EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol3.iss2.art5

Research Article

Application of K-Means for Clustering COVID-19 in Indonesian Private Hospitals

Kathina Deswiaqsa[1*], Endang Darmawan[1], Sugiyarto [2]

[1] Departmen of Pharmacy Clinic, Faculty of Pharmacy, Ahmad Dahlan University, Yogyakarta
[2] Mathematics Faculty of Applied Science and Technology, Ahmad Dahlan University, Yogyakarta

[*] *Corresponding author: kety.deswiaqsa16@gmail.com*

Received: 18 April 2022; Accepted: 22 August 2022; Published: 6 September 2022

Abstract: In December 2019, coronavirus (COVID-19) caused by SARS-CoV-2 was first discovered in Wuhan, China. This virus has a high transmission rate and can be transmitted through droplets, airborne, and aerosols. The clinical manifestations are very diverse ranging from mild, moderate, and severe. Therefore, this study aims to conduct a clustering of the spread of the Covid-19 pandemic to facilitate the identification and handling. The method of the K-Means algorithm can be used as a method to obtain the desired clustering. The implementation and evaluation were conducted using RapidMiner tools and Davies Bouldin Index (DBI) respectively. Furthermore, the data sources by Kangdra (2020) were used with a total sample of 110 for the period March-June 2020. The results showed that the optimal cluster is located at k: 2 with a DBI value: 0,094 as the lowest value. Therefore, the cluster is strong since a smaller DBI value gives a better cluster. The clustering obtained is Cluster 1 and 2 with mild and moderate severity. The results are expected to facilitate a better zone identification of the COVID-19 severity level and rising people awareness.

Keywords: COVID-19, Clustering, K-means, Severity Level, Medan

Introduction

The WHO has reported 57,882,183 and 1,377,395 confirmed and death cases as of November 2020. Currently, the pandemic has spread almost all over the world, and in the province of North Sumatra, as of June 3, 2020, there were 444 positive confirmed cases, 43 deaths, and 159 recoveries.

The high number of confirmed positive cases as many as 444 cases as of June 3, 2020 based on the background of cases in North Sumatra Province, has a different severity of COVID-19 and can be seen during the initial examination at the hospital. Thus, patients who are hospitalized with high severity should be given more attention and special care. This is one of the efforts to reduce the spread of Covid-19

Common signs and symptoms of the infection are acute respiratory disorders such as fever (83% -98%), cough (76% - 82%), and shortness of breath (31% - 55%) [1]. The incubation period is 5-6 days, and according to WHO (2020), COVID-19 sufferers are divided based on mild, moderate, and severe symptoms The data with similar characteristics were obtained and grouped through the clustering method, and it was used to easily identify the severity of COVID-19. One of the algorithms in clustering often encountered and most often used is K-Means. It is simple, easy to use, efficient, and widely used [2].

Comorbidities and complications are used to determine the severity of covid-19[3] One of the methods used is clustering or grouping. Clustering is a technique for identifying and grouping data that share similar characteristics. One of the algorithms in clustering that is often encountered and most

often used is K-Means because it has the advantages of being simple, easy to use, and efficient. Due to the simplicity of the K-means algorithm, it has been used by a large number of previous researchers [4]

In the study of Azarafza Mehdi et al. entitled "*Clustering method for spread pattern analysis of coronavirus (COVID-19) infection in Iran*," grouping data based on the K-Means algorithm on the prevalence of the coronavirus in Iran produces Clustering 0 to 5. The lower the value, the lower the infection risk; for example, a value of 0 indicates a low risk of infection, while cluster 5 indicates a high risk of infection [5]

In another study, namely the grouping of covid-19 with the K-means algorithm by entering data from age, gender, and comorbidities which were grouped into mild, moderate, and weight (Armstrong, Zhu, Hirdes, & Stolee, 2012). In Indraputra's research (2020) with the title K-Means Clustering Covid-19 using the Data Mining method, namely K-Means Clustering Microsoft Excel software and Weka and KNIME Data Mining software, two data clusters were obtained, including those with higher values. Number of infections and deaths. compared to cluster 1, the handling of this cluster area needs to be prioritized [6].

In the clustering study of COVID-19 disease in each province in Indonesia, due to the diverse topography and population of Indonesia, there are variations in the number of cases between provinces. Therefore, it is necessary to do clustering to compile a map of COVID-19 cases using the K-Means method according to each province [7].

In cluster research, many researchers used the K-Means algorithm because it has the advantage that it is faster and easier to implement [8]. There has been no study related to the clustering of the pandemic based on symptoms in North Sumatra, Medan. Therefore, this study aims to cluster the level of symptoms of COVID-19 patients at Mitra Medika Amplas Hospital in Medan, North Sumatra Province, Indonesia, and the results are expected to be a reference for the world of health.

Methods

A retrospective cohort design from Windy Yoanna Kangdra's study entitled Clinical Characteristics and Comorbid Factors in Patients Under Surveillance (PDP) for Coronavirus Disease 2019 (COVID-19) was used [7]. The attributes or variables used are genders, age, profession, shortness of breath, cough, painful swallowing, fever, nausea and vomiting, diarrhea, abdominal pain, comorbidities, Rapid Test and PCR. Sampling was carried out using the total sampling method, with a population of 122 people, and after being selected, it became 110 people with the inclusion criteria provisions, namely all medical records of PDP COVID-19 at Mitra Medika Amplas Hospital with complete data. Repeated or incomplete medical records at Mitra Medika Amplas Hospital were excluded as exclusion criteria.

Results and Discussion

Data Clustering Analysis

The clustering of COVID-19 severity was taken from the thesis data of Windy Yoanna Kangdra entitled Clinical Characteristics and Comorbid Factors in Patients Under Surveillance (PDP) for Coronavirus Disease 2019 (COVID-19) at Mitra Medika Amplas Hospital [9]. The data used was 110 patient medical records in March - June 2020 [9]. Furthermore, K-Means algorithm clustering analysis was conducted by the Rapidminer application.

Determine K as the Number of Clusters Formed

The data were clustered using the K-Means algorithm, and in the Rapidminer, the cluster was determined from the smallest Davies-Bouldin Index (DBI) value. Therefore, the smaller the DBI value obtained (non-negative), the better the cluster using the K-Means algorithm. Table1 showed the results obtained by analysing the Davies Bouldin Index (DBI) value.

Table 1. Davies Bouldin Index *(*DBI)

Cluster	DBI Rapidminer
C2	0,094
C3	0,109
C4	0,125
C5	0,137

The results of the DBI values showed (C2;0,094), (C3;0,109), (C4;0,125), and (C5;0,137) with the smallest being 0.094 in cluster 2. Therefore, the smaller the DBI value, the better the cluster and cluster 2 were used. The 110 data and 13 total attributes used were genders, age, occupation, fever, shortness of breath, cough, fever, nausea, and vomiting, diarrhea, abdominal pain, comorbidities, rapid tests, and PCR. Furthermore, the data was used to perform a clustering analysis of the K-Means.

Determining Initial Centroid

The initial centroid is the first cluster center point obtained randomly from the training data. The selection will affect the results of the cluster, and the following are the initial centroid used:

C0 : 1 ;2 ;2 ;1 ;1 ;0 ;1 ;0 ;0 ;0 ;1 ;0 ;1
C1 : 1; 2; 7; 1; 1; 0 ;1 ;1 ;1 ;0 ;1 ;1 ;1

Data Clustering and Machining Iteration

The clustering of data using the closest centroid was based on the comparison of the distance between the data, and the calculation was conducted until the 2nd iteration. In iterations 1 and 2, the data did not change the cluster membership, therefore, the iteration was stopped in the 2nd.

Implementation and Testing

Rapidminer software with the K-Means algorithm was used to group data based on the severity of COVID-19. Furthermore, the test used 110 medical record data which resulted in the following Table 2.

Table 2. Distribution of Cluster Data for COVID-19 patients

Cluster	Patient	Number of Patients
Cluster 1	2,5,6,8,11,13,22,24,25,26,29,30,31,32,33,35,37,43,44,45,46,50,53,63,64,68,69,70,71,73,74,75,76,80,82,84,85,86,90,91,92,93,95,96,98,99,100,101,102	49
Cluster 2	1,3,4,7,9,12,14,15,16,17,18,19,20,21,23,27,28,34,36,38,39,40,41,42,47,48,49,51,52,54,55,56,57,58,59,60,61,62,65,66,67,72,77,78,79,81,83,87,88,89,94,97,103,104,105,106,107,108,109,110	61

Table above showed that the determination of the best number of clusters is seen from the lowest value of the Davies-Bouldin index (DBI) using Rapidminer software, which is 0.094 in cluster 2. Meanwhile, Table 3 showed that C2, namely DBI 0.094, is the smallest value close to 0. Therefore, the smaller the DBI value obtained (non-negative >: 0), the better the cluster from clustering using the K-Means algorithm

E-ISSN: 2720-9326
P-ISSN: 2716-0459

Table 3. Patient Demographic Characteristics

Variable	Description	Amount	Cluster 1	Cluster 2
Gender	1 : male	1 : 65	1 : 29 (59,1%)	1 : 36 (59%)
	2 : female	2 : 45	2 : 20(40,8%)	2 : 25 (40,9%)
Age	1 : < 46 years old	1 : 3	1 : 2 (4,08%)	1 : 1 (1,63%)
	2 : > 46 years old	2 : 107	2 : 47 (96%)	2 : 60 (98,6%)
Occupation	1 : Housewife	1 : 15	1 : 0 (0%)	1 : 15 (24,5%)
	2 : Civil servant	2 : 10	2 : 0 (0%)	2 : 10 (16,4%)
	3 : Entrepreneur	3 : 36	3 : 0 (0%)	3 : 36 (16%)
	4 : Employee	4 : 20	4 : 20 (40,8%)	4 : 0 (0%)
	5 : Teacher	5 : 6	5 : 6 (12,2%)	5 : 0 (0%)
	6 : Farmer	6 : 10	6 : 10 (20,4%)	6 : 0 (0%)
	7 : Retired	7 : 13	7 : 13 (26,5%)	7 : 0 (0%)

Table 3 showed the demographic characteristics of patients in clusters 1 and 2, where the male variables were the most dominant of 20 and 36 patients, respectively. This is consistent with the latest national data dated November 22, 2020, which showed that the percentage of COVID-19 is dominated by 50.6% male and 49.4% female (COVID-19 Handling Task Force, 2020). A study by Zhang et al. (2020) also showed the dominance of the male gender as many as 295 people (51.5%). Subsequently, Guan et al. (2020) found the dominance of the male with a percentage of 58.1% and female at 41.9%.

The age group >46 years was the most dominant with the number in clusters 0 and 1 being 47 and 60 years, respectively. Deny Hidayati's study showed a population that was confirmed positive for COVID-19 in the province of Jakarta as the epicenter of coronavirus transmission attacking all age groups. However, people over 46 years old are more vulnerable than other age groups [10]. The type of occupations in clusters 1 and 2 were dominated by employees and entrepreneurs with 20 and 36 patients, respectively.

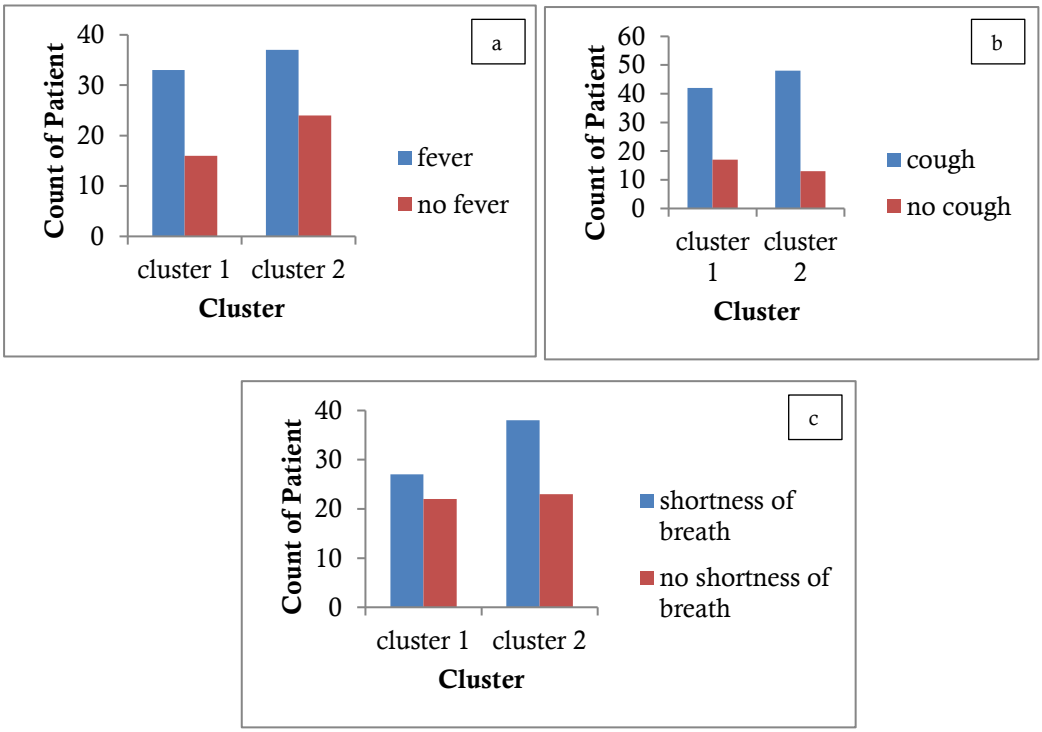

Figure 1. Cluster of Clinical Characteristics of COVID-19 Patients a. Fever b. Cough c. Shortness of breath

Figure 1 showed the clinical characteristics of fever in clusters 1 and 2 involving 42 (85.9%) and 48 (94.1%) patients. In clinical characteristics of cough, 61 (78.6%) and 49 (85.7%) patients had cough

E-ISSN: 2720-9326
P-ISSN: 2716-0459

in cluster 1 and 2. This is consistent with the study of Jacek Baj 2020 where the majority of COVID-19 patients showed general symptoms, such as coughing either with or without phlegm [11]. This is possible because occular infection can be caused through the hands or eyes when coughing. It encourages nasopharyngeal secretions from the nasolacrimal duct to the conjunctival sac [12]. Furthermore, these data were supported by Li et al (2020) which showed that the clinical symptoms of patients upon admission to the hospital, with cough symptoms in 67.8% of COVID-19 patients, fever in 43.8%, and shortness of breath were found in 18.7% of patients [13].

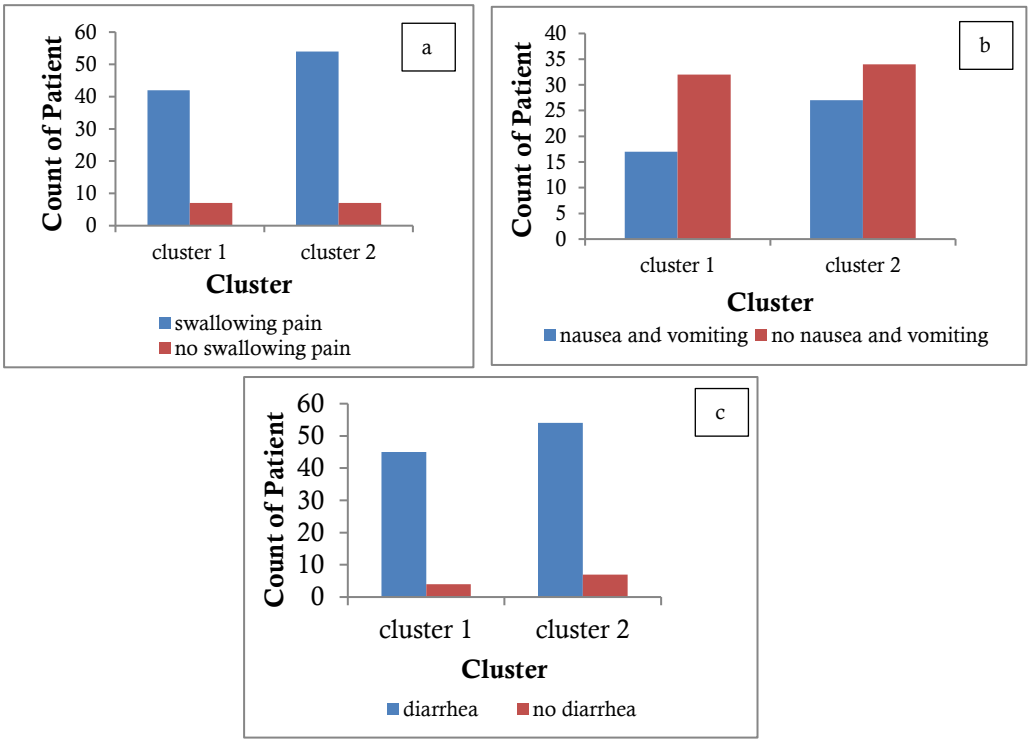

Figure 2. Cluster of Clinical Characteristics of COVID-19 Patients a). Swallowing pain b). Nausea vomiting, and c). Diarrhea

Figure 2 showed the clinical characteristics of patients with swallowing pain in cluster 1 and 2 involving 42 (85.7%) and 54 (88.0%) people respectively. This is in accordance with one of the main symptoms of this virus, such as swallowing pain experienced during the incubation period [14]

The relationship between COVID-19 and dysphagia is finding the COVID-19 virus entering the host cell via angiotensin-converting. Enzyme receptor 2 (ACE-2) in the lower respiratory tract of the human system. ACE-2, which is mainly associated with cardiovascular disease, is a membrane protein expressed in the lungs, heart, kidneys, and intestines, causing disturbances mainly in the pulmonary system, as well as the cardiovascular, gastrointestinal, and hepatic systems [15]. Dysphagia can cause aspiration pneumonia, there is a decrease in the ability to cough, it is important to do this in elderly patients with old age to avoid the severity [16]. Considering the clinical characteristics, 32 (65.3%) and 34 (55.7%) people in cluster 1 and 2 did not experience nausea and vomiting. The results of 29 studies analyzed reported that there were gastrointestinal disturbances as an initial finding in COVID-19 disease, one of which was nausea and vomiting [17]. Manifestations of this symptom can worsen prognoses such as an increased risk of acute respiratory distress syndrome (OR 2.96 [95% Cl 1.17-7.48]; p: 0.020) and liver injury (OR 2.71 [1.52 – 4.83]; P:0.0007) [18]

Figure 2, showed that 45 (91.8%) and 54 (88.5%) patients in cluster 1 and 2 experienced diarrhea. Therefore, there are gastrointestinal symptoms, such as diarrhea in COVID-19 patients. This is different in *MERS-CoV* or *SARS-CoV,* where the experience of this digestive disorders is reduced. Therefore, it is advisable to test urine and fecal samples as well as look for ways to inhibit and reduce the transmission [19].

E-ISSN: 2720-9326
P-ISSN: 2716-0459

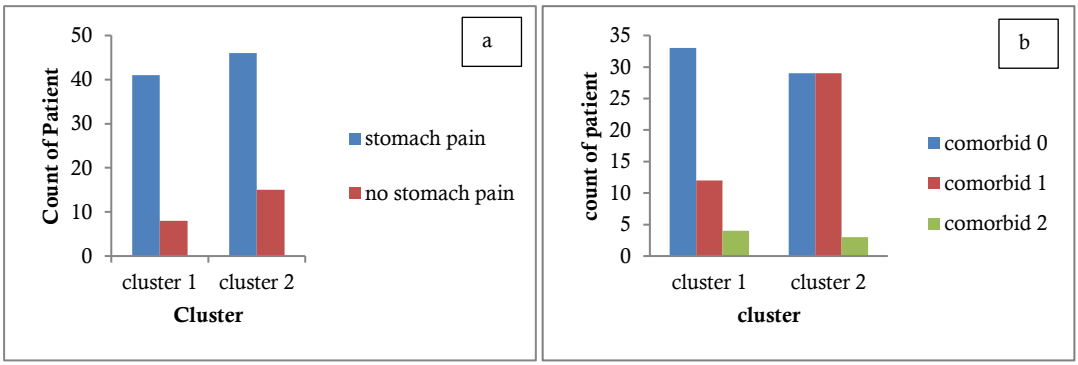

Figure 3. Cluster of Clinical Characteristics of COVID-19 Patients a). Stomach Pain b). Comorbid

Figure 3 showed that 41 (83.6%) and 46 (75%) people experienced abdominal pain in the cluster 1 and 2 of the clinical characteristics of COVID-19 patients. This indicates that most patients experience digestive disorders, one of which is abdominal pain. This is in line with a systematic review study with Meta-Analysis from Mao et al (2020) which found an odds ratio of 7.1. Therefore, patients with severe COVID-19 are 7.1 times more likely to experience abdominal pain than those with mild symptoms [20].

The clinical characteristics of patients with comorbidities in cluster C1 showed that 33 patients did not have comorbidities, while in cluster C2, as many as 29 patients had comorbidities. Furthermore, the medical record data of the disease most commonly found showed Diabetes Mellitus (DM) in as many as 22 patients. This is in line with the study, where patients with comorbidities were dominated by Diabetes Mellitus (20%) followed by hypertension and other heart diseases with a percentage of 15% [21]

Table 4. Comorbid Factors of COVID-19 Patients

Comorbid factors	Number of Patients
Diabetes mellitus	22 (20%)
Hypertension	18 (16,3%)
Cancer	2 (1,8%)
Chronic Obstructive Pulmonary Disease (COPD)	2 (1,8%)
Ischemic Stroke	6 (5,45%)
Pulmonary TB	4 (3,63%)
No comorbid	63 (57,2%)

Table 4 showed the comorbid factors for COVID-19 patients, consisting of the most diseases, namely diabetes mellitus, hypertension, cancer, COPD (Chronic Obstructive Pulmonary Disease), ischemic stroke, and pulmonary TB.

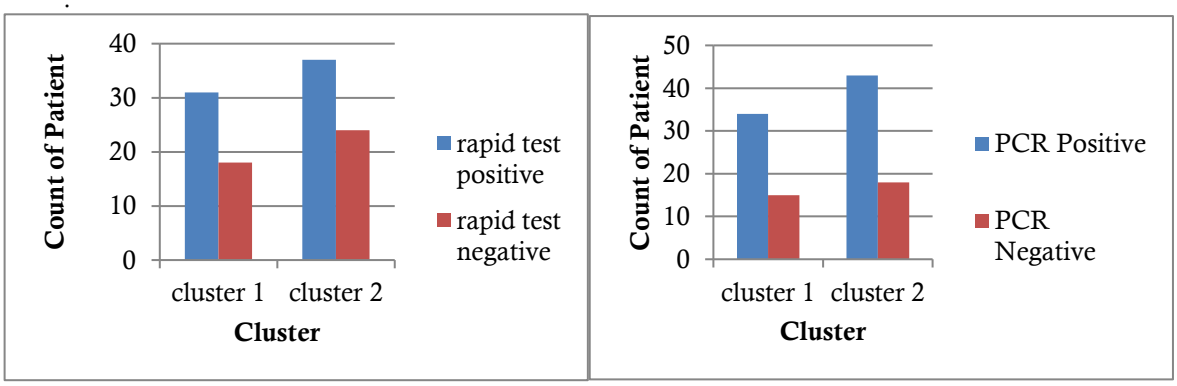

Figure 4. Cluster of Clinical Characteristics of COVID-19 Patients a. Rapid Test b. PCR

Figure 4 showed the clinical characteristics of COVID-19 patients with the results of the rapid test examination dominated by positive results in clusters C1 and C2 for 31 (63.2%) and 37 (60%) patients respectively. In addition, rapid test (RT) is used to detect the early stages of the pandemic in the body. It can also be used for someone without symptoms (OTG) or contact cases from confirmed patients and can detect COVID-19 cases in an area. According to Zainol Rashid and Li, it can be used for the initial screening of the virus to detect IgG and IgM antibodies and the test has a sensitivity of 72.7% and 88.6% [22].

The clinical characteristics of COVID-19 patients with PCR examinations were dominated by positive results, where clusters C1 and C2 consisted of 34 (49%) and 43 (70.4%) people respectively. The use of Rapid Test PCR (RT-PCR) is one of the most recommended methods for detecting the infection. Positive RT PCR is interpreted that the patient is currently infected with the COVID-19 virus and vice versa [23]

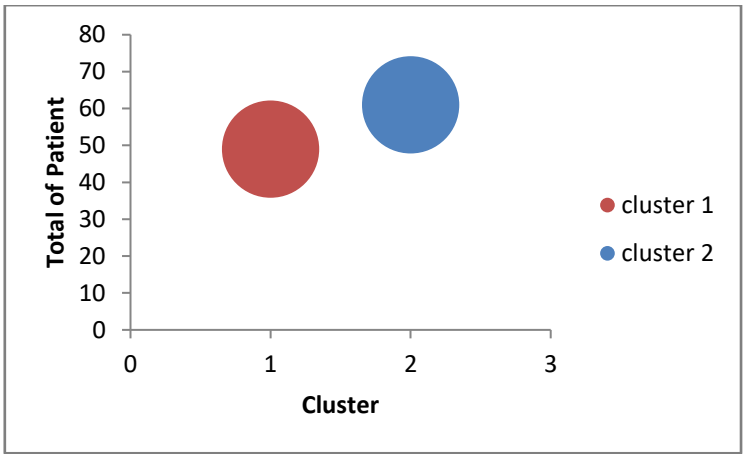

Figure 5. Clusters of COVID-19 Patients Based on the Severity of COVID-19

In Figure 5, the clustering of data using Rapidminer software shows Clusters 1 and 2 with the 2nd iteration, where the center point is no longer changing and there is no data moving between clusters. Cluster 1 shows a moderate degree, this is in accordance with the guidelines for the management of covid-19 edition 4 of 2022 [24], namely having symptoms of covid-19 and no comorbidities. Characteristics in cluster 1 are male dominant, age over 46 years old, majority of employees, symptoms of shortness of breath, cough, swallowing, fever, no nausea and vomiting, diarrhea, abdominal pain, no comorbidities and rapid test and PCR results positive.

Cluster 2 shows a moderate degree seen from the Covid-19 management guidelines edition 4 in 2022 with cluster 2 characteristics [24], namely male sex dominant, age over 46 years old, majority of entrepreneurs, symptoms of shortness of breath, cough, swallowing, fever, no nausea and vomiting, diarrhea, abdominal pain, have at least 1 comorbidity (diabetes, COPD, cancer, COPD, ischemic stroke or tuberculosis) and positive rapid test and PCR results. Therefore, there is a difference between the two where in the mild degree cluster 1 the majority of employees do not have comorbidities, but in the moderate 2 degree cluster the majority are entrepreneurs and have at least 1 comorbidity.

Conclusion

Based on the results of the test with a sample of 110 using Rapidminer, it produced 2 clusters, namely Cluster 1, a mild category with male dominant patient characteristics, aged over 46 years, the majority of employees, had symptoms of COVID-19, positive rapid antigen results, positive PCR but did not have any symptoms. The different severity of COVID-19 and the existence of this clustering can facilitate and speed up the initial examination at the hospital. Thus, patients who are hospitalized with high severity should be given more attention and provide special care. This is one of the efforts to reduce the spread of Covid-19.

References

[1] Y. C. Wu, C. S. Chen, Y. J. Chan, The outbreak of COVID-19: An overview, J. Chinese Med. Assoc., 83(3) (2020) 217–220.
[2] T. H. Sardar, Z. Ansari, An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm, Futur. Comput. Informatics J., 3(2) (2018) 200–209.
[3] X. Li, X. Zhong, Y. Wang, X. Zeng, T. Luo, Q. Liu, Clinical determinants of the severity of COVID-19: A systematic review and meta-analysis, PLoS One, 16(5) (2021) 1-21.
[4] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, M. Rahman, A dynamic K-means clustering for data mining," Indones. J. Electr. Eng. Comput. Sci., 13(2) (2019) 521–526.
[5] M. Azarafza, M. Azarafza, H. Akgün, Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran, J. Appl. Sci. Eng. Technol. Educ., 3(1) (2021) 1-6.
[6] R. A. Indraputra, R. Fitriana, K-Means Clustering Data COVID-19, 10(3) (2020) 275–282.
[7] F. Virgantari, Y. E. Faridhan, K-Means Clustering of COVID-19 Cases in Indonesia's Provinces, 5(2) (2020) 34–39.
[8] G. Liu, T. Wang, L. Yu, Y. Li, J. Gao, The improved research on K-means clustering algorithm in initial values, Proc. - 2013 Int. Conf. Mechatron. Sci. Electr. Eng. Comput., (2013).
[9] W. Y. Kangdra, Karakteristik Klinis Dan Faktor Komorbid Pada Pasien Dalam Pengawasan (Pdp) Coronavirus Disease 2019 (Covid-19) Di Rs Mitra Medika Amplas, 2019 (2021).
[10] D. Hidayati, Profil Penduduk Terkonfirmasi Positif Covid-19 Dan Meninggal: Kasus Indonesia Dan Dki Jakarta, J. Kependud. Indones., 2902 (2020) 93.
[11] Baj, J., Karakuła-Juchnowicz, H. Teresiński, G. Buszewicz, G. Ciesielka, M. Sitarz, R. Forma, A. Karakuła, K. Flieger, W. Portincasa, P. Maciejewski, R., COVID-19: Specific and non-specific clinical manifestations and symptoms: The current state of knowledge, J. Clin. Med., 9(6) (2020) 1–22.
[12] World Health Organization Europe (WHO Europe), Transmission of SARS-CoV-2: implications for infection prevention precautions. Scientific brief, (2020).
[13] Q. Li et al., Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia, N. Engl. J. Med., 382 (2020) 1199-1207.
[14] B. Araújo, D. Domenis, T. Ferreira, C. Merelles, T. Lima, COVID-19 and dysphagia: practical guide to safe hospital care number 1, Audiol. Commun. Res., 25(1) (2020) 1–5.
[15] S. Eyigör, E. Umay, Dysphagia management during covid-19 pandemic: A review of the literature and international guidelines, Turkish J. Phys. Med. Rehabil., 67(3) (2021) 267-274.
[16] S. Carda et al., The role of physical and rehabilitation medicine in the COVID-19 pandemic: The clinician's view, Ann. Phys. Rehabil. Med., 63(6) (2020) 554–556.
[17] H. Xu et al., High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa, Int. J. Oral Sci., 12(1) (2020) 1–5.
[18] M. Goswami, S. Chawla, Time to restart: A comparative compilation of triage recommendations in dentistry during the Covid − 19 pandemic, J. Oral Biol. Craniofacial Res., 10(4) (2020) 374-384.
[19] H. A. Rothan, S. N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, J. Autoimmun., (2020).
[20] Mao R, Qiu Y, He JS, Tan JY, Li XH, Liang J, Shen J, Zhu LR, Chen Y, Iacucci M, Ng SC, Ghosh S, Chen MH, Manifestations and prognosis of gastrointestinal and liver involvement in patients with COVID-19: a systematic review and meta-analysis, Lancet Gastroenterol. Hepatol., 5(7) (2020) 667-678.
[21] C. Huang et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet, 395(10223) (2020) 497–506.
[22] Z. Zainol Rashid, S. N. Othman, M. N. Abdul Samat, U. K. Ali, K. K. Wong, Diagnostic performance of COVID-19 serology assays, Malays J Pathol.,42(1) (2020) 13-21.
[23] H. Tombuloglu, H. Sabit, E. Al-Suhaimi, R. Al Jindan, K. R. Alkharsah, Development of multiplex real-time RT-PCR assay for the detection of SARS-CoV-2, PLoS One, 16(4) (2021) 1-11.
[24] E. Burhan et al., Cedera miokardium pada infeksi COVID-19, (2022).