ATCSJ
Applied Technology and Computing Science Journal

Open Access

p-ISSN 2621-4458 | e-ISSN 2621-4474

Available online to journal2.unusa.ac.id

UNUSA

S4-Accredited – SK No.200/M/KPT/2021

Journal Page is available to https://journal2.unusa.ac.id/index.php/ATCSJ

Implementation of K-Means Algorithm for Diseases Clustering in Elderly Posyandu Participants

Muhammad Iqbal Firdaus[1], Putri Aisyiyah Rakhma Devi[2]

[1,2]*Muhammadiyah Gresik University, Indonesia*

Jl. Sumatera No.101, Gn. Malang, Randuagung, Kec. Kebomas, Gresik

[1*]muhammadiqbalf11@gmail.com, [2]deviaisyiyah@umg.ac.id

Article history:

Received 20 april 2022
Revised 10 May 2022
Accepted 28 May 2022
Available online 10 June 2022

Keywords:

Chronic Diseases
Cluster
Elderly posyandu participants
K-Means Algorithm
Posyandu

Abstract

The Posyandu of Tirem Village is one of the integrated service posts for the elderly, where they can get proper health services. To get the right health services, Posyandu officers group elderly posyandu participants who suffer from chronic diseases for counseling and treatment. The problems that occur during the data collection and counseling process carried out by Posyandu officers are in the calculations that are still basic and are carried out alternately on the elderly Posyandu participants in Tirem village. So this method has the risk of inaccurate data collection and inconsistent handling for the treatment of elderly residents due to different health histories among the elderly. This study aims to classify the data of elderly posyandu participants in Tirem Village who suffer from chronic diseases with predetermined attributes. This grouping process uses the Clustering method using the K-Means Algorithm. The data used in the form of 40 elderly posyandu participant data in October 2022. The results of data processing using the K-Means algorithm with Microsoft Excel tools and using RapidMiner obtained the same results, namely Cluster 1 and cluster_0 have a total of 32 data from 40 data, Cluster 2 and cluster_1 have a total of 8 data from 40 data.

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2021 by author.

I. INTRODUCTION

In the era of globalization, the application of technology is growing rapidly in all areas of life, one of which is in the health sector. The application of this technology can be used to generate data about human diseases. Health is very important for everyone and everyone can experience health problems [1], especially for residents who have entered old age. The Elderly Posyandu in Tirem Village is one of the integrated service posts for the elderly, where they can get proper health services. To get proper health services, Posyandu officers group elderly Posyandu participants with chronic disease for counseling and treatment.

The problems that occurred during the data collection and counseling process carried out by Posyandu officers were in the calculations which were still basic and carried out alternately for elderly Posyandu participants in Tirem Village. So this method has a risk of inaccurate data collection and inconsistent treatment for posyandu staff due to different health histories among elderly posyandu participants. In order

[1] Corresponding author

to make the data management process easier, we need a system that can be used to determine the grouping of chronic diseases based on intermediate and advanced levels in elderly Posyandu participants in Tirem Village [2]. Therefore, it is necessary to analyze the health data of the elderly Posyandu participants in Tirem Village to find out the distribution of the population with a history of chronic diseases. Data analysis can be done in various ways, one of which is using Data Mining.

II. RELATED WORKS

Data mining is a data processing method to find hidden important patterns in data. The results of the data processing can be useful information for the future . Data mining has various methods or even models that can be used. For this study we used the clustering technique. Clustering is a method of grouping data, where each data will be combined into groups that have data similarity characteristics to one another [4]. One of the famous algorithms in the clustering method is K-Means. K-Means is one method for analyzing data. This algorithm determines the number and value of clusters (k) randomly. This value becomes the initial center of the cluster or can be called the centroid [5]. In the clustering method, the K-Means algorithm is popular because the algorithm is relatively simple and efficient to use. This algorithm is included in the unsupervised learning category where we do not need to carry out the training process or in other words there is no learning phase, so it can be applied to various fields. [6].

The studies that have used the K-Means algorithm with clustering techniques have often been used, one of which is for disease grouping problems, including Tanty et al. in 2021 conducted research on the grouping of diseases based on the age of patients who are often affected by the disease at the Bahorok Health Center [7]. Adiputra I., in 2022 conduct research on grouping DHF data at Dharma Kerti Hospital to find out in-depth information about DHF that occurs, so that hospitals can make the right decisions based on data [8]. In 2022, Ariyanto Analyze and classify data using the K-Means algorithm to be able to group data on patients with ARI symptoms quickly and accurately [9]. In 2020, Al-Rizki et al. apply data mining to determine the spread of TB disease with a case study at RSU Aisyiyah Ponorogo [10]. In 2020, Haryadi & Atmaja conducted a study to classify risk levels for heart disease by age using the K-Means Clustering Algorithm [11]. In 2020, Ordila et al. conduct research to find out what types of diseases exist in the PT.Inecda Polyclinic based on region, gender and age classification [12]. In 2020, Sari et al. conducted research to classify areas where tuberculosis was spread in Karawang Regency to find out which areas had high, moderate, and low levels of tuberculosis cases [13]. As for those related to problems in other fields, among others Wahyu & Rushendra in 2022 conducted research to find out how the impact of earthquakes that occurred on the island of Java used the K-Means Algorithm [14]. Qomariasih N. in 2021, grouping or clustering sub-districts in DKI Jakarta province into sub-district groups with very high, moderate, or low cases using the K-Means Clustering method [15]. Suhartini et al., in 2020 grouping drug data based on monthly reports which can be used as a reference for planning drug supplies in the following year [16]. Haviluddin et al., In 2021 grouping to recommend final project research areas for students based on grade data [17]. Normah et al., In 2021 discusses the grouping of clothing stock data to determine sales of which clothes are best selling, selling well and not selling well [18]. Syahputra et al., In 2022 implement K-Means clustering in grouping each sub-district area based on 5 health degree mortality indicator variables to make it easier for the Pontianak City Health Office to know the level of public health in each sub-district area [19]. Based on this, several researchers have used the K-Means Algorithm for the purpose of grouping in several different fields. So that in this study the clustering method was used using the K-Means Algorithm which aims to group data on elderly Posyandu participants in Tirem Village who have chronic diseases. And later it can be useful for Posyandu officers in minimizing errors when collecting data on the diseases of elderly Posyandu participants and can improve the performance of the elderly Posyandu so that they become more efficient and accurate [20].

III. METHODS

This research is a quantitative descriptive study using data obtained from elderly Posyandu participants. The data collection method was carried out by way of direct interviews and data that had been compiled by

posyandu officers who would later process the data using the K-Means Clustering method to group diseases based on middle-level chronic disease groups and advanced chronic disease groups.

At the data collection stage, several criteria were obtained that would be used for data processing. These attributes are NIK, name, date of birth, age, disease and length of sickness. The data obtained is data from the elderly Posyandu in October 2022, which totals 40 data.

The method that will be used in this study is the K-Means Clustering method, where this method will separate data that has the same characteristics into different groups by finding the shortest distance between the data and the centroid point. The first stage of the K-Means algorithm is to determine the number of clusters (k) to be used. After that, determine the initial centroid randomly taken from the data of elderly Posyandu participants. The next step is to calculate the distance of each data to each centroid point using the Euclidean Distance theory as in equation 1.

$$d(x_{i,}\, \mu_i) = \sqrt{\sum_{i=1}^{n} (x_{i,}\, \mu_i)^2} \tag{1}$$

Where, $D(x_{i,}\, y_i)$ is the distance between clustering x and the centroid point y in the i data. While x_i is the i weight in the cluster that you want to find the distance to. And y_i is the i data weight at the centroid point. And n which is the amount of data. After calculating the distance, the next step is to classify the data based on the closest distance between the data and the centroid. Then determine the value of the new initial cluster by calculating the average of the previously calculated clusters using equation 2.

$$C_k = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n} d_i} \tag{2}$$

Where n_k is the sum of all the data in the cluster (k). And d_i is the number of each cluster. After that, we repeat the steps to calculate the distance using the Euclidean Distance theory until the stage of determining the new cluster value. This repetition phase will be stopped if there is no data change in a cluster. While the stages of the K-Means algorithm are shown in Figure 1.

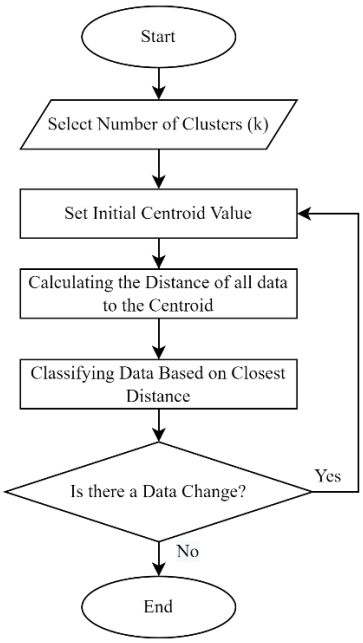

Fig. 1 K-Means Algorithm Flowchart

IV. RESULTS AND DISCUSSIONS

The data that has been obtained from research that comes from participants in the elderly Posyandu in Tirem Village is then stored in an Excel format. Then performed data cleaning to remove attributes that are less relevant. The removed attributes are ID Number, Name, Date of Birth. There is also a Disease Name attribute whose value must be changed into numeric form to make it easier to calculate and replace this attribute into a Disease Code as shown in table 1. The attributes that will be used to group data for elderly Posyandu participants in Tirem Village who suffer from chronic diseases are Age, Disease Code, and Length of Suffering with 40 Data Samples as shown in table 2.

Table 1. Disease data transformation

Disease Code	Disease Name
11	Hypertension
12	Asthma
13	Gastritis
14	Diabetes Mellitus
15	Osteoporosis
16	Hemorrhoids

Table 2. Sample data of elderly posyandu participants

Data no	Age (year)	Disease Code	Length of Sickness (year)
1	65	11	3
2	63	13	5
3	64	12	8
4	64	11	4
5	64	14	5
6	66	16	3
7	62	13	5
8	63	15	10
9	65	11	7
…	…	…	…
40	67	11	4

K-Means Algorithm Implementation

Before performing the calculation process using the K-Means method with the help of Microsoft Excel, in this study 2 clusters will be formed. By determining the initial center of the cluster taken randomly from the sample data in table 2. The initial center of the cluster can be seen in table 3. Then the data distance to the cluster center (centroid) will be calculated using the Euclidean Distance theory as in equation 1 to produce a new centroid. The results of calculating the distance between the data and the centroid can be seen in table 4.

Table 3. Initial centroid

Data No	Age (year)	Disease Code	Length of Sickness (year)
28	70	11	7
36	77	11	5

Table 4. First iteration data distance

Data No	Cluster 1	Cluster 2
1	6,4031	12,1655
2	7,5498	14,1421
3	6,1644	13,3791
4	6,7082	13,0384
5	7	13,3417
6	7,5498	12,2474
7	8,4853	15,1327
8	8,6023	15,3948
9	5	12,1655
…	…	…
40	4,2426	10,0499

The results of calculating the distance of the data in the first iteration which can also be seen in table 4 will select the shortest distance between the data and the nearest centroid. The results of data clustering in the first iteration can be seen in table 5.

Table 5. First iteration cluster results

Data No	C1	C2
1	v	
2	v	
3	v	
4	v	
5	v	
6	v	
7	v	
8	v	
9	v	
…	…	…
40	v	

Based on table 5 above, each cluster already has its own members where member data from each cluster will be calculated using equation 2 to get a new cluster center (centroid), where the results can be seen in table 6.

Table 6. New centroids 1

New Centroid Value			
C1	65	12,3333	5,3636
C2	80	11,8571	7,4286

The process of calculating the next iteration is done in the same way until each data in the cluster does not change. The results of calculating the distance between the data and the centroid in the second iteration can be seen in table 7.

Table 7. Second iteration data distance

Data No	Cluster 1	Cluster 2
1	2,7137	15,6636

2	2,1393	17,2106
3	2,8393	16,0108
4	2,1534	16,3857
5	1,9774	16,3245
6	4,4756	15,2570
7	3,0946	18,1990
8	5,7103	17,4783
9	2,1108	15,0306
…	…	…
40	2,7635	13,4718

The results of calculating the distance of the data in the second iteration which can also be seen in table 7 will select the shortest distance between the data and the nearest centroid. The results of data clustering in the second iteration can be seen in table 8.

Table 8. Second iteration cluster results

Data No	C1	C2
1	v	
2	v	
3	v	
4	v	
5	v	
6	v	
7	v	
8	v	
9	v	
…	…	…
40	v	

Based on table 8 above, each cluster already has its own members where member data from each cluster will be calculated using equation 2 to get a new cluster center (centroid), where the results can be seen in table 9.

Table 9. New centroids 2

New Centroid Value			
C1	64,7813	12,3750	5,2188
C2	79,0	11,75	7,75

In the second iteration, the values of C1 and C2 in table 9 are different from the values of C1 and C2 in the previous iteration, namely in table 6, so the third iteration is performed. The results of calculating the distance between the data and the centroid in the third iteration can be seen in table 10.

Table 10. Third iteration data distance

Data No	Cluster 1	Cluster 2
1	2,6195	14,8029
2	1,9004	16,2827
3	2,9131	15,0042
4	1,9966	15,4798
5	1,8163	15,4151

6	4,4214	14,4784
7	2,8590	17,2663
8	5,7379	16,4810
9	2,2608	14,0401
…	…	…
40	2,8808	12,5946

The results of calculating the distance of the data in the third iteration which can also be seen in table 10 will select the shortest distance between the data and the nearest centroid. The results of data clustering in the third iteration can be seen in table 11.

Table 11. Second iteration cluster results

Data No	C1	C2
1	v	
2	v	
3	v	
4	v	
5	v	
6	v	
7	v	
8	v	
9	v	
…	…	…
40	v	

Based on table 11 above, each cluster already has its own members where member data from each cluster will be calculated using equation 2 to get a new cluster center (centroid), where the results can be seen in table 12.

Table 12. New centroids 2

	New Centroid Value		
C1	64,7813	12,3750	5,2188
C2	79,0	11,75	7,75

In the third iteration, the values of C1 and C2 in table 9 are the same as the values of C1 and C2 in the previous iteration, namely in table 9, so the iteration is stopped.

RapidMiner Implementation

Based on the results obtained from data processing using the K-Means algorithm with Microsoft Excel tools, the results are obtained in the form of membership of each data in each cluster. Where there are 2 clusters formed in this study. Cluster 1 which is a group of medium-level chronic diseases. Cluster 2 which is a group of advanced chronic diseases. Cluster 1 has 32 data and cluster 2 has 8 data.

Data processing will be continued using RapidMiner using the same number of clusters, namely 2 clusters. The processing results can be seen in Figure 2, where cluster_0 has 32 data, cluster_1 has 8 data and a total of 40 data.

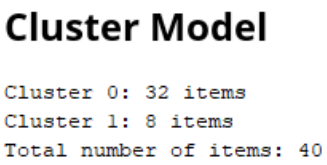

Fig. 2 Cluster results on RapidMiner

The results of calculating the average distance to the cluster can also be seen in Figure 3, where cluster_0 is a group of middle-level chronic diseases, cluster_1 is a group of advanced chronic diseases. From these results no differences were found with the results of data processing assisted by Microsoft Excel which can be seen in table 12. The results of performance measurements using the Davies Bouldin Index (DBI) can be seen in Figure 4

Attribute	cluster_0	cluster_1
Age	64.781	79
Disease Code	12.375	11.750
Length of Sickness	5.219	7.750

Fig. 3 Distance averaging results on RapidMiner

Fig. 4 Davies Bouldin Index (DBI)

From the results of data processing using the K-Means Algorithm with Microsoft Excel and using RapidMiner, the results obtained are Cluster 1 and cluster_0 having a total of 32 data out of 40 data, which has an average age of 64.78 years, with a disease code of 12.38 and long suffered 5.2 years. Cluster 2 and cluster_1 have a total of 8 out of 40 data, which has an average age of 79 years, with a disease code of 11.75 and a duration of 7.8 years.

V. CONCLUSIONS AND RECOMMENDATIONS

The results of data processing that has been carried out through several stages show that the clustering process using the K-Means Algorithm stops at the third iteration because the position of the members in each cluster does not change and gets the same new centroid value. There are 2 clusters formed in this study. Cluster 1 which is a group of medium-level chronic diseases with a total of 32 data and cluster 2 which is a group of advanced chronic diseases with a total of 8 data. Testing using RapidMiner has the same results, that is, each cluster has members that are in accordance with manual calculations assisted by Microsoft Excel. This proves that the K-Means Clustering Algorithm is able to classify data on elderly Posyandu participants who have advanced chronic disease or middle-level chronic disease appropriately and efficiently.

VI. REFERENCES

[1] A. Bastian, H. Sujadi, and G. Febrianto, "PENERAPAN ALGORITMA K-MEANS CLUSTERING ANALYSIS PADA PENYAKIT MENULAR MANUSIA (STUDI KASUS KABUPATEN MAJALENGKA)," *Jurnal Sistem Informasi*, vol. 14, no. 1, pp. 26–32, 2018.

[2] C. A. Sugianto, A. H. Rahayu, and A. Gusman, "Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Cigugur Tengah," *Journal of Information Technology*, vol. 2, no. 2, pp. 39–44, Aug. 2020, doi: 10.47292/joint.v2i2.30.

[3] Y. Prayoga, H. S. Tambunan, and I. Parlina, "Penerapan Clustering Pada Laju Inflasi Kota Di Indonesia Dengan Algoritma K-Means," *BRAHMANA: Jurnal Penerapan Kecerdasan Buatan*, vol. 1, no. 1, pp. 24–30, Dec. 2019, doi: 10.30645/brahmana.v1i1.4.

[4] M. Simanjuntak, E. Manik, and P. Ratna Sari, "PENERAPAN DATA MINING PENGELOMPOKKAN PENYAKIT MENULAR SEKSUAL (PMS) MENGGUNAKAN METODE CLUSTERING," *Jurnal Mahajana Informasi*, vol. 4, no. 1, 2019.

[5] J. Nasir, "PENERAPAN DATA MINING CLUSTERING DALAM MENGELOMPOKAN BUKU DENGAN METODE K-MEANS," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 11, no. 2, pp. 690–703, Oct. 2021, doi: 10.24176/simet.v11i2.5482.

[6] M. R. Nugroho, I. E. Hendrawan, and P. P. Purwantoro, "Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI," *NUANSA INFORMATIKA*, vol. 16, no. 1, pp. 125–133, Jan. 2022, doi: 10.25134/nuansa.v16i1.5294.

[7] T. Tanty, B. S. Ginting, and M. Simanjuntak, "Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode K-Means Clustering (Studi Kasus: Puskesmas Bahorok)," *ALGORITMA: JURNAL ILMU KOMPUTER DAN INFORMATIKA*, vol. 5, no. 2, 2021.

[8] I. N. M. Adiputra, "CLUSTERING PENYAKIT DBD PADA RUMAH SAKIT DHARMA KERTI MENGGUNAKAN ALGORITMA K-MEANS," *INSERT : Information System and Emerging Technology Journal*, vol. 2, no. 2, p. 99, Jan. 2022, doi: 10.23887/insert.v2i2.41673.

[9] D. Ariyanto, "Data Mining Menggunakan Algoritma K-Means untuk Klasifikasi Penyakit Infeksi Saluran Pernafasan Akut," *Jurnal Sistim Informasi dan Teknologi*, pp. 13–18, Feb. 2022, doi: 10.37034/jsisfotek.v4i1.117.

[10] M. F. I. Al-Rizki, I. Widaningrum, and G. A. Buntoro, "Prediksi Penyebaran Penyakit TBC dengan Metode K-Means Clustering Menggunakan Aplikasi Rapidminer," *JTERA (Jurnal Teknologi Rekayasa)*, vol. 5, no. 1, p. 1, Jul. 2020, doi: 10.31544/jtera.v5.i1.2019.1-10.

[11] D. Haryadi and D. M. U. Atmaja, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Tingkat Risiko Penyakit Jantung," *Journal of Informatics and Communication Technology (JICT)*, vol. 3, no. 2, pp. 51–66, Dec. 2021, doi: 10.52661/j_ict.v3i2.85.

[12] R. Ordila, R. Wahyuni, Y. Irawan, and M. Yulia Sari, "PENERAPAN DATA MINING UNTUK PENGELOMPOKAN DATA REKAM MEDIS PASIEN BERDASARKAN JENIS PENYAKIT DENGAN ALGORITMA CLUSTERING (Studi Kasus : Poli Klinik PT.Inecda)," *Jurnal Ilmu Komputer*, vol. 9, no. 2, pp. 148–153, Oct. 2020, doi: 10.33060/JIK/2020/Vol9.Iss2.181.

[13] Y. P. Sari, A. Primajaya, and A. S. Y. Irawan, "Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang," *INOVTEK Polbeng - Seri Informatika*, vol. 5, no. 2, p. 229, Nov. 2020, doi: 10.35314/isi.v5i2.1457.

[14] A. Wahyu and R. Rushendra, "Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 8, no. 1, pp. 174–179, 2022.

[15] N. Qomariasih, "Implementasi K-Means Clustering Analysis untuk Mengelompokkan Kelurahan-Kelurahan Di DKI Jakarta Berdasarkan Jumlah Positif Covid-19.," *Jurnal Syntax Transformation*, vol. 2, no. 7, pp. 1003–1011, Jul. 2021, doi: 10.46799/jst.v2i7.336.

[16] S. Suhartini, L. Kerta Wijaya, and N. Arini Pratiwi, "Penerapan Algoritma K-Means Untuk Pendataan Obat Berdasarkan Laporan Bulanan Pada Dinas Kesehatan Kabupaten Lombok Timur," *Infotek : Jurnal Informatika dan Teknologi*, vol. 3, no. 2, pp. 147–156, Aug. 2020, doi: 10.29408/jit.v3i2.2315.

[17] H. Haviluddin, S. J. Patandianan, G. M. Putra, N. Puspitasari, and H. S. Pakpahan, "Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir," *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, vol. 16, no. 1, p. 13, Mar. 2021, doi: 10.30872/jim.v16i1.5182.

[18] N. Normah, S. Nurajizah, and A. Salbinda, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," *Jurnal Teknik Komputer*, vol. 7, no. 2, pp. 158–163, Jul. 2021, doi: 10.31294/jtk.v7i2.10553.

[19] I. Syahputra, I. Ilhamsyah, S. Rahmayuda, and F. Febrianto, "SISTEM KLASTERISASI DATA KESEHATAN PENDUDUK UNTUK MENENTUKAN RENTANG DERAJAT KESEHATAN DAERAH MENGGUNAKAN K-MEANS," *Jurnal Khatulistiwa Informatika*, vol. 10, no. 1, pp. 66–73, Jul. 2022, doi: 10.31294/jki.v10i1.12872.

[20] Isy Karima Fauzia, Budi Arif Dermawan, and Tesa Nur Padilah, "Penerapan K-Means Clustering pada Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Kabupaten Karawang," *Jurnal Sistem dan Informatika (JSI)*, vol. 15, no. 1, pp. 81–87, Nov. 2020, doi: 10.30864/jsi.v15i1.350.