, "METODE NON HIERARCHY ALGORITMA K-MEANS DALAM MENGELOMPOKKAN TINGKAT KELARISAN BARANG (STUDI KASUS : KOPERASI KELUARGA BESAR SEMEN PADANG)," vol. 1, no. Senatkom, 2015.

[6] B. M. Metisen and H. L. Sari, "ANALISIS CLUSTERING MENGGUNAKAN METODE K-MEANS DALAM PENGELOMPOKKAN PENJUALAN PRODUK PADA SWALAYAN FADHILA," vol. 11, no. 2, pp. 110–118, 2015.

[7] S. Agustina, D. Yhudo, H. Santoso, N. Marnasusanto, A. Tirtana, and F. Khusnu, "CLUSTERING KUALITAS BERAS BERDASARKAN CIRI FISIK MENGGUNAKAN METODE K-MEANS Algoritma," *Clust. K-Means*, 2012.

[8] Sugiono, S. Nurdiani, S. Linawati, R. A. Safitri, and E. P. Saputra, "Pengelompokan Perilaku Mahasiswa Pada Perkuliahan E-Learning dengan K-Means Clustering," *J. Kaji. Ilm. Univ. Bhayangkara Jakarta Raya*, vol. 19, no. 2, pp. 126–133, 2019.