Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11603
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification of Covid-19 Patient Spread Rate By Age and Region With K-Means Algorithm

Adya Zizwan Putra[1)*], Ryan Wijaya Pinem[2)], Sehat Silalahi[3)], Fendianu Gulo[4)], Juan Antonio Adityo Liukhoto[5)]
[1,2,3,4,5)]Universitas Prima Indonesia, Indonesia
[1)] adyazizwanputra@unprimdn.ac.id, [2)] rw.pinem0608@gmail.com, [3)]Shtsilalahi4@gmail.com, [4)]fendianugulo123@gmail.com, [5)]raisersama@gmail.com

Submitted : July 30, 2022 | **Accepted** : Aug 4, 2022 | **Published** : Aug 5, 2022

Abstract: The Covid-19 virus is a new type of disease, the first case of covid-19 was found in Wuhan Province, China in 2019 with general symptoms such as pneumonia. This virus can grow rapidly and can cause serious infections and even death. Due to the very fast transmission of the virus, the WHO declared the Covid-19 virus a pandemic on March 11, 2020. Anyone can be infected with the covid-19 virus, from small children to the elderly. However, various ways have been done, but the cases of covid-19 continue to increase. Various ways have been done to reduce the spread of COVID-19 so that the Covid-19 virus does not spread quickly. Then data mining techniques are needed by implementing the K-Means algorithm because the K-Means algorithm can group data. In this study, 790 patient data were used for COVID-19 patients. The test resulted in 3 clusters grouped based on low, medium, and high categories with a DBI value of -0.332. In cluster 0 with a low category there are 3 districts, in cluster 1 with a medium category there is 1 sub-district, in cluster 2 with a high category, there are 6 districts. From the results of the test, it can be seen that the age susceptible to COVID-19 is 26 to 45 years.

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

INTRODUCTION

The Covid-19 virus is a new type of disease and has never been found to attack humans. The first case of covid-19 was found in Wuhan Province, China in 2019 (Sindi et al., 2020). Pneumonia with general flu-like symptoms is suspected to be the first outbreak of the COVID-19 virus. However, unlike the flu, the Covid-19 virus can develop quickly and cause serious infections and even death. Due to the high-speed transmission of the virus, WHO declared the Covid-19 virus a pandemic on March 11, 2020. The status of the pandemic indicates that the spread of Covid-19 is ongoing (Alvina Felicia Watratan et al., 2020). Anyone can get this virus, from small children to the elderly. The government has taken steps to prevent the Covid-19 virus from spreading quickly in Indonesia. Not only the government but all parties are doing various ways in order to stop the covid-19 virus because this virus many people are exposed because cases are getting higher every day (Darmansah, 2021). However, various ways have been done, but the cases of Covid-19 continue to increase. Therefore, to limit the spread of covid-19, it is necessary to classify areas with a high level of vulnerability to the spread of covid-19 (Gayatri & Hendry, 2021). Each region affected by the COVID-19 virus has a different spread, ranging from the number of positive cases, the number of recovered cases, and the number of cases treated to death. Different conditions require different handling from each region (Arifandi, 2021).

Therefore, classification is needed to classify the number of patients affected by Covid-19 based on age and region using data mining processing techniques. Data mining is a method for processing data on a large scale, data processed with this technique will produce new knowledge that is used for decision making (Sugianto et al., 2020). To determine the level of spread of Covid-19, data mining techniques are needed. Some of the methods used to determine the level of spread of covid-19 are Naïve Bayes (Alvina Felicia Watratan et al., 2020), K-Medoids (Sindi et al., 2020), C4.5 (Salsabila & Intani, 2021), KNN (Mustaghfiroh et al., 2022) and K-Means (Dwitri et al., 2020). One of the data mining algorithms used is the K-Means algorithm. K-Means is an unsupervised learning algorithm that groups data into several partitions. The data grouped using the k-means algorithm have the same properties, but the other groups have different properties (Muliono & Sembiring, 2019). K-means has advantages compared to other algorithms, namely simple, easy to implement, not slow in testing, and easy to adjust. K-means is also commonly used in data mining processing (Sari et al., 2020). The k-means

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v7i3.11603 p-ISSN : 2541-044X

algorithm was used in the research conducted by Suriani to classify criminal data at the Poldasu in knowing patterns of vulnerability in criminal activity. Because the K-Means algorithm is a simple and effective algorithm for finding groups in the data (Suriani, 2020). This study will implement the k-means algorithm to determine the level of spread of covid-19 based on age and region. It is hoped that this research can help the government in making decisions to tackle the spread of COVID-19. So that the decisions taken can break the chain of the spread of COVID-19.

LITERATURE REVIEW

Lestandy & Syafa'ah's research applies the KNN algorithm to predict covid-19 cases. This study uses 260 data with 13 parameters and 80% training data and 20% testing data. The results of this test resulted in an accuracy value of 72.3337% and an MSE of 0.007 (Lestandy & Syafa'ah, 2020). Research conducted by Dwitri et al applies the K-Means algorithm to determine the level of spread of Covid-19 in Indonesia. This study uses data on the number of Covid-19 viruses spreads on May 9, 2020, obtained from the Ministry of Health of the Republic of Indonesia. Tests were carried out using rapidminer to produce 3 clusters. Cluster 1 has 5056 cases of positive patients and 427 cases of death, for cluster 2 has 4525 positive cases and 348 cases died, while cluster 3 has 4043 positive cases and 184 cases died (Dwitri et al., 2020). A study conducted by Gunawan et al in his research used the K-Medoids algorithm to cluster provinces in Indonesia that were affected by the Covid-19 virus. In his research, he used patient data from March 2, 2020, to June 30, 2020, using 3 variables: confirmed, died, and recovered. The test results showed 3 clusters, cluster 1 had 12259 confirmed cases, 793 cases died and 5631 recovered cases. Cluster 2 had 2632 confirmed cases, 108 cases died and 1077 recovered cases. While cluster 3 had 388 confirmed cases, dead cases 1 and 210 cases were recovered (Gunawan et al., 2020).

n a study conducted by R Sianipar et al regarding the application of the k-means algorithm to determine the level of satisfaction of online learning during Covid-19. In his research using data filled in through a questionnaire link, the results of the tests carried out resulted in 3 clusters, namely the first cluster which stated that they agreed to online learning was low and those who disagreed was high, the second cluster that agreed to online learning was moderate and those who stated no agree is moderate, while cluster 3 which agrees on online learning is high and those who disagree are low (R Sianipar et al., 2020). Research conducted by Sunia et al uses the k-means algorithm in their research to find data about the poor. The study used data obtained from the Jambi City BPS in March 2017 with a total of 286.55 thousand poor people. The calculation results produce 512 samples of data with 5 clusters. The first cluster has 13 residents, the second cluster has 153 residents, the third cluster has 129 residents, the fourth cluster has 138 residents, and the fifth cluster has 79 residents (Sunia et al., 2019). Hasyrif SY conducted a study using the k-means algorithm to classify the spread of diarrhea in Makassar City. This study uses health data from 2016. The test results get 2 clusters. The second cluster is the cluster with the highest area (Sy et al., 2019). esearch conducted by Abdullah et al in his research used the k-means algorithm to classify provinces affected by the COVID-19 pandemic. The data used in this study used data obtained on April 19, 2020, from the COVID-19 task force. The testing in this study resulted in 3 groups, namely group 1 there were 4 provinces, group 2 had 28 provinces while in group 3 there were 1 province (Abdullah et al., 2022). Research conducted by Salsabila & Intani in his research carried out a combination of algorithms to determine the level of spread of Covid-19, the algorithms used were K-Means and C4.5. The data used is data obtained on January 10, 2021, from the website of the Ministry of Health of the Republic of Indonesia with a total of 34 records. The study resulted in 4 clusters in the black category, red zone, yellow zone, and green zone with a DBI value of 0.110. The result of the combination of algorithms to be used in providing new knowledge is information on the number of the spread of Covid-19 cases (Salsabila & Intani, 2021).

METHOD

In this study, the research procedure was as follows:

Data Collection

Used in this study were obtained from patient data obtained from the puskesmas. The data collected is 790 patient data.

Pre-Processing Data

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

a. Cleaning Data
   In this process, duplicate data is removed.
b. Data Transformation
   At this stage, data selection is carried out for attribute selection.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11603
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Testing

The test was carried out using Rapidminer software. The test also implemented the k-means algorithm. The k-means algorithm uses the perfomance vector parameter, namely the Davies Bouldin Index. DBI is the ratio of the number of distances between clusters and the distance between clusters. To obtain good clustering results, the distance between clusters must be large and the inter-cluster distances small, therefore a lower DBI value is required to indicate good clustering results.

RESULT

Pre-Processing Data

Before the data preprocessing stage is carried out, the initial data collected is shown in the table below.

Table 1
Dataset

Nama	Usia	Kelurahan Domisili	Hasil Pemeriksaan Awal	Hasil Pemeriksaan Akhir
Pasien 1	52 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 2	75 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 3	44 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 4	47 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 5	52 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 6	23 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 7	25 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 8	79 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 9	79 Tahun	Helvetia Tengah	Positif	Sembuh
Pasien 10	22 Tahun	Helvetia Tengah	Positif	Sembuh
…	…	…	…	…
Pasien 790	1 Tahun	Helvetia Tengah	Positif	Sembuh

In the early stages of data preprocessing, data cleaning is carried out, this stage is carried out to remove duplicate data. After the data cleaning stage is complete, the data transformation stage is carried out, this stage is carried out to select the data and the selection of attributes that will be used in the test. The following is a display of the transformation table below.

Table 2
Hasil Transformasi

Kelurahan	Balita	Anak	Remaja	Dewasa	Lansia
Cinta Damai	2	8	19	21	16
Dwikora	3	5	8	27	13
Hamparan Perak	0	0	0	2	0
Helvetia	0	17	28	28	27
Helvetia Tengah	8	28	87	63	93
Helvetia Timur	8	16	16	57	30
Sei Sikambing C II	4	2	17	19	12
Sei Sikambing D	0	0	0	2	0
Simpang Selayang	0	0	0	0	2
Tanjung Gusta	12	7	23	39	21

Testing

In the table that has been preprocessing the data, the data will be processed using Rapidminer software by implementing the k-means algorithm. The test is carried out to find out areas that have a high number of cases and ages that are susceptible to COVID-19. Testing using Rapidminer software can be seen in the image below.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11603
e-ISSN : 2541-2019
p-ISSN : 2541-044X

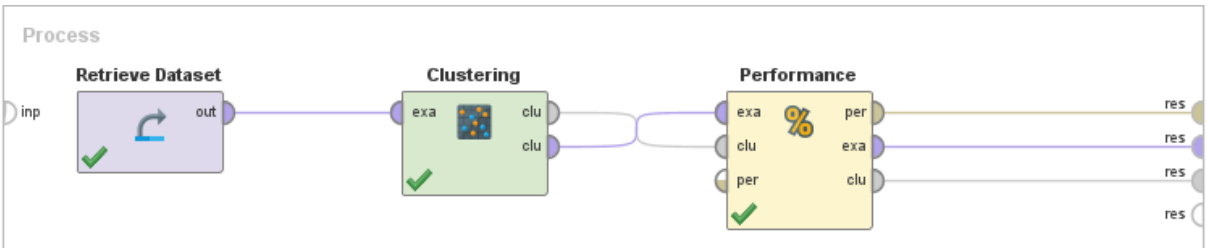

Fig. 1 Display of the application of the k-means algorithm on Rapidminer

In this test, the clustering technique is used with a total of 790 datasets to determine the number of clustering. For testing in rapidminer, measure types are used, namely numerical measure with Euclidean measure measurement, while the performance parameter used in this study is Davies Bouldin. After testing the dataset, the number of clusters obtained is 3, namely:

Table 3
Test Results

Atribut	Cluster 0	Cluster 1	Cluster 2	DBI
Balita	0	8	4,833	-0.332
Anak	0	28	9,167	
Remaja	0	87	18,500	
Dewasa	1,333	63	31,833	
Lansia	0,667	93	19,833	

DISCUSSIONS

Based on the test results on patient data, it produced 3 clusters, namely low clusters, medium clusters, high clusters. Table 4 shows that cluster 0 is a low cluster consisting of Hamparan Perak, Sei Sikambing D, and Simpang Selayang Districts, cluster 1 is a medium cluster consisting of Helvetia Tengah District, and cluster 2 is a high cluster consisting of Cinta Damai District, Dwikora District, Helvetia District, East Helvetia District, Sei Sikambing C II District, and Tanjung Gusta District. And in this test, it can be seen that adults and the elderly are the age most exposed to Covid-19 which can be caused by congenital diseases, or often move outside the home so they are easily exposed to Covid-19. The following shows the results of the centroid plot.

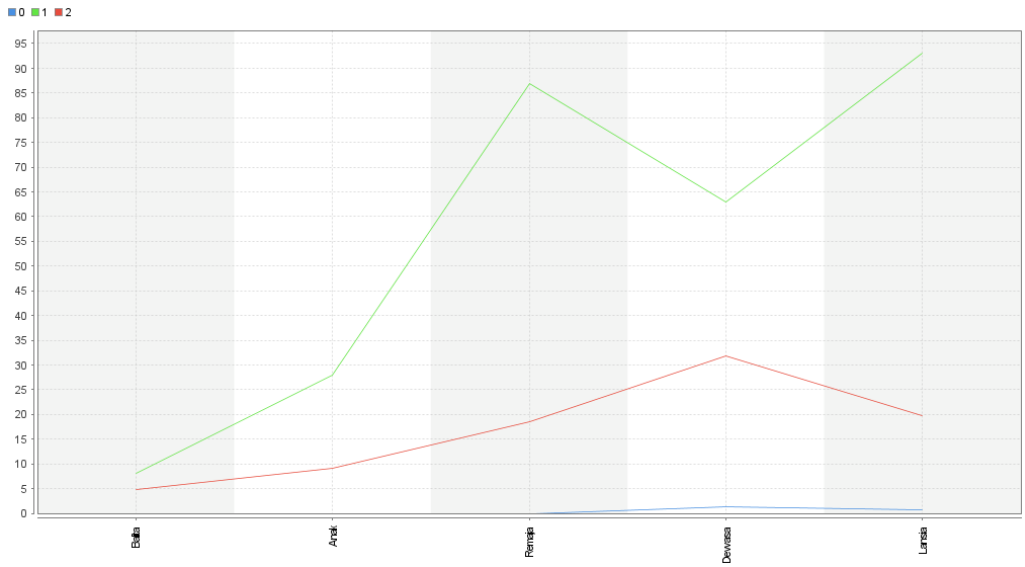

Fig. 2 Centroid Plot Results

CONCLUSION

Based on the results of studies that have been carried out on covid-19 patient data, as many as 790 datasets were tested with the K-Means algorithm. The test produced 3 clusters with a DBI value of -0.332, of which cluster 0 is a low category consisting of 3 sub-districts. Cluster 1 is a medium category consisting of 1 sub-

*name of corresponding author

district, while cluster 2 is a high category consisting of 6 sub-districts. From the results of testing in several sub-districts, almost all ages are exposed to covid-19 but the age that is vulnerable to covid-19 is adulthood and the elderly. With the testing that has been carried out by implementing the k-means algorithm, this algorithm can classify the level of spread of covid-19.

REFERENCES

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality and Quantity*, *56*(3), 1283–1291. https://doi.org/10.1007/s11135-021-01176-w

Alvina Felicia Watratan, Arwini Puspita. B, & Dikwan Moeis. (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

Arifandi, M. H. A. H. A. A. D. A. D. (2021). Implementasi algoritma K-Medoids untuk clustering wilayah terinfeksi kasus COVID-19 di DKI Jakarta. *JTT (Jurnal Teknologi Terapan)*, *7*(2), 120–128. https://jurnal.polindra.ac.id/index.php/jtt/article/view/353

Darmansah, D. D. (2021). Analisis Penyebaran Penularan Virus Covid-19 di Provinsi Jawa Barat Menggunakan Algoritma K-Means Clustering. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(3), 1188–1199. https://doi.org/10.35957/jatisi.v8i3.1034

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

Gunawan, I., Anggraeni, G., Rini, E. S., & Mustofa, Y. (2020). Klasterisasi provinsi di Indonesia berbasis perkembangan kasus Covid-19 menggunakan metode K-Medoids. *Seminar Nasional Matematika Dan Pendidikan Matematika (5thSENATIK)*, 301–306.

Lestandy, M., & Syafa'ah, L. (2020). Prediksi Kasus Aktif Covid-19 Menggunakan Metode K-Nearest Neighbors. *Seminar Nasional Teknologi Dan Rekayasa (SENTRA) 2020*, 45–48.

Muliono, R., & Sembiring, Z. (2019). Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. *CESS (Journal of Computer Engineering, System and Science)*, *4*(2), 2502–2714.

Mustaghfiroh, L., Ariani, M. H., Info, A., Neighbor, K., Mustaghfiroh, L., Informatika, P. S., Informatika, F. T., Tinggi, S., & Pati, T. (2022). KLASIFIKASI PASIEN COVID-19 DI INDONESIA MENGGUNAKAN METODE K-NEAREST NEIGHBOR. *Jurnal Nasional AMRI (Analisa, Metode, Rekayasa, Informatika)*, *1*(1), 16–21. https://doi.org/10.12487/AMRI.v1i1.xxxxx

R Sianipar, K. D., Wanti Siahaan, S., Siregar, M., & Fikrul Ilmi Zer, P. R. (2020). Penerapan Algoritma K-Means Dalam Menentukan Tingkat Kepuasan Pembelajaran Online Pada Masa Pandemi Covid-19. *Jurnal Teknologi Informasi*, *4*(1), 101–105.

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

Sindi, S., Ningse, W. R. O., Sihombing, I. A., R.H.Zer, F. I., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia. *Jurnal Teknologi Informasi*, *4*(1), 166–173. https://doi.org/10.36294/jurti.v4i1.1296

Sugianto, C. A., Rahayu, A. H., & Gusman, A. (2020). Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Cigugur Tengah. *Journal of Information Technology*, *2*(2), 39–44. https://doi.org/10.47292/joint.v2i2.30

Sunia, D., Kurniabudi, & Alam Jusia, P. (2019). Penerapan Data Mining Untuk Clustering Data Penduduk Miskin Menggunakan Algoritma K-Means. *Jurnal Ilmiah Mahasiswa Teknik Informatika*, *1*(2), 121–134.

Suriani, L. (2020). Pengelompokan Data Kriminal Pada Poldasu Menentukan Pola Daerah Rawan Tindak Kriminal Menggunakan Data Mining Algoritma K-Means Clustering. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *1*(2), 151. https://doi.org/10.30865/json.v1i2.1955

Sy, H., Rismayani, & Syam, A. (2019). Data Mining Menggunakan Algoritma K-Means Pengelompokan Penyebaran Diare di Kota Makassar. *SISITI : Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *8*(1), 73–82.

*name of corresponding author