Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11740
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data Mining using clustering method to predict the spread of Covid 19 based on screening and tracing results

Allwin M. Simarmata[1)*], Riwanto Manik[2)], Ourent Chrisin Renatta Simanjorang[3)], Dymas Ferpian Purba[4)]
[1,2,3,4)]Universitas Prima Indonesia, Indonesia
[1)] allwinsimarmata@unprimdn.ac.id, [2)] riwantomanik74@gmail.com, [3)]parknata38@gmail.com, [4)]dimaspurba12@gmail.com

Submitted : Aug 26, 2022 | **Accepted** : Aug 27, 2022 | **Published** : Oct 3, 2022

Abstract: Coronavirus is a virus that causes disease in humans and animals. The virus was discovered in Wuhan, China in December 2019. Initially, it was suspected to be pneumonia, with general symptoms similar to the flu. However, unlike influenza, coronaviruses can progress rapidly, leading to more severe infections and organ failure. The number of COVID-19 sufferers in Indonesia is increasing every month. Anticipation and reducing the number of people infected with the coronavirus in Indonesia have been carried out in all regions. Including providing policies that limit activities outside the home. Indonesia has a very wide area, so it is necessary to classify the spread of Covid-19 based on regions or regions in Indonesia. This grouping provides a central point for the spread of Covid-19 pandemic cases in Indonesia. In testing data using data mining, data mining allows users to find knowledge in databases that were previously unknown to the user. By using the Clustering technique and the K-Means algorithm to predict the spread of COVID-19 based on the results of screening and tracing. The Clustering method produces 3 clusters, Cluster 0 with a medium category with a total of 6 regions, Cluster 1 with a low category with a total of 3 regions, and Cluster 2 with a high cluster with a total of 7 regions, with a DBI value of -0.784.

Keywords: Spread, Covid-19, Data Mining, Classification, K-Means Algorithm

INTRODUCTION

Coronavirus is a virus that causes disease in humans and animals. The virus found in Wuhan, China in December 2019, was later named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2) and caused Corona Disease 2019 (Covid-19) (Gayatri & Hendry, 2021). The initial appearance was suspected to be pneumonia, with general symptoms similar to the flu. However, unlike influenza, the coronavirus can progress rapidly, leading to more severe infections and organ failure. This emergency occurs in patients with previous health problems or congenital diseases (Mona, 2020). Currently, in 2020, the evolution of the transmission of this virus is quite significant because it has spread throughout the world and all countries are feeling the impact, including Indonesia. Anticipation and reducing the number of people infected with the coronavirus in Indonesia have been carried out in all regions. Including providing policies that limit activities outside the home, stop school activities, work from home, and even worship activities. This has become a government guideline based on considerations that of course have been researched (Yunus & Rezki, 2020). Indonesia has a very wide area, so it is necessary to classify the spread of Covid-19 based on regions or regions in Indonesia. This grouping provides a central point for the spread of Covid-19 pandemic cases in Indonesia (Salsabila & Intani, 2021).

Data mining is the process of extracting large amounts of data and information that were previously unknown, but understandable and useful for making very important decisions. Data mining allows users to discover knowledge in databases that were previously unknown to the user. Data mining is the process of finding useful information automatically in big data storage (Watratan et al., 2020). One of the algorithms that will be used in this research is the K-Means algorithm. K-Means is one of the clustering algorithms included in unsupervised learning, used to group data into several with a partition system. This algorithm accepts input in the form of data without class labels. In the K-Means algorithm, the computer first gets data whose class is unknown and then groups them. The K-Means algorithm was used for several studies, such as research (Sari et al., 2020) conducting clustering of the spread of tuberculosis in Karawang Regency, research (Akramunnisa &

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fajriani, 2020) conducting clustering for the distribution of the unemployment rate in South Sulawesi Regency/City, and research that conducted (Gustientiedina et al., 2019) conducted clustering of drug data at Pekanbaru Hospital. The K-Means algorithm has advantages and disadvantages. One of the shortcomings of the K-Means Algorithm is the initialization of the initial centroid value which is random. It is very sensitive to the final results of clustering (Mursalim et al., 2021). Behind the shortcomings, the K-Means algorithm has advantages, namely being able to group large objects and object outliers very quickly so that it speeds up the grouping process (Bastian et al., 2018). So in this study, we will implement the K-Means algorithm to determine the level of spread of Covid-19 by using the results of screening and tracing. It is hoped that this research will help medical personnel and the government in tackling the spread of COVID-19.

LITERATURE REVIEW

The research conducted by Bu'ulolo & Purba in his research used the k-medoids algorithm to form clusters of the covid-19 spread zone, in his research it produced 3 clusters, namely cluster 1 with a red zone, cluster 2 with a yellow zone and cluster 3 with a red zone. green zone (Bu'ulolo & Purba, 2021). The research of Sari et al, applied the K-Means algorithm to determine the spread of tuberculosis in Karawang Regency. This study uses data obtained from the Health Office in 2018. The results of the study resulted in 3 clusters, namely cluster 0 consisting of 7 sub-districts, cluster 1 consisting of 9 sub-districts, and cluster 2 consisting of 14 districts with an SSE value of 2.4402 and Silhouette 0.5629. The area with a high number of tuberculosis cases is the cluster 0 region (Sari et al., 2020). A study by Akramunnisa & Fajriani, in their research using the K-Means algorithm to classify the unemployment rate in the city of South Sulawesi. The results of the study resulted in 2 clusters with low and high categories. cluster 0 is a high cluster, there are 3 areas of high unemployment, and cluster 1 is a low cluster, and there are 21 areas of the lowest unemployment (Akramunnisa & Fajriani, 2020).

In a study conducted by Gustientiedina et al, to collect drug data in Pekanbaru Hospital. The dataset used is from usage reports and a list of drug supplies at Pekanbaru Hospital in 2017 with the parameters of the drug name, unit, and a number of drug items every month for 1 year. The results of clustering produce 3 clusters. In cluster 1, the average use of drugs is 18,000, in cluster 2, the average need for drugs is 18,000 to 70,000, while in cluster 3, for high use, the need for drugs is on average 70,000 (Gustientiedina et al., 2019). Research conducted by Asroni et al, grouped prospective new student data with the K-Means algorithm. The study resulted in 4 groups, namely group 0 majoring in nursing science with a total of 15% registrants, group 1 majoring in medical education with a total of 33% registrants, group 2 majoring in dentistry with a total of 30% registrants, and group 3 majoring in physicians with total registrants. 22% (Asroni et al., 2018). Research conducted by Dwitri et al, carried out the level of pandemic spread in Indonesia using the K-Means algorithm. Data were obtained from the Indonesian Ministry of Health on May 9, 2020. Generated 3 clusters. Cluster 1 is a large spread because it has a number of positive cases of 5056 and 427 cases died. Cluster 2 is a medium spread because it has a number of positive cases of 4525 and 348 cases of death. Meanwhile, cluster 3 has a small spread because it has a number of positive cases of 4043 and 184 cases of death (Dwitri et al., 2020). Research conducted by Rimelda Adha is to compare the DBSCAN and Kmeans algorithms to cluster COVID-19 cases in the world. The test results using the DBSCAN algorithm obtained the best cluster with an Eps value of 0.2 and Minpts 3, while the K-Means obtained the best cluster with a total of k as many as 8 and has been validated with a Silhouette Index of 0.6902. Based on testing the k-means algorithm has a better cluster validation than DBSCAN (Adha et al., 2021).

In a previous study conducted by (Virantika et al., 2022) in his research entitled "Evaluation of Test Results for the Clusterization Level of the Application of the K-Means Method in Determining the Spread of Covid-19 in Indonesia". The similarity between this research and previous research is that the object of research is the same level of spread of Covid-19 and uses the K-Means algorithm. While the difference between previous research and this research is that the subject in the previous study used the level of spread of covid-19 in Indonesia, while in this study the subject used the rate of spread of covid-19 in the Kenangan Community Health Center. In previous studies, the focus of the research was on evaluating test results, while in this study the focus was on predicting the spread of COVID-19. The results of the previous study resulted in 3 clusters with 12 provinces in the low category, 18 provinces in the middle category, and 4 provinces in the high category.

METHOD

In this study, the research procedure was as follows:

Data Collection

The data used in this study was obtained from patient data obtained from the health center. The data collected was 400 data.

Pre-Processing Data

Before testing the data, the data preprocessing stage is carried out. The stages of data preprocessing carried out is:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11740
e-ISSN : 2541-2019
p-ISSN : 2541-044X

a. Cleaning Data
   Data cleaning is used to remove duplicate data.
b. Data Transformation
   Data transformation, used to define the attributes to be used for testing.

Pengujian

The test was carried out using Rapidminer software. The test also implemented the k-means algorithm. In the k-means algorithm, several parameters are used, namely:

a. Clustering
   In the clustering operator technique used measure types, namely numerical measures with euclidean distance measurements.
b. Performance
   The performance operator uses a parameter, namely the Davies Bouldin Index. Davies Bouldin Index is used to determine the number of clusters, this is necessary because the smaller the value of the davies bouldin index, the better it will produce a good value compared to other clusters.

K-Means Algorithm

K-Means is a data grouping method that can group data into two or more groups. K-Means is the simplest algorithm and the most widely used clustering algorithm. In calculating the distance to I (xi) at the center of the group k (ck), which is called (dik), the Euclidean formula can be used with equation (1), (Adha et al., 2021):

$$d_{ik}\sqrt{\Sigma_{j=1}^{m}\left(c_{ij}-x_{ik}\right)^{2}} \quad (1)$$

RESULT

Pre-Processing Data

Prior to the data preprocessing stage, the data collected was obtained from patient data at the puskesmas, and 400 datasets were collected. The following table dataset is below.

Table 1
Dataset

Report Date	Name	Gender	Age	Hamlets	Village	Method
04/02/2021	Patient 1	L	39	Dusun-IX	Kenangan	PCR
07/02/2021	Patient 2	L	28	Dusun-VIII	Kenangan	PCR
12/02/2021	Patient 3	L	48	Dusun-VI	Kenangan	PCR
14/02/2021	Patient 4	L	24	Dusun-V	Kenangan	RDT
14/02/2021	Patient 5	L	63	Dusun-VI	Kenangan	PCR
14/02/2021	Patient 6	L	35	Lingkungan IX Kiwi	Kenangan	PCR
14/02/2021	Patient 7	L	53	Dusun-VI	Kenangan	PCR
15/02/2021	Patient 8	P	24	Dusun-VI	Kenangan	RDT
17/02/2021	Patient 9	L	38	Lingkungan IX Kiwi	Kenangan	PCR
17/02/2021	Patient 10	P	34	Lingkungan VI Enggang	Kenangan	PCR
…	…	…	…	…	…	…
26/03/2021	Patient 400	L	40	Lingkungan V	Kenangan	Swab

In the data preprocessing stage, data cleaning and data transformation are carried out. Cleaning data is used to remove duplicate data while data transformation is used to select data for the selection of attributes that will be used for testing in Rapidminer. The following table displays the results of the transformation below.

Table 2
Transformation Results

Hamlets	PCR	RDT	SWAB
Dusun IX	1	20	11
Dusun V	0	9	7
Dusun VI	2	25	12

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11740
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dusun VIII	2	7	3
Lingkungan I Seriti	13	6	0
Lingkungan II Jalak A	6	2	1
Lingkungan III Jalak B	13	0	5
Lingkungan IV Belibis	20	6	2
Lingkungan IX Kiwi	21	5	6
Lingkungan V	5	17	1
Lingkungan V Parkit	12	9	7
Lingkungan VI Enggang	19	20	7
Lingkungan VII Kenari A	22	20	9
Lingkungan VIII Kenari B	7	1	1
Lingkungan X	2	0	0
Lingkungan X Pelikan	20	10	4

Testing

After the transformation process is completed, the next stage is carried out testing. In this test, it used a rapidminer by implementing the K-Means algorithm. Testing is carried out to determine which areas have a high number of cases. The following tests using Rapidminer software can be seen in the picture below.

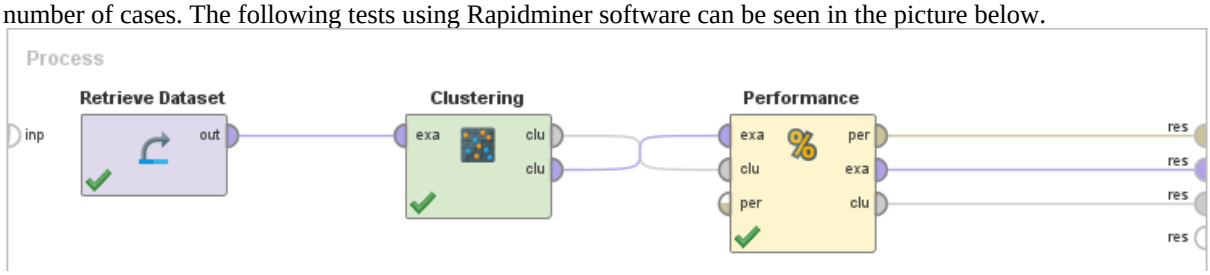

Fig. 1 Display of the application of the k-means algorithm on Rapidminer

In this test, it used a clustering technique with a total of 400 datasets to determine the number of clusterings. For testing in rapidminer using measure types, namely numerical measures with euclidean distance measurements while the performance parameters used in this study are the Davies Bouldin Index. After testing the dataset using the DBI parameter, the number of clusters was obtained as many as 3, because it became the best cluster with the smallest value of -0.784. The following is a table of test results, namely:

Table 3
Test Results

Atribut	Cluster 0	Cluster 1	Cluster 2
PCR	19	2,667	6,143
RDT	11,667	20,667	3,571
Swab	5,833	8	2,429

DISCUSSIONS

The results of the analysis on cluster 0 consist of 6 areas, namely Environment IV Belibis, Environment IX Kiwi, Environment V Parkit, Environment VI Enggang, Environment VII Kenari A, Environment X Pelikan. With the categories of screening and tracing results, namely PCR, RDT and Swab. The results of the analysis in cluster 1 consist of 3 regions, namely Hamlet IX, Hamlet VI, and Environment V. With the categories of screening and tracing results, namely PCR, RDT and Swab. While the results of the analysis in cluster 2 consist of 7 regions, namely Hamlet V, Hamlet VII, Environment I Seriti, Environment II Jalak A, Environment III Jalak B, Environment VIII Kenari B and Environment X. With the categories of screening and tracing results, namely PCR, RDT and Swab. Berikut tampilan hasil centroid plot dan performance vector

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11740
e-ISSN : 2541-2019
p-ISSN : 2541-044X

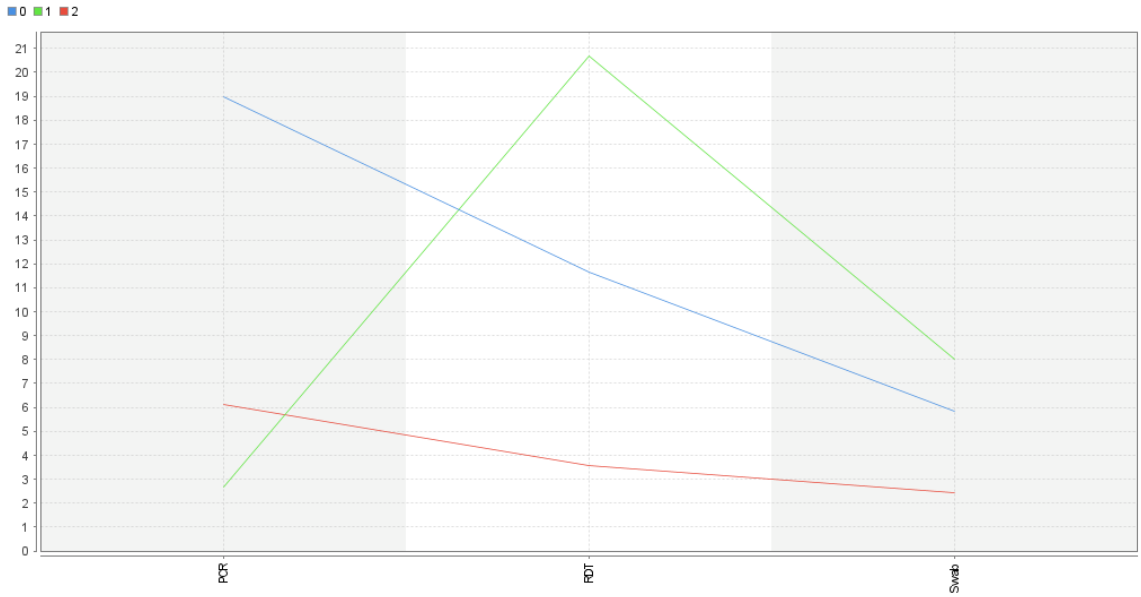

Fig. 2 Centroid Plot Results

From the centroid plot, it can be seen that the number of areas with positive cases of covid-19 using PCR testing is much higher than that of RDT and Swab.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -45.487
Avg. within centroid distance_cluster_0: -53.361
Avg. within centroid distance_cluster_1: -38.444
Avg. within centroid distance_cluster_2: -41.755
Davies Bouldin: -0.784
```

Fig. 3 Results of Performance Vector

CONCLUSION

Based on the results of studies that have been carried out on covid-19 patient data as many as 400 datasets tested with the K-Means algorithm. 2. The test produced 3 clusters consisting of low, medium and high categories. Cluster 0 belongs to the medium category, cluster 1 belongs to the low category and cluster 2 belongs to the high category. 3. The test also yielded a DBI value of -0.784. On cluster 0 has 6 regions, cluster 1 has and cluster 2 has 7 regions. With the testing that has been carried out by implementing the k-means algorithm, this algorithm can classify the spread of covid-19 based on the results of screening and tracing.

REFERENCES

Adha, R., Nurhaliza, N., Sholeha, U., & Mustakim, M. (2021). Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *18*(2), 206–211.

Akramunnisa, & Fajriani. (2020). K-Means Clustering Analysis pada PersebaranTingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. *Jurnal Varian*, *3*(2), 103–112. https://doi.org/10.30812/varian.v3i2.652

Asroni, A., Fitri, H., & Prasetyo, E. (2018). Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). *Semesta Teknika*, *21*(1), 60–64. https://doi.org/10.18196/st.211211

Bastian, A., Sujadi, H., & Febrianto, G. (2018). Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka). *Jurnal Sistem Informasi (Journal of Information System)*, *14*(1), 26–32.

Bu'ulolo, E., & Purba, B. (2021). Algoritma Clustering Untuk Membentuk Cluster Zona Penyebaran Covid-19.

*name of corresponding author

Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi, *12*(1), 59–67. https://doi.org/10.31849/digitalzone.v12i1.6572

Dwitri, N., Tampubolon, J. A., Prayoga, S., R.H Zer, F. I., & Hartama, D. (2020). Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Indonesia. *Jurnal Teknologi Informasi*, *4*(1), 128–132. https://doi.org/10.36294/jurti.v4i1.1266

Gayatri, L., & Hendry, H. (2021). Pemetaan Penyebaran Covid-19 Pada Tingkat Kabupaten/Kota Di Pulau Jawa Menggunakan Algoritma K-Means Clustering. *Sebatik*, *25*(2), 493–499. https://doi.org/10.46984/sebatik.v25i2.1307

Gustientiedina, Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *5*(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24

Mona, N. (2020). Konsep Isolasi Dalam Jaringan Sosial Untuk Meminimalisasi Efek Contagious (Kasus Penyebaran Virus Corona Di Indonesia). *Jurnal Sosial Humaniora Terapan*, *2*(2), 117–125. https://doi.org/10.7454/jsht.v2i2.86

Mursalim, Purwanto, & Soeleman, M. A. (2021). Penentuan Centroid Awal Pada Algoritma K-Means Dengan Dynamic Artificial Chromosomes Genetic Algorithm Untuk Tuberculosis Dataset. *Techno.Com*, *20*(1), 97–108. https://doi.org/10.33633/tc.v20i1.4230

Salsabila, F., & Intani, S. M. (2021). Implementasi Algoritma K-Means Dan C4.5 Dalam Menentukan Tingkat Penyebaran Covid-19 Di Indonesia. *Jurnal Siliwangi*, *7*(1), 25–30.

Sari, Y. P., Primajaya, A., & Irawan, A. S. Y. (2020). Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang. *INOVTEK Polbeng - Seri Informatika*, *5*(2), 229. https://doi.org/10.35314/isi.v5i2.1457

Virantika, E., Kusnawi, & Ipmawati, J. (2022). Evaluasi Hasil Pengujian Tingkat Clusterisasi Penerapan Metode K-Means Dalam Menentukan Tingkat Penyebaran Covid-19 di Indonesia. *Jurnal Media Informatika Budidarma*, *6*(3), 1657–1666. https://doi.org/10.30865/mib.v6i3.4325

Watratan, A. F., B, A. P., & Moeis, D. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology*, *1*(1), 7–14. https://doi.org/10.52158/jacost.v1i1.9

Yunus, N. R., & Rezki, A. (2020). Kebijakan Pemberlakuan Lockdown Sebagai Antisipasi Penyebaran Corona Virus Covid-19. *SALAM: Jurnal Sosial Dan Budaya Syar-I*, *7*(3), 227–238. https://doi.org/10.15408/sjsbs.v7i3.15083

*name of corresponding author