Data Mining Using Clustering Methods to Identify The Type of Lung Disease (Case Study: Siti Hajar Hospital)

Rifqi Agung Priambodo¹, Opim S. Sitompu², Muhammad Zarlis²
¹,²,³ Departement Technology and Computer Science, University North Sumatera, Medan, Indonesia

Abstract. In the process of grouping types of lung disease, doctors and staff Siti Hajar hospital experience the various obstacles encountered, including difficulties in grouping the data types of lung disease. This deployment planning analysis will help the hospital in analyzing data classification type of lung disease. This method of making use of data collection by taking data at the hospital

1. Introduction
Data mining is frequently described as a way to decipher and search for invention in the form of knowledge in a database. Data mining is the process of selection or "mining" knowledge of a bunch of data in great numbers. [1]

Lung disease is a disease related to the respiratory system in humans, it can be bad if it does not immediately dealt with seriously. In the hospital sometimes patients who want to see a lot of that queue up and wait for the specialists came so waits too long to know what diseases suffered by these patients. The lungs as the pump only to respiratory system is an organ that is essential for the continuation of life.

1.1. Problem Identification
In order not to deviate from the objectives given the limitations issue as follows:
   a. There is still also a difficulty in logging on a symptom or diagnosis that occurs in patients of lung disease.
   b. absence of calculation data mining using clustering method, so the nurse or doctor is very difficult in determining diagnosis/symptoms of patients
   c. It is not yet known how to analyze this type of disease through the clustering process

1.2. Problem Limitation
In order not to deviate from the objectives given the limitations issue as follows:
   a. the authors discuss only logging patients in poly Hospital lung Siti Hajar terrain.
   b. Authors discuss only logging on types of diseases or symptoms in patients suffering lung disease.
   c. build analysis capable of informing the patient data and conduct the process of clustering of the data.
1.3. Research objective
As for the purpose of this research is:

- a. contribute to Hospital in order to be able to give the patient record information and perform data clustering process.
- b. create a system that can provide convenience for patients in search of a list of doctors and also ease create a doctor in search of the patient's history list.
- c. analyze data patients lung disease and shape it into the desired cluster.

2. Theoretical Basis
Information system is a system in an organization to meet the needs of daily transaction processing support functions of the nonprofit organization operating with managerial strategy activities of an organization to be able to provide to certain outside parties with the necessary reports [2].

2.1. Understanding and Stage of Data Mining
Data mining is a set of processes to explore the added value from a set of data in the form of knowledge that is not known. It is worth remembering that the word mining itself means a little effort to get valuable items from a large number of basic material. Therefore Data Mining actually has a long root of the fields such as artificial intelligence (artificial intelligent), machine learning, statistics and data base. Some of the techniques that are often cited in the literature Data Mining, among others: clustering, classification, association rule mining, neural network, dan genetic algorithm.[3] Data Mining has several stages, as for the stages is: [4]

![Figure 1. The Stage of Data Mining](image)

According to Sando Romarios (2013) to know more about the stages in the process of knowledge in databases, the following description of the stages according to the:

- **a. Database**
  A collection of interrelated data for use together then designed to meet the information needs of the Organization.

- **b. Data Cleaning**
  In General, the data obtained from the database of an enterprise as well as experiments, have stuffing imperfect field such as lost data, invalid data or just a typo. Data that is not relevant it also better thrown away because his presence could reduce the quality or accuracy of the results of data mining. Data cleanup will also affect the performasi of the system data mining because the data will be handled will be reduced in number and their complexes.

- **c. Data integration**
  Data integration is performed on the attribute table-attribute table that identifies the unique entities such as employee attribute table name, number, place of birth, religion and so on. Data integration needs to be done carefully because of an error in data integration could yield data that distorted and misleading even taking action later.
d. Task relevant data

After all of the data source are merged or integrated into one overall database, the next step is to perform a task relevant data. At this stage do the relevancy of the attribute table of the data that is relevant or appropriate to the target or output to be generated.

e. Data transformation

Coding is the process of transformation in the data that have been chosen, so the data sesuain to the process of data mining. The process of coding in KDD is a creative process and depends greatly on the type of information or patterns that will be searched in the database.

f. Data mining

Data mining is the process of exploring and analyzing a large amount of data which aims to find a pattern or an interesting information from data stored in large numbers by the use of a particular method or technique. Technique, method, or the right algorithm depends heavily on the purpose and overall KDD process. This stage is the core of the stages of KDD that are done to analyze data that has been cleaned.


g. Pattern evaluation

In this stage, the result of data mining techniques in the form of the distinctive patterns or model predictions were evaluated to assess whether the hypothesis is that there are indeed achieved. When it turns out that the results obtained are not in accordance with the hypothesis, there are some alternatives that can be taken as feedback to improve menjadikannnya data mining other more appropriate, or accept the results as an outcome that is beyond suspicion that might be useful.

h. Knowledge

The last stage of the data mining process is how to formulate decisions or actions from the analysis results obtained. There are times when this should involve people who don't understand data mining. Therefore the presentation of the results of data mining in the form of knowledge that can be understood by everyone is atu the stages required in the data mining process. In this presentation, visualization can also help communicate the results of data mining.

2.2. Clustering Method

Basically it is a clustering method for searching and categorizing data that has a semblance of characteristic (similarity) between one data with other data. Clustering is one of the methods of data mining is without direction (unsupervised), means that the method is applied in the absence of exercise (training) and without any of the teacher and does not require the target output. In data mining there are two types of clustering method used in grouping the data, is hierarchical clustering dan non-hierarchical clustering.[5]

Clustering is a group activity in a wide range of data records, observation, or approach an object with data from the other objects have in common. A cluster is a set of data rows that have in common and the dissimilarity to record into another cluster.[4]

The steps do clustering is as follows:

a. Determine the Number of Clusters You Want to Form From a Data Set
b. Determine the location of the center or cluster centroid randomly
c. For each row of data, find the cluster closest to the center of the cluster.
d. For each cluster, find the cluster centroid and update the location of each cluster center to get the new centroid value
e. Repeat steps 3 through step 5 until a convergent cluster is encountered as the end of clustering

The formula for calculating the distance between two points with Euclidean distance space uses a formula

\[ d_{\text{Euclidean}}(x, y) = \sqrt{\sum (x_i - y_i)} \] \hspace{1cm} (1)

where:

\( x \) and \( y \): representation of attribute values of two records

Whereas to generate a new centroid can use the following centroid generation formula:
\[ C = \frac{\sum_{n} m}{n} \] 

Where:

- \( C \) : Centroid Data
- \( M \) : Data Members Included into Specific Centroid
- \( N \) : Amount of Data Becomes Specific Centroid Members

Example

- Variable

| No | Information | Value |
|----|-------------|-------|
| 1  | Very Often  | 80    |
| 2  | Often       | 70    |
| 3  | Rafery      | 60    |

- Patient data

| No | No. Pasien | Nama Pasien | Gejala             | Frekwensi | Nilai |
|----|------------|-------------|--------------------|-----------|-------|
| 1  | 001        | Eudi Santoso| Batuk Kering       | Sering    | 80    |
|    |            |             | Batuk Berdebat    | Sering    | 80    |
|    |            |             | Sesak Napas       | Sering    | 70    |
| 2  | 002        | Dwihusen Situmorang | Sesak Napas     | Sering    | 80    |
|    |            |             | Mengi             | Sering    | 20    |
|    |            |             | Gejala Tinubul Dini dina    | Sering    | 80    |
| 3  | 003        | Rani Aryani Lubis | Sesak Napas    | Sering    | 70    |
|    |            |             | Mengi             | Sering    | 80    |
|    |            |             | Batuk             | Sering    | 80    |
| 4  | 004        | Puwasti Sunanjato | Gejala Tinubul Dini   | Jarang    | 70    |
|    |            |             | Mengi             | Sering    | 30    |
|    |            |             | Batuk             | Jarang    | 30    |
- Variable Value

| No | Pasien | V1 | V2 | V3 |
|----|--------|----|----|----|
| 1  | 001    | 160| 70 | 0  |
| 2  | 002    | 240| 0  | 0  |
| 3  | 003    | 160| 70 | 0  |
| 4  | 004    | 0  | 70 | 120|

The steps in clustering the dataset are:

- Determine the number of clusters, for example 2 cluster
- Generate initial centroids randomly as many as clusters

Centroid-1 (0.0,70.0,120.0)
Centroid-2 (160.0,70.0,00.0)

Calculate the distance space data to each centroid with the Euclidean formula, then enter the data into the cluster with the closest distance space.

- calculate the first iteration
  - 001 to Centroid-1 = \(\sqrt{(160.0-00.0)^2 + (70.0-70.0)^2 + (00.0-120.0)^2} = 200.0\)
  - 001 to Centroid-2 = \(\sqrt{(160.0-160.0)^2 + (70.0-70.0)^2 + (00.0-00.0)^2} = 00.0\)
  - 002 to Centroid-1 = \(\sqrt{(240.0-00.0)^2 + (00.0-70.0)^2 + (00.0-120.0)^2} = 277.3\)
  - 002 to Centroid-2 = \(\sqrt{(240.0-160.0)^2 + (00.0-70.0)^2 + (00.0-00.0)^2} = 106.3\)
  - 003 to Centroid-1 = \(\sqrt{(160.0-00.0)^2 + (70.0-70.0)^2 + (00.0-120.0)^2} = 121.06\)
  - 003 to Centroid-2 = \(\sqrt{(160.0-160.0)^2 + (70.0-70.0)^2 + (00.0-00.0)^2} = 00.0\)
  - 004 to Centroid-1 = \(\sqrt{(00.0-00.0)^2 + (70.0-70.0)^2 + (120.0-120.0)^2} = 0.0\)
  - 004 to Centroid-2 = \(\sqrt{(00.0-160.0)^2 + (70.0-70.0)^2 + (120.0-00.0)^2} = 200.0\)

Cluster groups in the first iteration

| No | Pasien | V1 | V2 | V3 | Group Awal |
|----|--------|----|----|----|------------|
| 1  | 001    | 160| 70 | 0  | 0          |
| 2  | 002    | 240| 0  | 0  | 0          |
| 3  | 003    | 160| 70 | 0  | 0          |
| 4  | 004    | 0  | 70 | 120| 0          |

| No | Pasien | V1 | V2 | V3 | Group Baru |
|----|--------|----|----|----|------------|
| 1  | 001    | 160| 70 | 0  | B          |
| 2  | 002    | 240| 0  | 0  | B          |
| 3  | 003    | 160| 70 | 0  | B          |
| 4  | 004    | 0  | 70 | 120| A          |

Because the first and second iterations have not been the same, then it continues again until it has the same results.
Repeat generate the centroid:
\[ C = \frac{\sum m}{n} \]

Centroid-1 \((\frac{0}{5}, \frac{70}{2}, \frac{120}{5}) = (0.0, 70.0, 120.0)\)

Centroid-2 \((\frac{560}{5}, \frac{140}{2}, \frac{6}{5}) = (186.6, 46.6, 0.0)\)

- Calculate the second iteration
  - 001 to Centroid-1 = \(\sqrt{(160.0-0.0)^2 + (70.0-70.0)^2 + (0.0-120.0)^2} = 200.0\)
  - 001 to Centroid-2 = \(\sqrt{(160.0-186.6)^2 + (70.0-46.6)^2 + (0.0-0.0)^2} = 34.97\)
  - 002 to Centroid-1 = \(\sqrt{(240.0-0.0)^2 + (0.0-70.0)^2 + (0.0-120.0)^2} = 277.3\)
  - 002 to Centroid-2 = \(\sqrt{(240.0-186.6)^2 + (0.0-46.6)^2 + (0.0-0.0)^2} = 70.87\)
  - 003 to Centroid-1 = \(\sqrt{(160.0-0.0)^2 + (70.0-70.0)^2 + (0.0-120.0)^2} = 121.06\)
  - 003 to Centroid-2 = \(\sqrt{(160.0-186.6)^2 + (70.0-46.6)^2 + (0.0-0.0)^2} = 34.97\)
  - 004 to Centroid-1 = \(\sqrt{(0.0-0.0)^2 + (70.0-70.0)^2 + (120.0-120.0)^2} = 0.0\)
  - 004 to Centroid-2 = \(\sqrt{(0.0-186.6)^2 + (70.0-46.6)^2 + (120.0-0.0)^2} = 149.82\)

Group of clusters in the first iteration

| No | Patien | V1 | V2 | V3 | Group Awal |
|----|--------|----|----|----|------------|
| 1  | 001    | 160| 70 | 0  | E          |
| 2  | 002    | 240| 0  | 0  | E          |
| 3  | 003    | 160| 70 | 0  | E          |
| 4  | 004    | 0  | 70 | 120| A          |

| No | Patien | V1 | V2 | V3 | Group Barn |
|----|--------|----|----|----|------------|
| 1  | 001    | 160| 70 | 0  | E          |
| 2  | 002    | 240| 0  | 0  | E          |
| 3  | 003    | 160| 70 | 0  | E          |
| 4  | 004    | 0  | 70 | 120| A          |

Because the first and second iterations have the same results, then no need to continue. But if it's not the same, then it's repeated again until it has the same results.

3. Design

In global design, the proposed system process design is described in the form of system modeling using UML (Unified Modeling Language) which in this method is designed using UseCase Diagram design, Class Diagram, Activity and Sequence Diagram.

3.1. Problem analysis

Problem Analysis is the decomposition of a complete information system into its component parts in order to identify and evaluate problems, opportunities and constraints that occur and the expected needs so that they can be proposed and improvements. System analysis is done to find out the advantages and disadvantages of the current system.

Patient Lung Disease Information System at Siti Hajar Hospital Medan includes recording and reporting that already uses computer programs, but still uses a simple application, namely microsoft office (word and excel), so that it cannot be used properly and correctly, where is the data still stored and input into Microsoft Office Excel.
3.2. Use case Diagram

In the preparation of a program, a data model in the form of a diagram is needed that can explain a system process flow that will be built. In writing this thesis the author uses the UML method in which the method the writer uses the Use Case diagram. Broadly speaking, the business process system that will be designed is illustrated by the usecase diagram contained in

![Figure 2. Use Case Diagram](image)

3.3. Class Diagram

The diagram used to display several classes and packages in the system / software that we are developing. • The class diagram (Class Diagram) gives us a picture (static diagram) of the system / software and the relationships in it. Class Diagram form of the system built can be seen in the picture below.

![Figure 3. Class Diagram](image)

4. Discussion

Every analysis that is made certainly has advantages and disadvantages, as well as from the system that will be designed by the author later. At this point the author will discuss the advantages of the designed system, there are also advantages of this system compared to the old system

a. Can save time and be more efficient because it does not need to consult directly with experts
b. An easy-to-use system
c. The level of trust generated is quite accurate.
d. Adding user knowledge about symptoms and types of lung disease

5. Lucky Analysis
Inevitably the system that will be designed will still have shortcomings. The following are the shortcomings of the system designed.
   a. This system only discusses the classification of types of lung diseases and their solutions
   b. The resulting output is only temporary in inputting patient data

6. Conclusion
After doing research on the design of data mining applications using clustering method to identify the type of lung disease and adjusting the system designed to the desired needs in the field, the conclusions that can be drawn from the author’s thesis in making data mining applications with clustering methods to identify diseases the lungs are as follows:
   a. The application of data mining used by the Clustering method can help classify the number of patients based on the type of lung disease and also based on the symptoms they suffer.
   b. With this system can later help to speed up data processing in classifying the number of patients based on the type of lung disease.
   c. Data Mining correctly performs Clustering calculations and comparisons with manual calculations provide accurate results.

Reference
[1]. Ardhyanti, Joanna. Kusumawati, Yupie. Data Mining Dengan Metode Clustering Untuk Pengolahan Informasi Persediaan Obat Pada Puskesmas Pandanaran Semarang, Semarang.
[2]. Lindawati. 2008, Data Mining dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara, Yogyakata, Indonesia.
[3]. Romario, Sandro. 2013, Penerapan Data Mining Pada RSUP DR.Moh Hoesin Sumatera Selatan Untuk Pengelompokkan Hasil Diagnosa Patient Pengguna Asuransi Kesehatan Miskin (ASKIN), Palembang
[4]. Oskar, Johan. 2013, Implementasi Algoritma K./ Means Clustering Untuk Menentukan Strategi Markleting President University, Bekasi