Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13551
e-ISSN : 2541-2019
p-ISSN : 2541-044X

K-Medoids Algorithm to Clustering COVID-19 patients with Various Age Levels at Hospitals in Yogyakarta Province

Pratamasari Noor Insani[1)], Endang Darmawan[2)], Sugiyarto[3)]
[1,2,3)] Ahmad Dahlan University of Yogyakarta, Indonesia
[1)]pratamasari.insani8@gmail.com, [2)] endang.darmawan@pharm.uad.ac.id, [3)]sugiyarto@math.uad.ac.id

Submitted : Feb 29, 2024 | **Accepted** : Apr 4, 2024 | **Published** : Apr 5, 2024

Abstract: COVID-19 can cause a wide spectrum of symptoms, such as mild upper respiratory infection or life-threatening sepsis. From 20.2% of cases of COVID-19 progressed to severe disease with a mortality rate of 3.1% where 60%-90% of patients with comorbidities were hospitalized. The purpose of this study was to find out that cluster analysis using K-Medoids can distinguish COVID-19 patients at various age levels which analytical method has sensitivity and specificity values in analyzing clustering in COVID-19 patients. This study uses a cohort retrospective design conducted at five hospitals in Yogyakarta Province. The study used patient medical record data from March 2020 – September 2021 with a total of 916 patient data that met the inclusion criteria. Cluster analysis will be carried out using Google Colaboratory with the Python programming language. The clustering results are divided into 2 cluster groups where cluster 1 consists of 558 patients and cluster 2 consists of 358 patients with various age levels. The test resulted in 2 clusters with a DBI value of 5,191631. The results of statistical tests showed that there was a significant relationship (*p*-value = 0,023) between age, recovery rate, and patient mortality. From the test results, it can be seen that ages 50 to 59 years are suspected of COVID-19

Keywords: Age Levels, COVID-19, Cohort Retrospective, Clustering, K-Medoids

INTRODUCTION

Data from Johns Hopkins University on May 31, 2021, recorded 170,074,768 positive confirmed cases of COVID-19 with 3,535,842 deaths worldwide, while China recorded 102,960 positive confirmed cases with 4,846 deaths. Meanwhile, the number of COVID-19 cases in Indonesia showed 1,809,926 with a death toll of 50,262 (John Hopkins University, 2021).

Critically ill patients progress rapidly to acute respiratory distress syndrome, metabolic acidosis, coagulopathy, septic shock, and multiorgan failure. Another research said from 53,000 hospitalized patients showed that 20.2% of COVID-19 cases developed severe disease with a mortality rate of 3.1% (Harapan et al., 2020). There is a significant increase in mortality in the elderly and those suffering from comorbidities, such as cardiovascular disease, chronic kidney disease, and chronic obstructive pulmonary disease. COVID-19 affects mostly people aged 30-79 years in China, while cases over 80 years and under 19 years are relatively rare with an average age range of 47-56 years, 15% of cases occur in smokers, 25-30 % of patients showed concomitant disease with 40% of them having cardiovascular disease (Pericàs et al., 2020). Numerous studies have identified comorbidities associated with the adverse prognosis of COVID-19, age, gender, and at least a few comorbidities are the strongest predictors of the prognosis of COVID-19 patients (Thakur et al., 2021).

Clustering refers to the assignment of patterns into groups (clusters) so that objects belonging to the same group are more similar to each other than those in different groups (Azar et al., 2013). The k-Medoids algorithm is one of the unsupervised learning methods. The K-Medoids algorithm uses objects as representatives (medoids) as cluster centers for each cluster (Kaur et al., 2014). The K-Medoids algorithm has advantages in overcoming weaknesses in the K-Means algorithm sensitive to noise and outliers, where an object with a large value that is lets deviates from the data distribution. Another advantage is the results clustering process does not depend on the sequence enter datasets (Pramesti et al., 2017). A study showed the performance of the K-Medoids algorithm is superior to the K-Means algorithm in terms of accuracy with an accuracy of 63,24% (Nurhayati et al., 2019). Study on diabetes data to know the characteristics and track the maximum number of patients suffering from diabetes based on the area calculation algorithm, it was found that K-Medoids is the best algorithm for generating clusters

*name of corresponding author

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

(Anuradha et al., 2014). The purpose of this study was to find out whether K-Medoids can distinguish COVID-19 patients at various ages level and can help the government to break the chain of transmission of COVID-19.

LITERATURE REVIEW

K-medoids is a grouping method in data mining that is part of partitional clustering. This method uses objects in the object pool to represent a cluster. The advantages of this method are being able to overcome the weaknesses of the k-means method which are sensitive to outliers and the results of the clustering process do not depend on the order in which the dataset is entered (Pramesti et al., 2017). K-medoids have a good performance more optimal if the amount of data used is small (Rofiqi, 2017). The K-Medoids algorithm is better than the K-Means algorithm in terms of accuracy, execution time, and time complexity (Nurhayati et al., 2019).

Another study produced a Silhouette Coefficient K-Medoids validity value of 0,5009 and a K-Means validity value of 0,1443. This shows that the K-Medoids algorithm is better at grouping data on the distribution of disabled children (Marlina et al., 2018). Sindi et al research, based on implementation and testing, the K-Medoids algorithm can group Covid-19 data on which areas are infected in each region with the best clustering done with 3 clusters. Of the 34 records obtained 1 record was in the first cluster, 2 records in the second cluster, and 31 records in the third cluster (Sindi et al., 2020). Another researcher for the K-Medoids algorithm for disease grouping in Pekanbaru Riau produces 4 clusters as the best grouping (Juninda et al., 2019). By clustering online applications such as WhatsApp, zoom, and moodle which are often used in the learning process for 100 students, can produce a grouping of applications that students like and dislike in the learning process. there are 2 clusterings in red and blue (Samudi et al., 2020).

Irwansyah et al research raised the topic of grouping cardiovascular disease patients. This research was conducted using the K-Medoids method to produce two clusters with a silhouette coefficient of 0,35 (Irwansyah et al., 2020). In the research conducted by Bu'ulolo and Purba, the K-Medoids clustering algorithm can be applied in the formation of clusters of Covid-19 distribution zones, especially in North Sumatra. Based on the data used, the spread of Covid-19 can be clustered into 3 (three) groups. The result of the formation of clusters of Covid-19 distribution zones is a cluster 1 (one) is a zone with high cases (red zone), cluster 2 (two) is a zone with moderate cases (yellow zone) and cluster 3 is a zone with low cases (green zone) (Bu'ulolo & Purba, 2021).

In previous study conducted by Ningrum et al, K-Medoids algorithm can be applied to data on the percentage of children affected by allergic diseases by province, so 34 provinces are obtained resulting in 21 provinces, namely low clusters, 12 provinces in medium clusters, and 1 province in high clusters allergy immunization percentage in each province (Ningrum et al., 2021).

METHOD

Data Collection

This type of research is an observational retrospective cohort. This study consisted of large-scale data sources from hospital admissions and patient clinical data from March 2020 to September 2021. The study criteria included patients with laboratory-confirmed SARS-Cov-2 infection treated at five hospitals in Yogyakarta Province.

Pre-Processing Data

a. Data Cleaning
Data cleaning is used to remove duplicate data, remove unnecessary columns, and remove irrelevant observations and errors.

b. Data Transformation
The process of changing and transforming the data format into a numeric data format / initialization is carried out.

c. Normalization
Normalize data by reducing the weight value per cluster with the Min value in the cluster then dividing it by the result of the reduction between the Max and Min cluster values

Statistical Analysis

The data that has been obtained will then be clustered using the K-Medoids algorithm which is then processed using Google Colaboratory with the Python programming language. The K-Medoids algorithm uses a performance vector parameter, namely the Davies Bouldin Index (DBI). To get good clustering results, the distance between clusters must be large and the distance between clusters must be small, therefore a lower DBI value is needed to show good clustering results.

Flowchart Diagram

The research flowchart is structured as follows

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13551
e-ISSN : 2541-2019
p-ISSN : 2541-044X

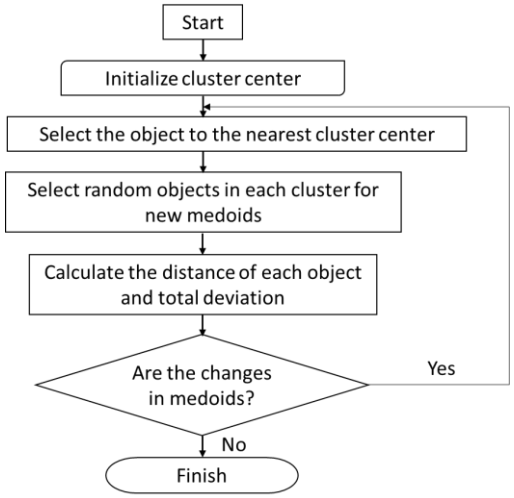

Fig. 1 K-medoids algorithm flowchart

RESULT

Descriptive of Indonesian Covid-19 Patients

The frequency distribution diagram for the characteristics of Covid-19 patients in the hospitals in the period March 2020 - September 2021, the largest age group is the 50-59 years age group, namely 254 patients (27,73%) and the 60-69 years (20,96%) age group as the second largest age group. This is in line with previous research on the characteristics of Covid-19 patients which stated that the number of Covid-19 cases based on age was dominated by the age group >30 years (Karyono & Wicaksana, 2020) (Li et al., 2020). See Fig. 2.

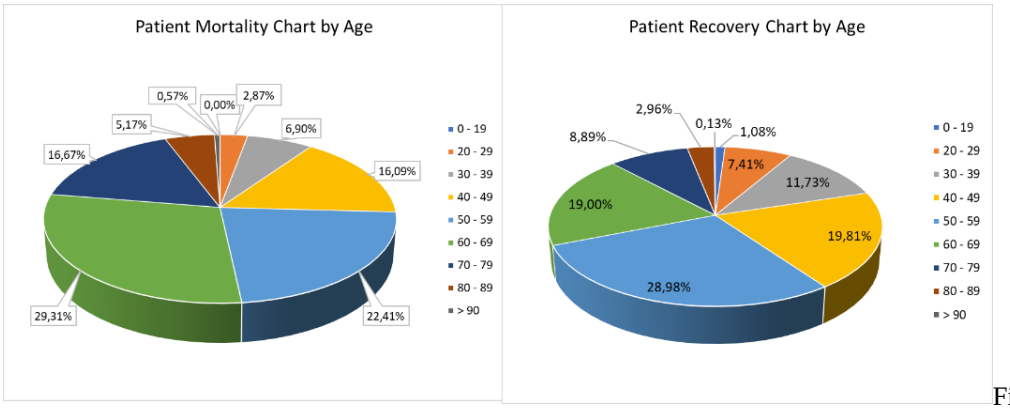

Fig. 2 Distribution diagram of covid patients at various ages level

In this test, a clustering technique was used with a total of 916 patients' medical record data to determine the number of clusters. For testing on Google Colaboratory, the type of measurement used is a numerical measurement with Euclidean measurements while the performance parameters used in this study are Davies Bouldin Index. After testing the dataset, the number of clusters obtained is 2, namely:

Table 1. Clustering result

Variable	Cluster 0	Cluster 1	Sig. (*p*-value)	Silhoutte Coefficient	DBI
0-19 years	5	3	0,023	0,025644	5,191631
20-29 years	42	18			
30-39 years	56	43			
40-49 years	97	78			
50-59 years	147	107			
60-69 years	127	65			
70-79 years	62	33			
80-89 years	21	10			
>90 years	1	1			

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13551
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

Based on the results of data processing using Google Colaboratory, it produced 2 clusters in Figure 3 which is cluster 0 consists of 558 patients, and cluster 1 consists of 358 patients, from each cluster, the age range of 50-59 years has a greater number of patients than other ages. The results of the age distribution in each cluster can be seen in Figure 4

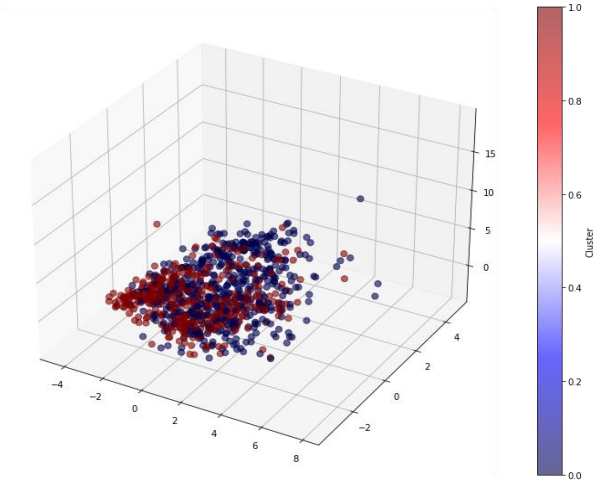

Fig. 3 Plot of Distribution of COVID-19 Patient Clusters using the K-Medoids Algorithm
(Blue: Cluster 0; Red: Cluster 1)

The value of p = 0,023 is lower than α = 0,05 (Table 1). So it can be said that age has a significant relationship to recovery and death in COVID-19 patients. The Silhouette Coefficient value is in the range (-1) to 1. The higher the silhouette value, the better the results will be. In this study, the K-Medoids silhouette coefficient was 0,025644. K-Medoids algorithm gets a DBI value of 5,191631. The DBI value range is 0 to 1, the smaller the value obtained, the higher the similarity of the data in one group.

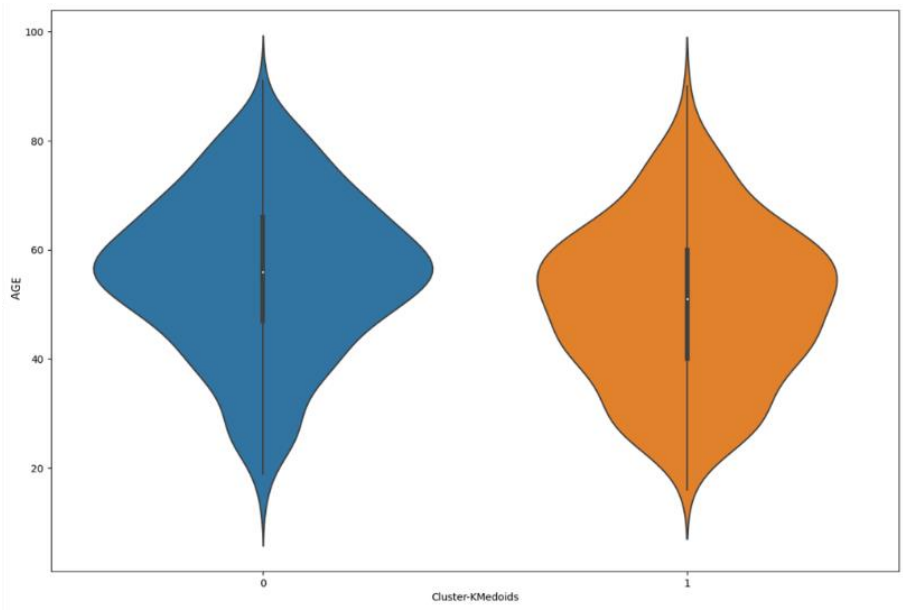

Fig. 4 The age distribution in each cluster

CONCLUSION

Based on the results of data processing that has been carried out on COVID-19 patient data, as many as 916 COVID patient medical record data were tested with the K-Medoids algorithm. The test resulted in 2 clusters with a DBI value of 5,191631 where cluster 0 consisted of 558 patients and cluster 1 consists of 358 patients. From the results of these tests, almost all ages are exposed to COVID-19, but the age that is vulnerable to COVID-19 is the age range of 50-59 years. Tests that have been carried out by implementing the K-Medoids algorithm, this his algorithm can group patients who are at risk of contracting COVID-19 based on various age levels.

*name of corresponding author

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

REFERENCES

Anuradha, S., Jyothirmai, P., Tirumala, Y., Goutham, S., & Hariprasad, V. (2014). Comparative Study of Clustering Algorithms on Diabetes Data. *International Journal of Engineering Research & Technology (IJERT)*, *3*(6), 922–926.

Azar, A. T., El-Said, S. A., & Hassanien, A. E. (2013). Fuzzy and hard clustering analysis for thyroid disease. *Computer Methods and Programs in Biomedicine*, *111*(1), 1–16. https://doi.org/10.1016/j.cmpb.2013.01.002

Bu'ulolo, E., & Purba, B. (2021). Algoritma Clustering Untuk Membentuk Cluster Zona Penyebaran Covid-19. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *12*(1), 59–67. https://doi.org/10.31849/digitalzone.v12i1.6572

Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., Megawati, D., Hayati, Z., Wagner, A. L., & Mudatsir, M. (2020). Coronavirus disease 2019 (COVID-19): A literature review. *Journal of Infection and Public Health*, *13*(5), 667–673. https://doi.org/10.1016/j.jiph.2020.03.019

Irwansyah, E., Pratama, E. S., & Ohyver, M. (2020). *Clustering of Cardiovascular Disease Patients Using Data Mining Techniques with Principal Component Analysis and K-Medoids Clustering of Cardiovascular Disease Patients Using Data Mining Techniques with Principal Component Analysis and K-Medoids*. *August*. https://doi.org/10.20944/preprints202008.0074.v1

John Hopkins University. (2021). *Coronavirus Resource Center*. https://coronavirus.jhu.edu/region/indonesia

Juninda, T., Mustasim, & Andri, E. (2019). Penerapan Algoritma K-Medoids untuk Pengelompokan Penyakit di Pekanbaru Riau. *Seminar Nasional Teknologi Informasi, Komunikasi Dan Industri*, *11*(1), 42–49.

Karyono, D. R., & Wicaksana, A. L. (2020). Current prevalence, characteristics, and comorbidities of patients with COVID-19 in Indonesia. *Journal of Community Empowerment for Health*, *3*(2), 77. https://doi.org/10.22146/jcoemph.57325

Kaur, N. K., Kaur, U., & Singh, D. (2014). K-Medoid Clustering Algorithm- A Review. *International Journal of Computer Application and Technology (IJCAT) Volume 1 Issue 1 (April 2014) ISSN: 2349-1841*, *1*(1), 42–45.

Li, H., Wang, S., Zhong, F., Bao, W., Li, Y., Liu, L., Wang, H., & He, Y. (2020). Age-Dependent Risks of Incidence and Mortality of COVID-19 in Hubei Province and Other Parts of China. *Frontiers in Medicine*, *7*(April), 1–6. https://doi.org/10.3389/fmed.2020.00190

Marlina, D., Lina, N., Fernando, A., & Ramadhan, A. (2018). Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *4*(2), 64. https://doi.org/10.24014/coreit.v4i2.4498

Ningrum, H., Irawan, E., & Lubis, M. R. (2021). Implementasi Metode K-Medoids Clustering Dalam Pengelompokan Data Penyakit Alergi Pada Anak. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *6*(1), 130. https://doi.org/10.30645/jurasik.v6i1.277

Nurhayati, Sinatrya, N. S., Wardhani, L. K., & Busman. (2019). Analysis of K-Means and K-Medoids's Performance Using Big Data Technology. *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, *Citsm*, 1–5. https://doi.org/10.1109/CITSM.2018.8674251

Pericàs, J. M., Hernandez-Meneses, M., Sheahan, T. P., Quintana, E., Ambrosioni, J., Sandoval, E., Falces, C., Marcos, M. A., Tuset, M., Vilella, A., Moreno, A., & Miro, J. M. (2020). COVID-19: From epidemiology to treatment. *European Heart Journal*, *41*(22), 2092–2108. https://doi.org/10.1093/eurheartj/ehaa462

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 Vol. 1, No. 9, Juni 2017, Hlm. 723-732*, *1*(9), 723–732. https://doi.org/10.1109/EUMC.2008.4751704

Rofiqi, A. Y. (2017). Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat. *Jurnal Simantec*, *6*(1), 25–32.

Samudi, S., Widodo, S., & Brawijaya, H. (2020). The K-Medoids Clustering Method for Learning Applications during the COVID-19 Pandemic. *SinkrOn*, *5*(1), 116. https://doi.org/10.33395/sinkron.v5i1.10649

Sindi, S., Ningse, W. R. O., Sihombing, I. A., Ilmi R.H.Zer, F., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia. *Jti (Jurnal Teknologi Informasi)*, *4*(1), 166–173.

Thakur, B., Dubey, P., Benitez, J., Torres, J. P., Reddy, S., Shokar, N., Aung, K., Mukherjee, D., & Dwivedi, A. K. (2021). A systematic review and meta-analysis of geographic differences in comorbidities and associated severity and mortality among individuals with COVID-19. *Scientific Reports*, *11*(1), 1–13. https://doi.org/10.1038/s41598-021-88130-w

*name of corresponding author