Journal of Medical Records and Information Technology (JOMRIT)

Vol.2, No. 2, December 2024,
ISSN: XXX-XXXX,

Application of Data *Mining* with *K-Means Clustering* for Disease Classification at the Pagambiran Health Center

Rosi Rahmadhani[1] , Muhammed Ihksan[2] , Chamy Rahmatiqa[3]
[123]University of Syedza Saintika, Padang, West Sumatra, Indonesia

Article Info

Article history:

Received October 08, 2024
Revision October 12, 2024
Accepted December 28, 2024

Keywords:

K-Means Medical Records
Data *Mining*
Classification
Disease

ABSTRACT

The Pagambiran Health Center is a place for public health services in Pagambiran Village, Padang City. Every day, the Pagambiran Health Center serves various community diseases. Many diseases and with different symptoms from the community make it difficult for the health center to determine the right counseling or socialization in a disease caused. Proper counseling will minimize the possibility of people contracting a disease. In addition, the health center still has difficulties inmaking decisions so that the results of counseling and socialization are not optimal. This study aims to produce a group of patient disease data and know the parameters of each patient's disease. The data used in this study is the patient's disease data that is in the patient's medical record data. In this study, the data used was patient disease data for a period of 5months consisting of 268 patient disease categories. The data was processed using *the K-Means* method. The *K-Means* method groups data with a partition system. The results obtained were patient diseases as many as 4 *clusters*, where *cluster* 1 was a disease that occurred with highintensity as many as 1 category of disease, *cluster* 2 was a disease that occurred with moderate intensity as many as 7 categories of diseases, cluster 3 was a disease that occurred with low intensity as many as 45 categories of diseases, and *cluster* 4 are diseases with very low/rare intensity as many as 215 disease categories. This study produces classification based on patient diseases so that the health center can carryout prevention and provide community services effectively because this study has produced models and considerations to determine the right treatment or countermeasure to the symptoms of community diseases.

Corresponding Author:

Anisa Mafera
Health Information Management Study Program, University of Syedza Saintika
Address : Street Prof. Dr. Hamka No. 228 Air Tawar Timur, Padang, West Sumatra, Indonesia
E-mail: anisamaghfirajaya@gmail.com

1. INTRODUCTION

Puskesmas is one of the government institutions engaged in public health services at the sub-district level. One of the roles of health centers is in supporting the performance of health institutions above them, namely hospitals, as an effort to prevent and overcome public health problems. In an effort to improve the quality of better health services at the health center level in particular. [1]

Medical records or can be called ICD (*International Classification Diseases*) are records of patients who have been treated in hospitals or health centers. The medical language that is usually used by doctors in identifying a disease and then providing action for the disease is in the form of medical language (medical records) which are then encoded into ICD codes. This code is a standard language that is commonly used by

doctors in general, even though they are not specialists, in reading this code in accordance with the rules that have been set. [2]

Knowledge Discovery in Database (KDD) is a method or model used to obtain hidden knowledge from existing databases, the knowledge obtained can be a reference in making a decision [3] . *Databases* that are generally only used to store data but with KDD can find a hidden knowledge in it, from that knowledge can beused as a benchmark or guideline in analyzing something or taking an action. Broadly speaking, KDD consistsof several steps, namely data *set cleaning*, data *integration*, data *selection, transformation of* data, and data *mining.*

Data *mining* can be used to dig up information and hidden values from a data set, where this information cannot be found if done manually. Data [2] *Mining* is one of the important steps in the *Knowledge Discovery in Database process*. This stage can also be called the core process of KDD. Data *mining* is a process that uses statistical, mathematical, artificial intelligence, and *machine learning* techniques to extract and identify useful information and related knowledge from large *databases*. The term data *mining* has the essence of a discipline whose main purpose is to find, excavate, or mine knowledge from data or information owned.

K-Means is a partitioning method that is often used in the adjustment of *macine studies* and the analysis of a data pattern, this algorithm is highly affected by the *initial centroid*, but cannot guarantee until the final solution is found because the initial *centroid* is randomly searched for *a given* cluster. [4] *K-Means Clustering* is a non-hierarchical data *clustering* method that partitions available data into one or several *clusters*, so that it can produce data groups that have the same characteristics or *clusters* and data that have different characteristics are grouped into other groups. In determining the number of clusters to be used, the *elbow* method (elbow point) will be used to generate information in determining the value of K or the best number of *clusters* by paying attention to the *cluster* that has the most obvious visible angle. [5]

Every day, the Pagambiran health center serves various types of people. Many diseases and with different symptoms from the community make it difficult for the health center to determine the right counseling or socialization in a disease caused. Proper counseling will minimize the possibility of people contracting a disease. The patient's initial diagnosis can be a reference in knowing the cause of a disease. In addition, the health center still has difficulties in making decisions so that the results of counseling and socialization are not optimal.

One of the solutions that can be done in overcoming the problem here is to use data *mining techniques*. The data *mining* used is *Clustering* using the *K-Means Clustering algorithm*. The result of the application of *the Clustering algorithm* is in the form of a way to determine the group of diseases that occur in the communityso that from *the Clustering* the health center can determine effective and efficient counseling in order to preventthe cause of a disease.

The origin of the word "algorithm" comes from "*algoritsm*", the Latin form of alKhawarizmi, a Persian mathematician, scientist, astronomer, and geographer. The algorithm has several characteristics and parts that have been observed. With the existence of programming algorithms, things that need to be solved can be easily solved with the help of computers [6]

The *K-Means Clustering* algorithm has previously been carried out in grouping medical record data based on the type of disease at the PT. Inecda using *the K-Means Clustering* method. The data used is medical record data of 600 patients, so that data classification is obtained based on region, type of disease, and age. Identifying medical record data from Anwar Medika Hospital as many as 534 patient data with a completion time of 0.06 seconds by the system and also analyzing medical record data to be grouped into 4[2] *clusters* with 4 variables, namely sub-district variables, disease diagnosis, age and gender. Patient Disease Grouping Using[7] *the K-Means* Algorithm, the data used consisted of 3875 records and 5 attributes, namely Gender, Type of Participant, Diagnosis, Discharge Status, Address. This research resulted in *clustering*, namely *cluster* 1 with 710 data while *cluster* 2 with a total of 3165 data. The results of the study show that the use of 2 *clusters* to bethe *best cluster* with *a Silhouette Coefficient* value shows a result with an SC value of 0.646. Patient complaintdata that is in the patient's medical record data. The results of this study are patient complaints in 3[8] *clusters*, where *cluster* 1 is complaints that occur with high intensity as many as 9 categories of complaints, *cluster* 2 is complaints that occur with moderate intensity as many as 15 categories of complaints, and *cluster*3 is complaints that occur with low intensity as many as 48 categories of complaints. [9]

2. METHODS

In this study, a framework is used which can be seen in Figure 1.

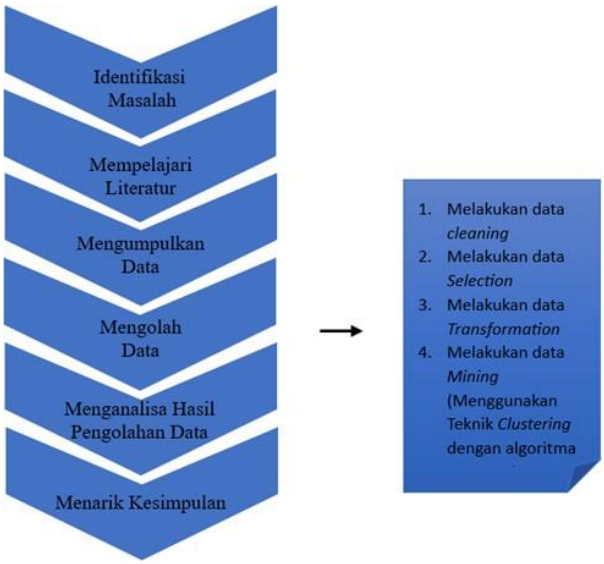

Figure 1. Research Framework

Based on the research framework in Figure 1, it can explain several frameworks or stages that will be carried out in this study.

1. Problem Identification

Problem identification is the stage where the researcher identifies the existing problem as a step to be able to understand the existing problem. There has been no classification of diseases so far. The application of data *mining* Clustering techniques using *the K-Means Clustering* algorithm helps in determining the groups of existing diseases.

2. Studying Literature

This stage will be searched and collected and studied a number of literature on theories and concepts that support problem solving in research. The literature used is in the form of reference books or supporting books, international and national journals and supporting concepts in completing this research. A literature study is carried out, namely by reading books that support the analysis of the data and information obtained. The literature studied is selected to be able to determine which literature will be used in the research.

3. Collecting Data

The data collection is carried out in various ways, namely:

a. The direct observation method is to make direct observations of medical record activities to obtain thedata needed.
b. The interview method is to hold interviews with parties directly related to the problem being researchedto obtain an overview and explanation.
c. The literature study method is a source that can be used as a reference from data sources or literature.
d. The browsing method is to collect references sourced from the internet.

4. Processing Data

The data required in the research has been collected, the next step is to process the data in several stages including:

a. Data *cleaning* is done to eliminate inconsistent or irrelevant data.
b. Data selection, is carried out to take appropriate data for analysis. In this study, the attributes

used were inthe form of disease data per period (per month).

c. Data *transformation*, so that data can be processed using *the K-Means algorithm.*

d. Data *Mining*, after the data is transformed, the data can be processed using *the K-Means Clustering*

algorithm.

5. Analyze data processing results

This stage is analyzed on the results of data processing in the form of disease *clusters* that have beenprocessed using *the K-Means Clustering* algorithm.

6. Drawing conclusions and recommendations

The last step after all the processes are completed, the researcher can draw a conclusion to the problembeing studied.

3. RESULTS AND DISCUSSION

In the process, data *mining* will extract valuable information by analyzing the existence of certain patterns or relationships from the data. There are also mining data processes to produce information including:

1. Data *Selection*

The data used in this study comes from disease data totaling 268 categories for 5 months starting fromJanuary-May 2024. The information used in this study is available in table 1 below:

Table 1. Data *Selection*

No.	Disease	Moon				
		January	February	March	April	May
1	*Abnormal uterine and vaginal bleeding*	1	1	2	1	3
2	*Abscess*	1	1	0	2	1
3	*Acute bronchiolitis*	1	3	3	2	1
4	*Acute gingivitis*	1	1	1	2	1
5	*Acute laryngopharyngitis*	1	3	3	1	2
6	*Acute lymphadenitis*	1	3	3	5	5
7	*Acute nasopharyngitis*	22	12	12	17	17
8	*Acute pharyngitis*	2	4	3	1	1
9	*Acute serous otitis media*	1	3	1	2	1
10	*Acute sinusitis*	1	5	12	3	1
11	*Acute suppurative otitis media*	1	2	2	1	1
12	*Acute upper respiratory infection*	17	1	1	4	15
13	*Allergic purpura*	1	2	1	1	1
14	*Allergic rhinitis*	2	1	1	1	1
15	*Amenorrhoea*	3	2	2	1	2
16	*Angina pectoris*	1	1	6	12	7
17	*Atherosclerotic cardiovascular disease*	1	2	1	2	1
18	*Atherosclerotic heart disease*	4	5	5	10	8
19	*Atopic dermatitis*	5	1	1	1	5
20	*Dst.*					
80	*Disorders of vestibular function*	1	3	2	3	1
81	*Disturbances in tooth eruption*	3	3	2	1	1
82	*Dizziness and giddiness*	1	2	1	2	1
83	*Dyspepsia*	7	20	21	12	12
84	*Embedded teeth*	2	6	3	4	4
85	*Epidemic myalgia*	1	3	1	2	2
86	*Epilepsy*	1	2	3	1	1
87	*Erosive (osteo)arthrosis*	1	2	3	1	1
88	*Essential (primary) hypertension*	21	34	35	42	51
89	*Female pelvic inflammatory disease*	1	3	1	3	3
90	*Dst.*					
253	*Tuberculosis of lung*	3	2	2	1	1
254	*Tympanosclerosis*	1	2	2	1	1
255	*Ulcer of lower limb*	1	1	2	2	1
256	*Unilateral or unspecified inguinal hernia*	1	1	1	3	3
257	*Unspecified chronic bronchitis*	2	2	2	3	1
258	*Unspecified injury of head*	1	1	1	2	1
259	*Unstable angina*	1	2	3	3	1
260	*Urinary tract infection*	1	2	3	2	2

No.	Disease	Moon				
		January	February	March	April	May
261	*Urticaria*	4	3	2	2	1
262	*Valgus deformity*	1	2	2	1	1
263	*Varicella [chickenpox]*	1	1	1	1	1
264	*Varus deformity*	3	3	3	2	1
265	*Vertigo of central origin*	5	3	3	3	1
266	*Viral conjunctivitis*	2	1	1	1	2
267	*Visual disturbance*	2	3	1	1	1
268	*Zoster [herpes zoster]*	1	3	2	2	2

2. Data *Transformation*

In the use of *the K-Means* grouping *method*, it is important to convert nominal data into *numerical* data,so that disease data can be processed using *K-Means algorithm* data and in accordance with the purpose of this study, this is because *K-Means* can only process *numerical data*, so data transformation is carried out. Thisstage of *transformation* is like consolidating all disease data per day into data per month. The results of the data *transformation* in Table 2 are as follows:

Tabel 2. Data *Transformation*

No.	Disease	Moon				
		January	February	March	April	May
1	*Abnormal uterine and vaginal bleeding*	1	1	2	1	3
2	*Abscess*	1	1	0	2	1
3	*Acute bronchiolitis*	1	3	3	2	1
4	*Acute gingivitis*	1	1	1	2	1
5	*Acute laryngopharyngitis*	1	3	3	1	2
6	*Acute lymphadenitis*	1	3	3	5	5
7	*Acute nasopharyngitis*	22	12	12	17	17
8	*Acute pharyngitis*	2	4	3	1	1
9	*Acute serous otitis media*	1	3	1	2	1
10	*Acute sinusitis*	1	5	12	3	1
11	*Acute suppurative otitis media*	1	2	2	1	1
12	*Acute upper respiratory infection*	17	1	1	4	15
13	*Allergic purpura*	1	2	1	1	1
14	*Allergic rhinitis*	2	1	1	1	1
15	*Amenorrhoea*	3	2	2	1	2
16	*Angina pectoris*	1	1	6	12	7
17	*Atherosclerotic cardiovascular disease*	1	2	1	2	1
18	*Atherosclerotic heart disease*	4	5	5	10	8
19	*Atopic dermatitis*	5	1	1	1	5
20	*Dst.*					
80	*Disorders of vestibular function*	1	3	2	3	1
81	*Disturbances in tooth eruption*	3	3	2	1	1
82	*Dizziness and giddiness*	1	2	1	2	1
83	*Dyspepsia*	7	20	21	12	12
84	*Embedded teeth*	2	6	3	4	4
85	*Epidemic myalgia*	1	3	1	2	2
86	*Epilepsy*	1	2	3	1	1
87	*Erosive (osteo)arthrosis*	1	2	3	1	1
88	*Essential (primary) hypertension*	21	34	35	42	51
89	*Female pelvic inflammatory disease*	1	3	1	3	3
90	*Dst.*					
253	*Tuberculosis of lung*	3	2	2	1	1
254	*Tympanosclerosis*	1	2	2	1	1
255	*Ulcer of lower limb*	1	1	2	2	1
256	*Unilateral or unspecified inguinal hernia*	1	1	1	3	3
257	*Unspecified chronic bronchitis*	2	2	2	3	1
258	*Unspecified injury of head*	1	1	1	2	1
259	*Unstable angina*	1	2	3	3	1
260	*Urinary tract infection*	1	2	3	2	2
261	*Urticaria*	4	3	2	2	1
262	*Valgus deformity*	1	2	2	1	1
263	*Varicella [chickenpox]*	1	1	1	1	1
264	*Varus deformity*	3	3	3	2	1
265	*Vertigo of central origin*	5	3	3	3	1
266	*Viral conjunctivitis*	2	1	1	1	2
267	*Visual disturbance*	2	3	1	1	1
268	*Zoster [herpes zoster]*	1	3	2	2	2

3. Data *Mining K-Means*

After converting all disease data in January-May 2024 into *a numerical* format, the next step to be taken is to apply the *K-Means Clustering method* to group the data. The stages of the *K-Means Clustering* algorithm

are to describe the steps of modeling *K-Means Clustering* from the initial step of the algorithm to be formed, namely determining the number of *clusters* to be formed, determining the initial *centroid*, and iterating until the cluster is generated. Here are the steps of the *K-Means* algorithm in detail in this study:

a. **Set the number of *clusters***

In determining the number of factors (number of *clusters*) to be clustered, researchers have grouped thedata into four clusters with diseases with high intensity (C1), moderate (C2), low (C3) and very low/rare (C4).

b. Determining the *initial* centroid

In this study, researchers set the initial center point of each *cluster* in a random way to determine it. The initial *centroid* data consists of C1, C2, C3 and C4, Table 3 shows the center point of the initial *cluster* obtained:

Table 3. Central point

Data to	*Centroid*	January	February	March	April	May
88	1	21	34	35	42	51
22	2	7	6	13	3	10
90	3	12	4	5	5	2
8	4	2	4	3	1	1

c. Calculating the Distance of the First Iteration

The first iteration process will produce a group of data in the first data processing. Researchers used theeucliedean formula to calculate the distance of the data to the *initial centroid*. Here are the calculation stages of the first iteration:

$$D_{(a,b)} = \sqrt{(xa - yb)^2 + (xa - yb)^2 + (xn - yn)^2}$$

Calculation of the distance between the first disease and the center of the first cluster. This calculation isalso carried out on the 2nd to 268th disease data. The results of the calculation of the distance between the disease and the four initial cluster centers . Here are the results of iteration 1 in table 4.

Table. 4 Iteration 1

Data to	C1	C2	C3	C4
1	81,012	15,330	12,490	3,873
2	82,547	17,664	12,845	4,472
3	80,529	15,067	11,662	1,732
4	82,128	16,941	12,490	3,873
5	80,418	14,595	11,916	1,732
6	76,616	13,191	11,619	5,831
7	52,868	22,517	24,125	32,512
8	80,411	14,629	11,000	0,000
9	81,345	16,462	12,166	2,646
10	76,099	10,909	13,266	9,327
11	81,817	16,062	12,288	2,449
12	70,718	17,176	14,832	21,048
13	82,225	16,763	12,570	3,000
14	82,383	16,703	11,916	3,606
15	80,740	14,866	10,488	2,646
16	71,875	14,142	14,318	13,266
17	81,731	16,673	12,288	3,162
18	70,021	11,269	11,225	11,790
19	79,360	14,213	9,950	6,164
20	81,578	15,716	11,402	2,236
80	80,443	15,716	11,662	2,646
81	80,963	15,199	10,392	1,732
82	81,731	16,673	12,288	3,162
83	54,854	18,574	26,192	29,103

Data to	C1	C2	C3	C4
84	76,302	12,728	10,630	4,690
85	80,734	15,937	12,124	2,828
86	81,419	15,395	12,083	2,236
87	81,419	15,395	12,083	2,236
88	0,000	68,308	75,173	80,411
89	79,637	15,427	11,958	3,742
90	75,173	12,689	0,000	11,000
253	81,351	15,427	10,536	2,449
254	81,817	16,062	12,288	2,449
255	81,719	16,248	12,207	3,464
256	80,436	15,937	12,288	4,690
257	80,592	15,588	10,863	3,000
258	82,128	16,941	12,490	3,873

Table 4. Advanced

Data to	C1	C2	C3	C4
259	80,430	15,264	11,576	3,000
260	80,306	14,731	11,747	2,646
261	80,243	14,866	9,165	2,646
261	80,243	14,866	9,165	2,646
262	81,817	16,062	12,288	2,449
263	82,620	17,029	12,767	3,742
264	80,056	14,387	9,798	1,732
265	79,133	13,928	7,681	3,742
266	81,780	16,186	11,874	3,742
267	81,603	16,217	11,576	2,236
268	80,318	15,199	11,832	2,236

In table 4. It contains the results of the calculation of the distance from each disease category to *the initial centroid*. In column C1 is the distance of the category to the initial centroid C1. Column C2 contains the results of the calculation of the distance of disease data to *the initial centroid C2*, *Column C3 contains theresults of the calculation of the distance of disease data to* the *initial centroid C3*, *Column C4 contains the results of the calculation of the distance of disease data to* the *initial centroid C4*.

d. First Iteration grouping results

In the calculation of the 1st data, there are 4 *centroid* values: CI=81.012, C2=15.330, C3=12.490, and C4=3.872. A data will be a member of a *centroid* (C1, C2, C3 and C4). Based on the minimum *centroid* of 3,872 which *is located in* centroid *4, the 1st data will be entered into* cluster 4. Here is table 5. Iteration 1 results

Table 5. Results of the First Iteration

Data to	C1	C2	C3	C4	*Cluster*
1	81,012	15,330	12,490	3,873	4
2	82,547	17,664	12,845	4,472	4
3	80,529	15,067	11,662	1,732	4
4	82,128	16,941	12,490	3,873	4
5	80,418	14,595	11,916	1,732	4
6	76,616	13,191	11,619	5,831	4
7	52,868	22,517	24,125	32,512	2
8	80,411	14,629	11,000	0,000	4
9	81,345	16,462	12,166	2,646	4
10	76,099	10,909	13,266	9,327	4
80	80,443	15,716	11,662	2,646	4
81	80,963	15,199	10,392	1,732	4
82	81,731	16,673	12,288	3,162	4
83	54,854	18,574	26,192	29,103	2
84	76,302	12,728	10,630	4,690	4
85	80,734	15,937	12,124	2,828	4
86	81,419	15,395	12,083	2,236	4
87	81,419	15,395	12,083	2,236	4

Data to	C1	C2	C3	C4	*Cluster*
88	0,000	68,308	75,173	80,411	1
89	79,637	15,427	11,958	3,742	4
90	75,173	12,689	0,000	11,000	3
253	81,351	15,427	10,536	2,449	4
254	81,817	16,062	12,288	2,449	4
255	81,719	16,248	12,207	3,464	4
256	80,436	15,937	12,288	4,690	4
257	80,592	15,588	10,863	3,000	4
258	82,128	16,941	12,490	3,873	4
259	80,430	15,264	11,576	3,000	4
260	80,306	14,731	11,747	2,646	4
261	80,243	14,866	9,165	2,646	4
262	81,817	16,062	12,288	2,449	4
263	82,620	17,029	12,767	3,742	4
264	80,056	14,387	9,798	1,732	4
265	79,133	13,928	7,681	3,742	4
266	81,780	16,186	11,874	3,742	4
267	81,603	16,217	11,576	2,236	4
268	80,318	15,199	11,832	2,236	4

Perform the process of iteration 2 and so on until there is no more shift in the centroid value that is being calculated with *the previous centroid*. Defining a new *centroid* for iterations 2 onwards is based on *the* values of the previous cluster, i.e. by:

C1 iteration 2 = (Sum of January values on C1 + Sum of February values on C1 + Sum of values for March on C1 + Sum of April values on C1 + Sum of May values on C1)

C2 iteration 2 = (Sum of January values on C2 + Sum of February values on C2 + Sum of March values on C2 + Sum of April values on C2 + Sum of May values on C2)

C3 iteration 2 = (Sum of January values on C3 + Sum of February values on C3 + Sum of values for March on C3 + Sum of April values on C3 + Sum of May values on C3)

C4 iteration 2 = (Sum of January values on C4 + Sum of February values on C4 + Sum of March values on C4 + Sum of April values on C4 + Sum of May values on C4

4. *Interpretation/Evaluation*

Researchers utilize *the RapidMiner* application to achieve research goals. The results obtained are as follows:

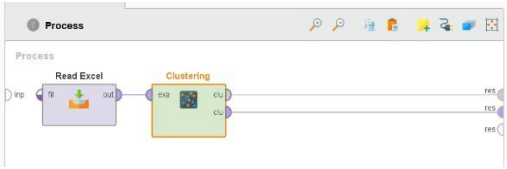

Figure 2. K-Means Clustering usage model

cluster	*avg. whitin centroid distance*
2	28.776
3	16.684
4	10.808
5	9.631
6	8.624
7	7.882

Figure 3. *Cluster* Testing with *elbow method*

In figure 3 above, it is clear that *cluster* 4 is the *cluster* that has the clearest angle, therefore the *cluster* to be selected is 4 *clusters.*

The results of the K-Means Data *Mining Clustering* process can be seen in the image below:

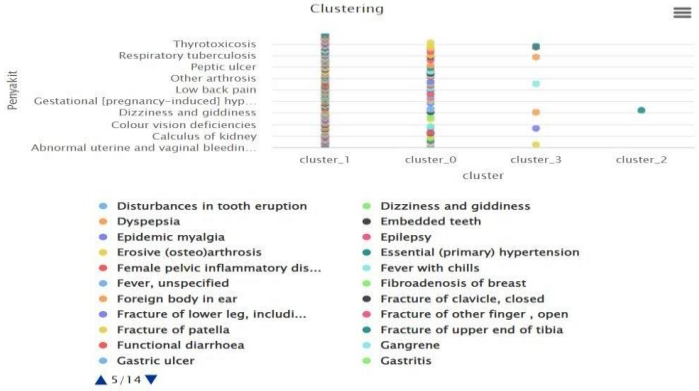

Figure 4. Visualization graph of the results of the distribution of *disease* clusters

The results of rapidminer calculations to determine *the cluster* of each disease at the PagembaranHealth Center. The detailed results can be seen in the table below:

Table 6. Hasil *K-Means Clustering*

No.	Disease	Moon					Result
		January	February	March	April	May	
1	*Abnormal uterine and vaginal bleeding, unspecified*	1	1	2	1	3	C4
2	*Abscess*	1	1	0	2	1	C4
3	*Acute bronchiolitis*	1	3	3	2	1	C4
4	*Acute gingivitis*	1	1	1	2	1	C4
5	*Acute laryngopharyngitis*	1	3	3	1	2	C4
6	*Acute lymphadenitis*	1	3	3	5	5	C3
7	*Acute nasopharyngitis*	22	12	12	17	17	C2
8	*Acute pharyngitis*	2	4	3	1	1	C4
9	*Acute serous otitis media*	1	3	1	2	1	C4
10	*Acute sinusitis*	1	5	12	3	1	C3
11	*Acute suppurative otitis media*	1	2	2	1	1	C4
12	*Acute upper respiratory infection*	17	1	1	4	15	C3
13	*Allergic purpura*	1	2	1	1	1	C4
14	*Allergic rhinitis*	2	1	1	1	1	C4
15	*Amenorrhoea*	3	2	2	1	2	C4
16	*Angina pectoris*	1	1	6	12	7	C3
17	*Atherosclerotic cardiovascular disease*	1	2	1	2	1	C4
18	*Atherosclerotic heart disease*	4	5	5	10	8	C3
19	*Atopic dermatitis*	5	1	1	1	5	C4
20	*Dst.*						
80	*Disorders of vestibular function*	1	3	2	3	1	C4
81	*Disturbances in tooth eruption*	3	3	2	1	1	C4
82	*Dizziness and giddiness*	1	2	1	2	1	C4
83	*Dyspepsia*	7	20	21	12	12	C2
84	*Embedded teeth*	2	6	3	4	4	C3
85	*Epidemic myalgia*	1	3	1	2	2	C4
86	*Epilepsy*	1	2	3	1	1	C4
87	*Erosive (osteo)arthrosis*	1	2	3	1	1	C4
88	*Essential (primary) hypertension*	21	34	35	42	51	C1
89	*Female pelvic inflammatory disease*	1	3	1	3	3	C4
90	*Dst.*						
253	*Tuberculosis of lung*	3	2	2	1	1	C4
254	*Tympanosclerosis*	1	2	2	1	1	C4
255	*Ulcer of lower limb*	1	1	2	2	1	C4
256	*Unilateral or unspecified inguinal hernia*	1	1	1	3	3	C4
257	*Unspecified chronic bronchitis*	2	2	2	3	1	C4
258	*Unspecified injury of head*	1	1	1	2	1	C4
259	*Unstable angina*	1	2	3	3	1	C4
260	*Urinary tract infection*	1	2	3	2	2	C4
261	*Urticaria*	4	3	2	2	1	C4
262	*Valgus deformity*	1	2	2	1	1	C4
263	*Varicella [chickenpox]*	1	1	1	1	1	C4
264	*Varus deformity*	3	3	3	2	1	C4
265	*Vertigo of central origin*	5	3	3	3	1	C4
266	*Viral conjunctivitis*	2	1	1	1	2	C4
267	*Visual disturbance*	2	3	1	1	1	C4
268	*Zoster [herpes zoster]*	1	3	2	2	2	C4

Based on the cluster results from Table 6 regarding disease data, researchers can conclude that, C1 (*Cluster* 1) is a disease that often occurs, with high intensity so that the health center pays special attention to the disease, namely by providing outdoor counseling and indoor counseling. C2 (*Cluster* 2) is a disease that occurs with moderate intensity so that the action that can be taken by the health center is by means of indoor counseling, C3 (*Cluster* 3) is a disease that occurs with low intensity so that the health center only gives directions or warnings when the patient is doing treatment, C4 (*Cluster* 4) is categorized as a disease with verylow intensity or rare, so the health center conducts periodic public health monitoring to detect the potential emergence of diseases and provide general information and education about disease prevention to the public through print or digital media.

DISCUSSION

In the test that has been carried out above, the researcher used 268 disease data using 4 clusters. Before determining the number of clusters, each cluster is tested and determined using the elbow method and 4 clusters are obtained that are the most appropriate to use. The results of processing using *the rapidminer* application produce as follows:

Cluster 1 : 1 disease

Cluster 2 : 7 diseases

Cluster 3 : 45 diseases

Cluster 4 : 215 diseases

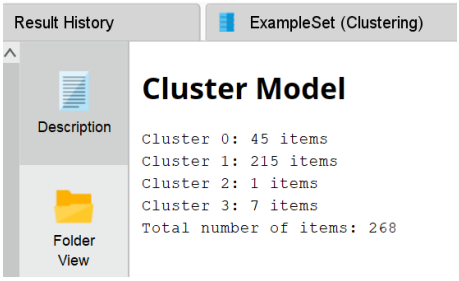

Figure 5. *Cluster Model*

From the results of the data mining process that has been carried out, the results are obtained, namely, 1 disease which is a disease with high intensity (C1), namely *Essential (primary) hypertension*, 7 diseases with medium intensity (C2), namely *Acute nasopharyngitis, Chronic kidney disease, Dyspepsia, Non-insulin- dependent diabetes mellitus, Pulpitis, Supervision of high-risk pregnancy, Supervision of other normalpregnancy*. 45 diseases with low intensity (C3) are *Acute lymphadenitis, Acute sinusitis, Acute upper respiratory infection, Angina pectoris, Atherosclerotic heart disease, Bell's palsy*. 215 diseases with very low intensity/rarely occurring (C4) namely *Abnormal uterine and vaginal bleeding, Abscess, Acute bronchiolitis, Acute gingivitis, Acute laryngopharyngitis, Acute pharyngitis.*

4. CONCLUSIONS

This study resulted in 4 *disease clusters*. *Cluster* 1 obtained 1 disease data, with a disease incidence percentage of 68.28% in January-May 2024. *Cluster* 2 obtained 7 disease data, with a disease incidence percentage of 20.52%. *Cluster* 3 obtained 45 disease data, with a disease incidence percentage of 7.96%. Meanwhile, *Cluster* 4 obtained 215 disease data with a disease incidence percentage of 3.24% in January-May 2024 from 268 processed sample data. The calculation of *K-Means Clustering* can help the health center in preventing and providing community services effectively because this study has produced models and considerations to determine the right treatment or countermeasures for disease symptoms in the community.

REFERENCES

[1] Nitra Eka Safitri, "Implementation of the Regulation of the Minister of Health of the Republic of Indonesia Number 75 of 2014 concerning Community Health Centers at the Sukamakmur Labuhan Batu Health Center," 2019.

[2] R. Ordila, R. Wahyuni, Y. Irawan, and M. Yulia Sari, "Application of Data Mining for Clustering of Patient Medical Record Data Based on Disease Type with Clustering Algorithm (Case Study: Pt.Inecda Polyclinic)," *Journal of Computer Science*, Vol. 9 No. 2 HLM. 148–153, Oct 2020, doi:10.33060/zik/2020/vol9.ISS2.181.

[3] Murnawan Dan U. Nugraha, "Classify Event Participants In Universities And Industries Using Knowledge Discovery In Databases," *Review Of International Geographical Education (Rigeo)*, Vol.11, No. 1, Hlm. 526–542, 2021, doi: 10.48047/Rigeo.11.1.36.

[4] S. Manochandar, M. Punniyamoorthy, Dan R. K. Jeyachitra, "Development Of New Seed With Modified Validity Measures For K-Means Clustering," *Comput Into Eng*, vol. 141, Hlm. 106290, Mar2020, Doi: 10.1016/J.Cie.2020.106290.

[5] M. Ihksan, H. Susilo, N. Abdillah, and S. S. Saintika, "Application of K-Means Data Mining Clusteringof Drug Needs at Medika Saintika Clinic," *Medical Health Journal of Science June 2023 |Vol 14 Number*, Vol. 14, No. 1, Hlm. 394, 2023, Doi: 10.30633/Jkms.V14i1.2581.

[6] Nurhaliza Khesya, "Getting to Know Flowcharts and Pseudocode in Algorithms and Programming," 2021.

[7] A. Ali, "Clustering of Patient Medical Record Data Using the K-Means Clustering Method at Anwar Medika Balong Bendo Sidoarjo Hospital," *Matrix : Journal of Management, Informatics Engineering and Computer Engineering*, Vol. 19, No. 1, Hlm 186–195, Nov. 2019, Doi: 10.30812/Matrik.V19i1.529.

[8] R. Anggraini, E. Haerani, J. Jasril, and I. Afrianty, "Patient Disease Classification Using K-Means Algorithm," *Jurikom (Journal of Computer Research)*, Vol. 9, No. 6, HLM. 1840, DES 2022, doi: 10.30865/Jurikom.V9i6.5145.

[9] M. A. V. Ideal, "Classification Of Patient Complaints Against Patient Medical Record Data Using TheK Means Method," *Journal of Information and Technology Systems*, Hlm. 1–6, Agu 2022, doi: 10.37034/jsisfotek.V5i1.151.