Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

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

Classification Of Diseases In Patients Based On Factors Environment Using The K-Means Algorithm At Puskesmas Subdistrict Selesai

Sisca Ramadhani [1] *, Marto Sihombing [2], Magdalena Simanjuntak [3]

[1,2,3] Information Systems, STMIK Kaputama
Jl. Veterans No. 4A-9A, Binjai, North Sumatra, Indonesia
E-mail: siscaramadhani3@gmail.com [1] , martosihombing45@gmail.com [2], magdalena.simanjuntak84@gmail.com [3]*

Abstract

Diseases caused by the environment are disease phenomena caused by the relationship between humans and environmental factors. Diseases that occur due to the environment that must be known by the public are such as ISPA, dermatitis, diarrhea, pulmonary TB, and so on. In the area of the Kecamatan Selesai, there are still many environmental conditions Not yet such as damaged roads and smoke from factories that cause air pollution, so with condition environment like This can affect public health. Puskesmas Selesai is Public health center Which located in region Kecamatan Selesai. The data of patients seeking treatment at this puskesmas are only used archives and to view the patient's medical history. The public should know aboutsymptoms of the disease in order to get appropriate services. In data mining techniques for clustering patient disease data can be used as new information useful for puskesmas or related as material counseling to society. The purpose of this study is to analyze the results of the application of data mining using K-Means Clustering in grouping patient diseases based on the environment with age, village and disease diagnoses variables.

***Keywords**: Data Mining, Patient, K-Means*

1. Introduction

Health become one factor most important besides education and income. Every person own right basic which the same for get good health services. Health conditions can be affected by several factor among them is environment and service health [1]. In the area of the Kecamatan Selesai, there are still many environmental conditions not yet such as damaged roads and smoke from factories that cause air pollution, so with condition environment like this can affect public health. At Puskesmas Selesai, the patient's data for treatment is only used as an archive and to view the patient's medical history. In data mining, patient data clustering techniques can be used as new information that is useful for the health center to carry out counseling to the community. The application of disease grouping based on the environment is in research on diseases related to air pollution or pollution other environment. This grouping is useful in helping identify patterns or correlations between disease and factors specific environment with the K-Means algorithm.

Data record medical can utilized for increase satisfaction patient to hospital. By using data mining method, data record medical which on at first only form data stack usually can made solution for look for information which contained in record medical activity medical. Lots of information can be taken from the data in hospital. Like information which can we know about deep pattern various kind of disease [2].

2. Research Methods

2.1. Data Mining

Data mining is the process of looking for patterns or interesting information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The selection of the right method or algorithm is highly dependent on the objectives and process of Knowledge Discovery in Database (KDD) as a whole. Knowledge Discovery in Database (KDD) is a method for obtaining knowledge from existing databases. In the database there are tables - tables that are interconnected / related. The results of the knowledge obtained in this process can be used as a knowledge base for decision making purposes [3]. In data mining there are steps that must be carried out to carry out the process of extracting data which can be seen in the following figure.

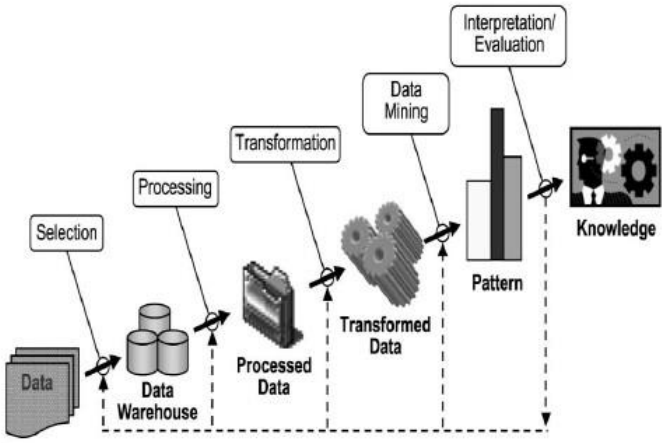

Figure 1: Stages of Data Mining

2.2. Clustering Method

Clustering is a process of grouping a number of data or objects into clusters (groups) so that each in the same cluster will contain data that is as similar as possible and different from objects in other clusters. An object/data that is grouped into a group will have the same characteristics based on certain criteria. One of the activities carried out in analyzing the data is the classification or grouping of data into several categories, groups or clusters. Data grouping of computer parts and accessories uses the k-means clustering method. The data obtained from this method will be grouped into several clusters based on consumer buying interest. Data will be grouped in one cluster if it has the same characteristics [4].

2.3. K-Means Algorithm

According to (Eko, 2012) said that the K-Means method partitions the data into groups so that data with the same characteristics is included in the same group and data with different characteristics are grouped into other groups. Data that has a representative value similarity in one group and data that has a difference in another group so as to allow the grouping of different data that has a small level of variation. The main principle of this technique is to arrange K partitions/centroids / averages from a set of data. The purpose of grouping this data is to minimize the objective function that is set in the grouping process, which generally tries to minimize variations within a group and maximize variations between groups.

The steps in the clustering process using the K-means algorithm are as follows:

1. Determine the value of K as the number of clusters to be formed.
2. Initialization of K cluster centers can be done in various ways, but the most often done is by random method taken from existing data.
3. Calculating the distance of each input data to each centroid using the Euclidean Distance formula until the closest distance is found for each data with an example. The following is the Euclidean Distance equation :

$$De = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2}$$

   Description: De is the Euclidian Distance, I is the number of objects, (x,y) is the object coordinates and (s,t) is the centroid coordinate.
4. Classify each data based on its proximity to the centroid.
5. Updating the centroid value, the new value is obtained from the average cluster concerned with the formula:

$$v_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj}$$

   Description: Vij is the ith cluster centroid for the j variable, Ni is the number of data members of the ith cluster , i,k is the cluster index , j is the index of the variable, xkj is the k data value in in the cluster for the j variable [5].

2.4. Disease Grouping

Grouping is a process for grouping objects, people, or concepts based on similarities and certain criteria. The purpose of grouping is to organize and structure these entities into more orderly and meaningful categories.

According to the World Health Organization (WHO), disease is an abnormal condition in a person's body or mind that causes disturbances in bodily or social functions. Disease can be caused by various factors, such as infection, heredity, environment, or an unhealthy lifestyle.

2.5. Patient

According to Permenkes Nomor 4 Tahun 2018 pasal 1, a patient is anyone who consults about their health problems to obtain the necessary health services, either directly or indirectly at the hospital. Patients have rights such as obtaining quality, humane, fair, honest and non-discriminatory health services in accordance with professional standards and standard operating procedures and so on [6].

2.6. Puskesmas

According to Permenkes number 43 of 2019 in article 1 Community Health Centers, hereinafter referred to as Puskesmas, are health service facilities that carry out public health efforts and first-level individual health efforts, with more priority on promotive and preventive efforts in their working areas. In article 2 paragraph 1, the health development held at the Puskesmas aims to create a healthy working area for the Puskesmas, with people who:

a. have healthy behavior which includes awareness, will, and ability to live a healthy life;
b. able to reach quality health services;
c. living in a healthy environment; and
d. have an optimal degree of health, both individuals, families, groups, and society [7].

2.7. Environmental Based Diseases

According to (Sang, 2016) disease based on the environment is a pathological condition in the form of abnormal function or morphology of an organ caused by human interaction with everything around it that has the potential for disease. Diseases based on the environment are still a problem today. ISPA and diarrhea which are environmental-based diseases are always included in the top 10 diseases in almost all health centers in Indonesia.

2.7.1. Factors Affecting Disease Based on the Environment

Factors that support the emergence of disease based on the environment include:

1. Availability and access to safe water
   Data from the Bappenas stated that in 2009 the proportion of the population with access to safe drinking water was 47.63 % . Sources of drinking water that are considered suitable include tap water, public taps, drilled wells or pumps, protected wells, protected springs and rainwater. The health impacts of not meeting basic needs for clean water and sanitation include children as a vulnerable age group. WHO estimates that in 2005, as many as 1.6 million children under five (an average of 4500 each year) died due to unsafe water and lack of hygiene.
2. Access to proper basic sanitation
   Ownership and use of toilet facilities is an important issue in determining the quality of sanitation. However, in reality, the 2009 Susenas data shows that almost 49% of Indonesian people do not have access to a latrine. This means that there are more than 100 million Indonesians who defecate openly and use latrines that are not of good quality. This figure is clearly a major factor resulting in the high incidence of diarrhea, especially in infants and toddlers in Indonesia.
3. Handling of garbage and waste
   Unorganized waste management will cause many disturbances both in terms of aesthetics in the form of piles and scattered garbage, air, soil and water pollution, potential release of methane gas (CH4) which contributes to global warming, siltation of rivers which leads to flooding and disturbances health such as diarrhea, cholera, typhus, skin disease, intestinal worms, or poisoning due to consuming food (meat/fish/plants) contaminated with toxic substances from waste.
4. Disease vector
   Disease vectors are increasingly difficult to eradicate, this is because disease vectors have adapted in such a way to environmental conditions, so that their ability to survive is even higher. This is supported by other factors that make vector breeding more rapid, including: changes in the physical environment such as mining, industry and housing development, piped clean water supply systems that have not yet reached the entire population so that containers are still needed for water supply, residential and urban drainage systems that are still do not meet the requirements, the waste management system does not meet the requirements, the use of pesticides that are not wise in vector control, global warming which increases the air humidity by more than 60% and is an ideal living condition and place for the spread of disease vectors.
5. Community Behavior
   Clean and healthy behavior has not been widely implemented by the community, according to a Basic Human Services (BHS) study in Indonesia in 2006, community behavior in washing hands 7 is (1) after defecating 12%, (2) after cleaning the feces of infants and toddlers 9 %, (3) before eating 14%, (4) before feeding the baby 7%, and (5) before preparing food 6%. Another BHS study on the behavior of household drinking water management showed that 99.20 % boiled water to get drinking water, but 47.50% of this water still contained Escherichia Coli . According to the 2006 Indonesia Sanitation Sector Development Program (ISSDP) study , 47% of people still defecate in rivers, rice fields, ponds, gardens and open areas [8].

3. Analysis And Design

3.1. Calculation Process

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 centroid , calculate the distance of the object to the centroid and then group it based on the closest distance, if no objects are moved or grouped then the iteration is finished. To determine the group of an object, the first thing to do is measure the Euclidean distance between two object points (X, Y and Z) which is defined as follows:

Table 1: Transformed Data

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

Then form a cluster into 3 groups (K=3) and determine the centroid center point . The clustering calculation process is as follows :
K=3 Centroids
C_1 = (2 ,6,4) taken from data A
C_2 = (3 ,5,4) taken from data B
C_3 = (4 ,2,7) is taken from the F data

Then proceed with the calculation process in Iteration 1:

1. A(2,6,4)
   C_1 =(2,6,4) = $\sqrt{(2-2)^2+(6-6)^2+(4-4)^2}$ =0
   C2 =(3,5,4) = $\sqrt{(2-3)^2+(6-5)^2+(4-4)^2}$ = 1,4
   C3 =(4,2,7) = $\sqrt{(2-4)^2+(2-2)^2+(4-7)^2}$ =5,4
2. B(3,5,4)
   C1 =(2,6,4) = $\sqrt{(3-2)^2+(5-6)^2+(4-4)^2}$ =1,4
   C2 =(3,5,4) = $\sqrt{(3-3)^2+(5-5)^2+(4-4)^2}$ = 0
   C3 =(4,2,7) = $\sqrt{(3-4)^2+(5-2)^2+(4-7)^2}$ =4,4
3. C(2,1,4)
   C1 =(2,6,4) = $\sqrt{(2-2)^2+(1-6)^2+(4-4)^2}$ =5
   C2 =(3,5,4) = $\sqrt{(2-3)^2+(1-5)^2+(4-4)^2}$ = 4,1
   C3 =(4,2,7) = $\sqrt{(2-4)^2+(1-2)^2+(4-7)^2}$ =3,7
4. And so on up to number 19
20. T(4,3,4)
   C1 =(2,6,4) = $\sqrt{(4-2)^2+(3-6)^2+(4-4)^2}$ =3,6
   C2 =(3,5,4) = $\sqrt{(4-3)^2+(3-5)^2+(4-4)^2}$ = 2,2
   C3 =(4,2,7) = $\sqrt{(4-4)^2+(3-2)^2+(4-7)^2}$ =3,2
   From the calculation above, the results of iteration 1 calculations are obtained, namely in the table below:

Table 2: Data from Iteration 1

Object	X	Y	Z	C1	C2	C3	Group
A	2	6	4	0.0	1,4	5,4	1
B	3	5	4	1,4	0.0	4,4	2
C	2	1	4	5.0	4,1	3,7	3

D	2	6	4	0.0	1,4	5,4	1
E	2	1	4	5.0	4,1	3,7	3
F	4	2	7	5,4	4,4	0.0	3
G	3	3	4	3,2	2.0	3,3	2
H	4	1	4	5,4	4,1	3,2	3
I	5	4	6	4,1	3.0	2,4	3
J	2	2	4	4.0	3,2	3,6	2
K	2	6	4	0.0	1,4	5,4	1
L	4	1	4	5,4	4,1	3,2	3
M	4	1	3	5,5	4,2	4,1	3
N	4	1	4	5,4	4,1	3,2	2
O	5	3	4	4,2	2,8	3,3	2
P	4	1	6	5,7	4,6	1,4	3
Q	1	6	4	1.0	2,2	5,8	1
R	4	4	4	2,8	1,4	3,6	2
S	2	1	7	5,8	5,1	2,2	3
Q	4	3	4	3,6	2,2	3,2	2

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

Old Groups: (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

New Groups : (1 2 3 1 3 3 2 3 3 2 1 3 3 2 2 3 1 2 3 2)

Because there was a change in the Group, the calculations were carried out in Iteration 2 which obtained the results in the following table :

Table 3: Data from Iteration 2

Object	X	Y	Z	C1	C2	C3	Group
A	2	6	4	0.3	3,4	4,9	1
B	3	5	4	1,6	2,1	3,7	1
C	2	1	4	5.0	2,5	1,8	3
D	2	6	4	0.3	3,4	4,9	1
E	2	1	4	5.0	2,5	1,8	3
F	4	2	7	5,5	3,2	2,1	3
G	3	3	4	3,3	0.6	1,9	2
H	4	1	4	5,5	2.0	1,2	3
I	5	4	6	4,3	2,7	3,2	2
J	2	2	4	4.0	1,9	1,8	3
K	2	6	4	0.3	3,4	4,9	1
L	4	1	4	5,5	2.0	1,2	3
M	4	1	3	5,6	2,3	2,1	3
N	4	1	4	5,5	2.0	1,2	3
O	5	3	4	4,4	1,4	2,4	2
P	4	1	6	5,8	2,9	1,2	3
Q	1	6	4	0.8	4.0	5,3	1
R	4	4	4	3.0	1,1	2,8	2
S	2	1	7	5,8	3,9	2,5	3
Q	4	3	4	3,8	0.4	1,9	2

After calculating using the cluster formula in iteration 2, the groups based on the minimum distance to the nearest centroid are:

Old Groups: (1 2 3 1 3 3 2 3 3 2 1 3 3 2 2 3 1 2 3 2)

New Groups : (1 1 3 1 3 3 2 3 2 3 1 3 3 3 2 3 1 2 3 2)

Because there was a change in the Group, the calculations were carried out in Iteration 3 which obtained the results in the following table :

Table 4: Data from Iteration 3

Object	X	Y	Z	C1	C2	C3	Group
A	2	6	4	0.3	3,4	4,9	1
B	3	5	4	1,6	2,1	3,7	1
C	2	1	4	5.0	2,5	1,8	3
D	2	6	4	0.3	3,4	4,9	1
E	2	1	4	5.0	2,5	1,8	3
F	4	2	7	5,5	3,2	2,1	3
G	3	3	4	3,3	0.6	1,9	2
H	4	1	4	5,5	2.0	1,2	3
I	5	4	6	4,3	2,7	3,2	2
J	2	2	4	4.0	1,9	1,8	3

K	2	6	4	0.3	3,4	4,9	1
L	4	1	4	5,5	2.0	1,2	3
M	4	1	3	5,6	2,3	2,1	3
N	4	1	4	5,5	2.0	1,2	3
O	5	3	4	4,4	1,4	2,4	2
P	4	1	6	5,8	2,9	1,2	3
Q	1	6	4	0.8	4.0	5,3	1
R	4	4	4	3.0	1,1	2,8	2
S	2	1	7	5,8	3,9	2,5	3
Q	4	3	4	3,8	0.4	1,9	2

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

Old Groups: (1 1 3 1 3 3 2 3 2 3 1 3 3 3 2 3 1 2 3 2)

New Groups : (1 1 3 1 3 3 2 3 2 3 1 3 3 3 2 3 1 2 3 2)

After calculating using the existing cluster formula, in iteration 3 it is the same as in iteration 2 and there is no data that moves groups again so the calculation can be stopped. So that a cluster graph can be made grouping of diseases in patients based on the environment using the K-Means Algorithm at the Puskesmas Kecamatan Selesai .

3.2. Clustering Graph

Based on the results of iteration calculations and data mining grouping of diseases in patients based on environmental factors at the Community Health Center, the graphical results are obtained as follows:

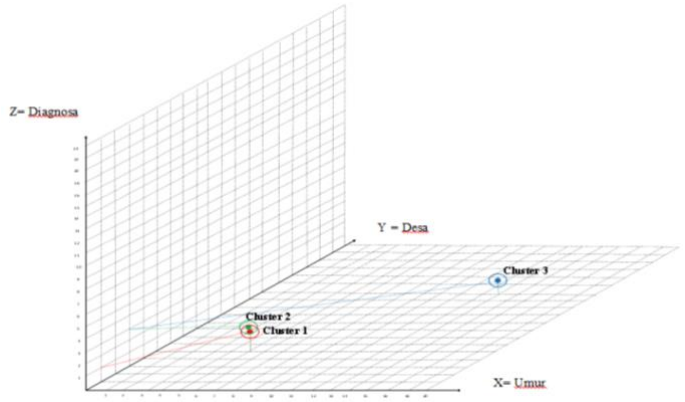

Figure 2: Cluster Graph

- Cluster 1 : 2, 5,4, 4 (2, 5, 4)
- Cluster 2 : 4,2, 3,4, 4,4 (4, 3, 4)
- Cluster 3 : 3,2, 1,2, 4,7 (3, 1, 5)

Of the 20 types of disease grouping data in patients based on the environment at the Puskesmas Kecamatan Selesai, 3 groups were obtained, cluster 1 contained 5 data, cluster 2 contained 5 data, and cluster 3 contained 10 data.

1. Cluster 1 There are 5 Data
   It can be seen in cluster 1 centered on 2, 5, 4, namely aged 5-11 years (children) in Mancang village with a diagnosis of dermatitis.
2. Cluster 2 There are 5 Data
   It can be seen in cluster 2 centered on 4, 3, 4, namely those aged 17-25 years (late adolescents) in the village of Kuta Parit with a diagnosis of dermatitis.
3. Cluster 3 There are 10 Data
   It can be seen that cluster 3 is centered on 3, 1, 5, namely those aged 12-16 years (early youth) in Bekulap village diagnosed with diabetes.

3.3 Interface Design

The description of the results is an overview of the results of the analysis that has been carried out. The results of this analysis will later be designed into an interface design so that it is easily understood by the user. Data mining interface design for grouping disease in patients based on the environment at the Puskesmas Kecamatan Selesai. Completed using the clustering method can be described in the menu structure. The following will explain clearly the description of the menu structure that will be made for grouping diseases in patients based on environmental factors at the Puskesmas Kecamatan Selesai.

1. home
   This page appears when you open the MATLAB application for the first time for data mining to be designed, on this page you will see the menus used in this application which can be seen in the picture as follows:

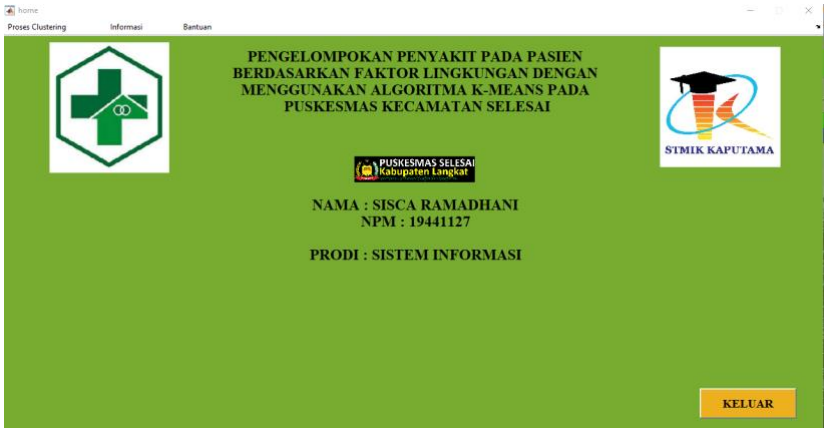

Figure 3: Menu Home

2. Cluster Process
   On this page you will see the entire data mining process up to the appearance of graphs and centroid information as a result of calculations using the clustering method . The cluster menu display can be seen as shown in the image below:

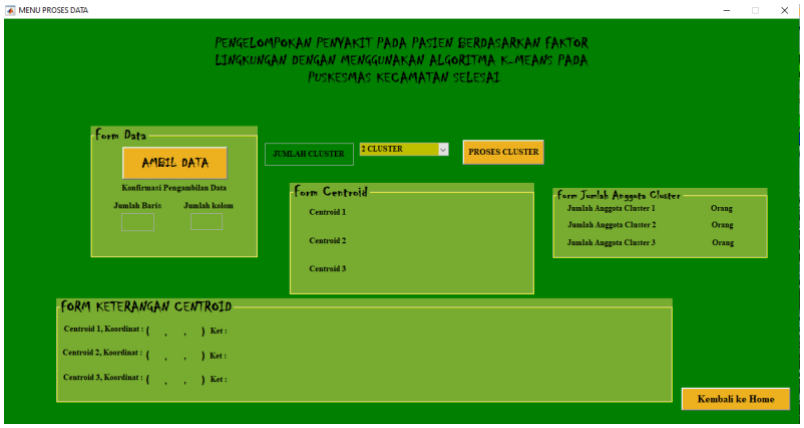

Figure 4: Process Clustering

3. Information
   On the Data Information page, patient disease grouping data will appear based on environmental factors at the Puskesmas Kecamatan Selesai according to categories stored in Microsoft Excel that has been connected to MATLAB, where the data undergoes a selection process based on disease, gender, and patient address on this menu page . The display of this menu is as shown below :

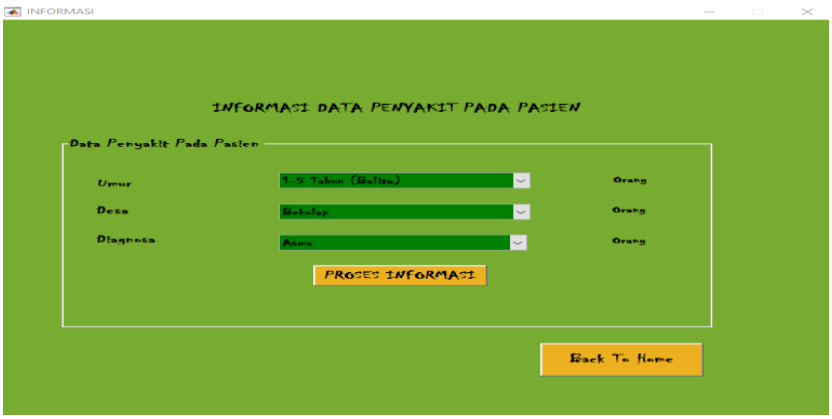

Figure 5: Information

4. Help
   The help page is used to display information on how to use the disease grouping system in patients based on environmental factors using the K-Means Algorithm at the Puskesmas Kecamatan Selesai . The display of the help menu is as follows :

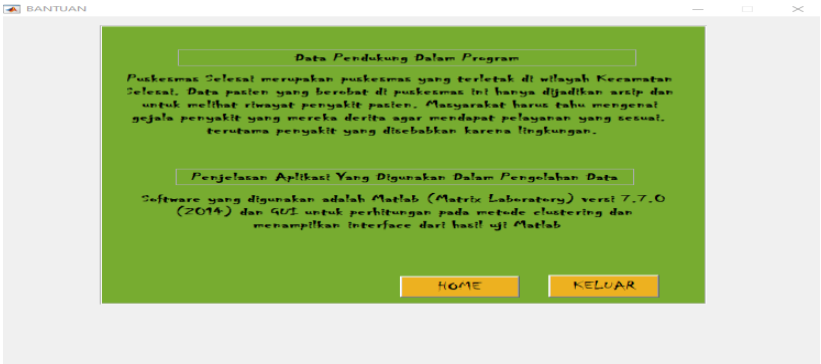

Figure 6: Help

4. Conclusion

From the results of the analysis of grouping patient disease data based on the environment, a conclusion can be drawn as follows:

1. From the tests carried out using the clustering method with the K-means algorithm, it can be seen that the patient's disease groups are based on the environment at the Puskesmas Selesai.
2. From 20 disease sample data, 3 groups were obtained , cluster 1 contained 5 data, cluster 2 contained 5 data, cluster 3 contained 10 data.
3. In cluster 1 centered on 2, 5, 4, namely those aged 5-11 years (children) in the village of Mancang with a diagnosis of dermatitis.
4. In cluster 2 centered on 4, 3, 4, namely those aged 17-25 years (late adolescents) in the village of Kuta Parit with a diagnosis of dermatitis.
5. In cluster 3 centered on 3, 1, 5, namely those aged 12-16 years (early youth) in Bekulap village diagnosed with diabetes.

References

[1] R. Anggraini, E. Haerani, and I. Afrianty, "Pengelompokkan Penyakit Pasien Menggunakan Algoritma K-Means," *Jurnal Riset Komputer*, vol. 9, pp. 1840–1849, Dec. 2022.
[2] T. Anjarsari, Megawaty, and A. Putra, "PENGELOMPOKAN PENYEBARAN PENYAKIT ISPA DI WILAYAH KOTA SEKAYU MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING (STUDI KASUS: RSUD SEKAYU)," *Bina Darma Conference on Computer Science 2019*, pp. 174–184, Jan. 2019.
[3] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik Informatika*, vol. 2, no. 2, pp. 213–219, 2017.
[4] S. Handoko, Fauziah, and E. T. E. Handayani, "IMPLEMENTASI DATA MINING UNTUK MENENTUKAN TINGKAT PENJUALAN PAKET DATA TELKOMSEL MENGGUNAKAN METODE K-MEANS CLUSTERING," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 25, no. 1, pp. 76–88, Apr. 2020.
[5] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *Jurnal Tekno Kompak*, vol. 15, no. 2, p. 25, Aug. 2021, doi: 10.33365/jtk.v15i2.1162.
[6] "Permenkes Nomor 4 Tahun 2018," Jakarta, Mar. 2018.
[7] "BERITA NEGARA REPUBLIK INDONESIA," 2019. [Online]. Available: www.peraturan.go.id
[8] S. G. Purnama, *Buku Ajar Penyakit Berdasarkan Lingkungan*. 2016.