Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2023. Vol. 3. No. 1; e-ISSN : 2808-4519

Data Mining Grouping Areas Of Diarrhea Disease In The City Medan Using K-Means (Clustering) Algorithm In Service Environment In Medan City

Borkat Parulian Pohan[1], Marto Sihombing[2], Suci Ramadani[3]

[1,2,3]Informatics Enginering, STMIK Kaputama
Jl. Veterans No. 4A-9A, Binjai, North Sumatra, Indonesia
Borkatpohan14@gmail.com[1]*

Abstract

Diarrhea is a contagious disease and is characterized by symptoms such as changes in the shape and consistency of the stool from becoming soft to liquefying and an increase in the frequency of defecation more than usual accompanied by vomiting, thus causing sufferers to experience a lack of fluids in the body or dehydration which in In the end, if you don't get help immediately it can cause seriousness and even death. Diarrhea can attack all ages with symptoms such as frequent bowel movements with a liquid or watery stool consistency, signs and symptoms of dehydration, fever, vomiting, anorexia, weakness, paleness, changes in vital signs (rapid pulse and breathing), urine output. decreased or absent. Of the 20 data grouping of diarrheal disease areas using the clustering method with the k-means algorithm, 3 groups were obtained, where the most cluster members were in cluster 1 with a total of 11 data centered on centroids 2, 5, 3, namely data on the spread of diarrhea in 6-10 sub-districts, the number 15,000-20,000 sufferers with sporadic distribution status.

Keywords: *Data Mining, Medan City Environment Service, K-means algorithm*

1. Introduction

Diarrhea is a contagious disease and is characterized by symptoms such as changes in the shape and consistency of the stool from becoming soft to liquefying and an increase in the frequency of defecation more than usual accompanied by vomiting, thus causing sufferers to experience a lack of fluids in the body or dehydration which in turn In the end, if you don't get help immediately it can cause seriousness and even death. Diarrhea can attack all ages with symptoms such as frequent bowel movements with a liquid or watery stool consistency, signs and symptoms of dehydration, fever, vomiting, anorexia, weakness, paleness, changes in vital signs (rapid pulse and breathing), urine output. decreased or absent. Diarrhea can occur due to poor hygiene and sanitation, malnutrition, crowded environments and poor medical resources. In 2018, the number of diarrhea sufferers of all ages (SU) who were served in health facilities. As population density increases in the city of Medan, diarrhea sufferers also increase due to the lack of public awareness of the importance of health and the environment. There is a need to group population health data to determine public health. In order to provide education at the center of the spread of Diarrhea disease, it can run well. So it is important and necessary to cluster areas where diarrheal disease spreads.

2. Metodologi

2.1. Data Mining

Data mining is the process of extracting useful information and patterns from very large data. Data mining includes data collection, data extraction, data analysis, and data statistics. Data mining is also known as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, and others. Data mining bertujuan untuk menemukan pola yang sebelumnya tidak diketahui. Jika pola-pola tersebut telah diperoleh maka dapat digunakan untuk menyelesaikan berbagai macam permasalahan. According to Erwin (2017, h.1) data mining is one of the main parts or processes of Knowledge Discovery in Database (KDD) whose form of activity is collecting and using past data to find regularities, patterns or relationships in a larger data set. Broadly speaking, KDD includes three stages, namely pre-processing, process (data mining) and post-processing .Data mining has now become a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. Through data mining, companies can find information in large amounts of data through precise and effective processing with various methods available in data mining so that in simple terms data mining can be described as a pattern or model or rule or knowledge generated from data mining, such as Figure 1.

Figure 1: Model or Knowledge Is the Output of Data Mining

2.2 Algoritma K-Means

The K-Means algorithm is a relatively simple algorithm for classifying or grouping a large number of objects with certain attributes into groups (clusters) of K. In the K-Means algorithm, the number of K clusters is predetermined. According to Baginda Harahap (2021) states that K-Means is a non-hierarchical grouping method that tries to partition data into clusters or groups so that data with the same characteristics will be included in the same cluster and data with different characteristics are grouped into other groups.

2.3 Clustering

Clustering is a data analysis method, which is often included as a data mining method, the purpose of which is to group data with the same characteristics. According to (Zulfadhillah et al., 2016) cluster analysis (clustering) is finding objects in one group that are the same (have a relationship) with others and are located (points have a relationship) with objects in other groups. The main goal of the Clustering method is to group a number of data or objects into clusters (groups) so that each cluster contains data that is as similar as possible. The clustering method seeks to place similar (closely spaced) objects in one group and make the distance between groups as far as possible. This means that objects in one group are very similar to each other and different from objects in other groups as can be seen in Figure II. 2.

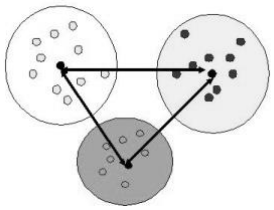

Figure 2: Examples of Clusters Formed

2.4 Diarrheal diseases

Diarrhea is a condition characterized by excessive loss of electrolytes through feces. Diarrhea is a condition where defecation occurs as usual with a frequency of > 3 times a day with the consistency of soft or liquid stools. Diarrhea is a disease characterized by changes in the form and concentration of stools that are softened to liquid with a frequency of more than five times a day. Diarrhea can be a very acute and dangerous disease because it often results in death if treated too late.

2.5 Causes of Diarrhea

Diarrhea can be caused by various infections, in addition to other causes such as malabsorption. Factors causing diarrhea are as follows.

1. Infection Factors
   Internal infections, namely infections of the digestive tract, are the main cause of diarrhea in children.
2. Malabsorption Factors
   Carbohydrate malabsorption: discharides (lactose, maltose and sucrose intolerance), monosaccharides (glucose, fructose and galactose intolerance). The most important and common problem in babies and children is lactrose intolerance.
3. Food Factor
   Food factors, namely: stale food, toxic, food allergies.
4. Psychological Factors
   Fear and anxiety. Although it is rare, it can cause diarrhea, especially in older children.

2.6 Diarrhea Classification

Acute diarrhea is diarrhea that attacks and usually lasts less than 14 days. The consequence that will arise from acute diarrhea is dehydration, while dehydration is the main cause of death for diarrhea sufferers.

2.7 Signs and Symptoms of Diarrhea

The clinical picture of diarrheal disease begins with the patient being whiny, restless, body temperature usually increases, appetite is reduced or absent, then diarrhea occurs. Liquid stools, possibly with mucus or mucus and blood. The color of the stool changes over time to greenish because it is mixed with bile. The anus and the surrounding area develop blisters because of the bowels and stools that are increasingly acidic as a result of more and more lactic acid, which comes from lactose, which is not absorbed by the intestine during diarrhea. Symptoms of vomiting can occur before or after diarrhea and can be caused by an inflamed stomach or due to disturbances in acid-base and electrolyte balance. When the patient has lost a lot of fluids and electrolytes, symptoms of dehydration begin to appear,

namely weight loss, reduced turgor, eyes and large fontanel become sunken (in infants), the mucous membranes of the lips and mouth and skin appear dry.

3. Results and Discussion

3.1. Calculation Clustering

In using the clustering method, the initial process carried out to form clusters is to transform data into numeric form with predetermined codes, then determine the number of groups (K), calculate the centroids, calculate the use of objects to the centroids and then group them based on work and place of residence against complaints that are felt, if no objects move or groups then the iteration is complete. Then transform the criteria data above to be calculated using the clustering method. The data transformation from the data above can be seen in the table below.

Table 1: Transformation data

No	Object	X	Y	Z
1	A	2	2	2
2	B	3	2	3
3	C	2	5	4
4	D	2	5	3
5	E	2	6	4
6	F	2	6	4
7	G	2	2	4
8	H	3	3	3
9	I	2	2	2
10	J	3	5	3
11	K	2	2	2
12	L	2	3	3
13	M	1	6	4
14	N	2	6	4
15	O	2	5	3
16	P	1	6	4
17	Q	2	5	3
18	R	2	5	3
19	S	2	6	4
20	T	2	5	3

Then form a cluster into 3 groups (K = 3) and determine the centroid center point. The clustering calculation process is as follows.
K=3 Centroid
C1= (2, 5, 4) taken from the data C
C2= (3, 3, 3) taken from the data H
C3= (1, 6, 4) taken from the data M
Then do the calculations like the calculation process below: Iteration 1 :

1. A (2,2,2)
   C_1= (2,5,4) = $\sqrt{(2-2)^2+(2-5)^2+(2-4)^2} = 3{,}61$
   C_2= (3,3,3)= $\sqrt{(2-3)^2+(2-3)^2+(2-32)^2} = 1{,}73$
   C_3= (1,6,4) = $\sqrt{(2-1)^2+(2-6)^2+(2-4)^2} = 4{,}78$
2. B (3,2,3)
   C_1= (2,5,4) = $\sqrt{(3-2)^2+(2-5)^2+(3-4)^2} = 3{,}32$
   C_2= (3,3,3)= $\sqrt{(3-3)^2+(2-3)^2+(3-3)^2} = 1$
   C_3= (1,6,4) = $\sqrt{(3-1)^2+(2-6)^2+(3-4)^2} = 4{,}58$
3. C (2,5,4)
   C_1= (2,5,4) = $\sqrt{(2-1)^2+(5-2)^2+(2-3)^2} = 0$
   C_2= (3,3,3)= $\sqrt{(2-3)^2+(2-1)^2+(2-2)^2} = 2{,}45$
   C_3= (1,6,4) = $\sqrt{(2-2)^2+(2-4)^2+(2-3)^2} = 1{,}41$
4. D (2,5,3)
   C_1= (2,5,4) = $\sqrt{(2-2)^2+(5-5)^2+(3-4)^2} = 1$
   C_2= (3,3,3)= $\sqrt{(2-3)^2+(3-1)^2+(3-3)^2} = 2{,}24$
   C_3= (1,6,4) = $\sqrt{(2-1)^2+(5-6)^2+(3-4)^2} = 1{,}73$
5. E (2,6,4)
   C_1= (2,5,4) = $\sqrt{(2-2)^2+(6-5)^2+(4-4)^2} = 1$
   C_2= (3,3,3)= $\sqrt{(2-3)^2+(6-3)^2+(4-3)^2} = 3{,}32$
   C_3= (1,6,4) = $\sqrt{(2-1)^2+(6-6)^2+(4-4)^2} = 1$
6. F (2,6,4)

C_1= (2,5,4) $= \sqrt{(2-2)^2+(6-5)^2+(4-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(6-3)^2+(4-3)^2} = 3{,}32$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(6-6)^2+(4-4)^2} = 1$

7. G (2,2,4)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(2-5)^2+(4-4)^2} = 3$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(2-3)^2+(4-3)^2} = 1{,}73$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(2-6)^2+(4-4)^2} = 4{,}12$
8. H (3,3,3)
C_1= (2,5,4) $= \sqrt{(3-2)^2+(3-5)^2+(3-4)^2} = 2{,}45$
C_2= (3,3,3)= $\sqrt{(3-3)^2+(3-3)^2+(3-3)^2} = 0$
C_3= (1,6,4) = $\sqrt{(3-3)^2+(3-3)^2+(3-3)^2} = 3{,}74$
9. I (2,2,2)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(2-5)^2+(2-4)^2} = 3{,}61$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(2-3)^2+(2-3)^2} = 1{,}73$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(2-6)^2+(2-4)^2} = 4{,}58$
10. J (3,5,3)
C_1= (2,5,4) $= \sqrt{(3-2)^2+(2-5)^2+(1-4)^2} = 1{,}41$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(5-3)^2+(4-3)^2} = 2$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(5-6)^2+(4-4)^2} = 2{,}45$
11. K (2,2,2)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(2-5)^2+(2-4)^2} = 3{,}61$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(2-3)^2+(2-3)^2} = 1{,}73$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(2-6)^2+(2-4)^2} = 4{,}58$
12. L (2,3,3)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(3-5)^2+(3-4)^2} = 2{,}24$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(3-3)^2+(3-3)^2} = 1$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(3-6)^2+(3-4)^2} = 3{,}32$
13. M (1,6,4)
C_1= (2,5,4) $= \sqrt{(1-2)^2+(6-5)^2+(4-4)^2} = 1{,}41$
C_2= (3,3,3)= $\sqrt{(1-3)^2+(6-3)^2+(4-3)^2} = 0$
C_3= (1,6,4) = $\sqrt{(1-1)^2+(6-6)^2+(4-4)^2} = 3$
14. N (1,3,1)
C_1= (2,5,4) $= \sqrt{(1-2)^2+(3-5)^2+(1-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(1-3)^2+(3-3)^2+(1-3)^2} = 3{,}32$
C_3= (1,6,4) = $\sqrt{(1-1)^2+(3-6)^2+(1-4)^2} = 1$
15. O (2,6,4)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(6-5)^2+(4-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(6-3)^2+(4-3)^2} = 2{,}24$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(6-6)^2+(4-4)^2} = 1{,}73$
16. P (1,6,4)
C_1= (2,5,4) $= \sqrt{(1-2)^2+(6-2)^2+(4-4)^2} = 1{,}41$
C_2= (3,3,3)= $\sqrt{(1-3)^2+(6-3)^2+(4-3)^2} = 3{,}74$
C_3= (1,6,4) = $\sqrt{(1-1)^2+(6-6)^2+(4-4)^2} = 0$
17. Q (2,5,3)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(5-5)^2+(3-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(5-3)^2+(3-3)^2} = 2{,}24$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(5-6)^2+(3-4)^2} = 1{,}73$
18. R (2,5,3)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(5-5)^2+(3-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(5-3)^2+(3-3)^2} = 2{,}24$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(5-6)^2+(3-4)^2} = 1{,}73$
19. S (2,6,4)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(6-5)^2+(4-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(6-3)^2+(4-3)^2} = 3{,}32$
C_3= (1,6,4) = $\sqrt{(2-1)^2+(6-6)^2+(4-4)^2} = 1$
20. T (2,5,3)
C_1= (2,5,4) $= \sqrt{(2-2)^2+(5-5)^2+(3-4)^2} = 1$
C_2= (3,3,3)= $\sqrt{(2-3)^2+(5-3)^2+(3-3)^2} = 2{,}24$

C_3= (1,6,4) = $\sqrt{(2-1)^2 + (5-6)^2 + (3-4)^2} = 1{,}73$

From the calculation above, the results of the iteration 1 calculation are obtained, which are as shown in the table below.

Table 2: Iteration Result Data 1

No	X	Y	Z	C1	C2	C3	Group
1	2	2	2	3,61	1,73	4,58	2
2	3	2	3	3,32	1,00	4,58	2
3	2	5	4	0,00	2,45	1,41	1
4	2	5	3	1,00	2,24	1,73	1
5	2	6	4	1,00	3,32	1,00	1
6	2	6	4	1,00	3,32	1,00	1
7	2	2	4	3,00	1,73	4,12	2
8	3	3	3	2,45	0,00	3,74	2
9	2	2	2	3,61	1,73	4,58	2
10	3	5	3	1,41	2,00	2,45	1
11	2	2	2	3,61	1,73	4,58	2
12	2	3	3	2,24	1,00	3,32	2
13	1	6	4	1,41	3,74	0,00	3
14	2	6	4	1,00	3,32	1,00	1
15	2	5	3	1,00	2,24	1,73	1
16	1	6	4	1,41	3,74	0,00	3
17	2	5	3	1,00	2,24	1,73	1
18	2	5	3	1,00	2,24	1,73	1
19	2	6	4	1,00	3,32	1,00	1
20	2	5	3	1,00	2,24	1,73	1

After calculating using the existing cluster formula, the groups based on the minimum distance to the nearest centroid are:

Group Lama : (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

Group Baru : (2 2 1 1 1 1 2 2 2 1 2 2 3 1 1 3 1 1 1 1)

There is a group change, followed by the following iteration:

Table 3: Iteration Result Data 2

No	X	Y	Z	C1	C2	C3	Group
1	2	2	2	3,58	0,75	4,58	2
2	3	2	3	3,47	0,88	4,58	2
3	2	5	4	0,67	3,09	1,41	1
4	2	5	3	0,50	2,82	1,73	1
5	2	6	4	0,92	4,02	1,00	1
6	2	6	4	0,92	4,02	1,00	1
7	2	2	4	3,35	1,33	4,12	2
8	3	3	3	2,54	1,17	3,74	2
9	2	2	2	3,58	0,75	4,58	2
10	3	5	3	1,12	2,93	2,45	1
11	2	2	2	3,58	0,75	4,58	2
12	2	3	3	2,33	0,88	3,32	2
13	1	6	4	1,36	4,19	0,00	3
14	2	6	4	0,92	4,02	1,00	1
15	2	5	3	0,50	2,82	1,73	1
16	1	6	4	1,36	4,19	0,00	3
17	2	5	3	0,50	2,82	1,73	1
18	2	5	3	0,50	2,82	1,73	1
19	2	6	4	0,92	4,02	1,00	1
20	2	5	3	0,50	2,82	1,73	1

After calculating using the cluster formula in iteration 3, the groups based on the minimum distance to the nearest centroid are:

Group Lama : (2 2 1 1 1 1 2 2 2 1 2 2 3 1 1 3 1 1 1 1)

Group Baru : (2 2 1 1 1 1 2 2 2 1 2 2 3 1 1 3 1 1 1 1)

The following is a cluster graph based on the calculation of the results of data mining iterations of grouping patient data based on work and place of residence for the complaints that are felt. The graphs obtained are as follows:

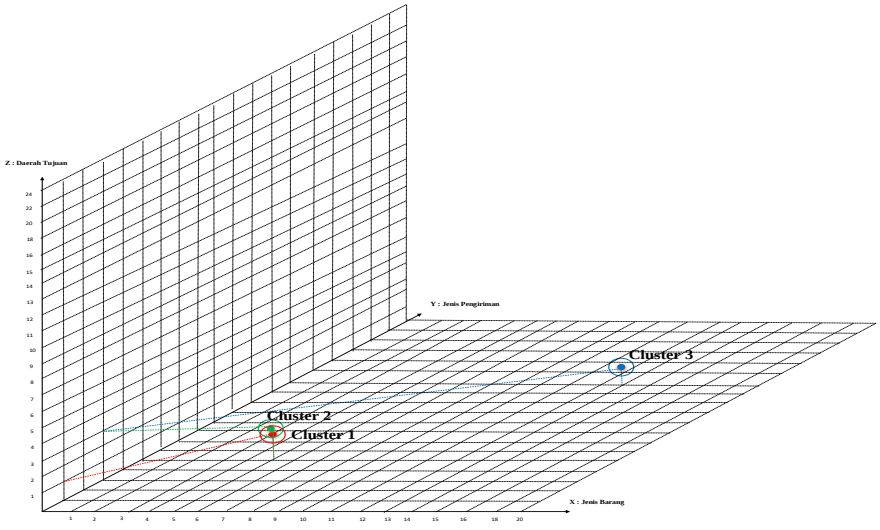

Figure 3: Cluster Chart

4. Conclusion

Of the 20 data groupings of diarrheal disease distribution areas, 3 groups were obtained, cluster 1 contained 11 data, cluster 2 contained 7 data, and cluster 3 contained 2 data.Cluster 1 There are 10 Data

1. Cluster 1 There are 11 Data
   It can be seen that cluster 1 is centered on 2, 5, 3, namely data on the spread of diarrhea in 6-10 sub-districts, the number of sufferers is 15,000 to 20,000 people with sporadic distribution status.
2. Cluster 2 There are 7 Data
   It can be seen that cluster 2 is centered on 2, 2, 3, namely data on the spread of diarrhea to 6-10 villages, the number of sufferers is 1,001 to 5,000 people with sporadic distribution status.
3. Cluster 3 There are 2 Data
   It can be seen that cluster 3 is centered on 1, 6, 4, namely data on the spread of diarrhea in 1-5 sub-districts, the number of sufferers is >20,000 people with endemic distribution status and requires more attention from the Medan city government.

References

[1] Anggit Hidayah, 2021, Implementasi data mining untuk clustering daerah penyebaran penyakit demam berdarah di Kota Tangerang selatan menggunakan algoritma K-Means, *Jurnal: Sistem Informasi Universitas Duta Bangsa*
[2] Arhami,Muhammad dan Muhammad Nasir. 2020. *Data Mining Algoritma dan Implementasi*. Yogyakarta: CV Andi Offset
[3] Bambang Riyanto, 2019, Penerapan Algoritma K-Medois Clustering Untuk Pengelompokan Penyakit Diare di Kota Medan, *Jurnal: Teknik nformatika STMIK Budi Darma Medan*
[4] Eko Prastyo, 2012, *Data Mining Konsep dan Aplikasi Menggunakan Matlab*, Yogyakarta: CV Andi Ofset
[5] Erwin Widiasmoro, 2017, *Inovasi Pembelajaran*, Yogyakarta: Parama Ilmu
[6] Fajar Astuti Hermawati, 2013, Data Mining, Yogyakarta: CV Andi Ofset
[7] Pudiastuti, 2011, *Penyakit Pemicu stroke*, Yogyakarta: Nuha Medika
.