**Data Mining Analysis For Medical Record Data Clustering Using K-Means Algorithm In Sylvani Binjai Hospital**

Wahyu Surya Nanda¹, Akim M.H. Pardede², Magdalena Simanjuntak³

¹,²,³ Fakultas Teknik Informatika, STMIK Kaputama, Binjai, Indonesia

**ABSTRACT**

Medical record is a record of the history of patients who take treatment in hospitals or clinics. RSU Sylvani has many patients, every month and makes patient history data accumulate in the medical record data, but there is no follow-up benefit from the available data. Even though these data have great potential to provide new information and valuable insights if explored with data mining using the k-means clustering method. The amount of data that was tested was 893 data and produced 4 groups from the variables of disease diagnosis, gender and address. Where group 1 totaled 268 data with a diagnosis center for hypertension and female gender at the Pepper Garden address. Group 2 totaled 289 data with a diagnosis center for Asthma and female gender at the Hero's address. Group 3 totaled 185 data with Gerd disease diagnosis center and male gender at Kebun Pepper address. group 4 totaling 151 data with a diagnosis center for Prostate Enlargement disease and male sex at Kebun Pepper address.

This is an open access article under the CC BY license.
1. INTRODUCTION

In this digital era, everything is efficient with the use of information technology. Medical records or commonly referred to in the health sector are ICD (International Classification of Diseases) which are records of the history of patients taking treatment in hospitals or clinics. The medical language used by doctors in making a diagnosis then provides action on the disease received by the patient in the form of medical language (medical record) which then the results are coded by a medical record expert into ICD codes. This code is a standard language that can be used by all doctors, even if they are not specialists, to read it in accordance with the rules that apply to the code (ICD). In addition, the use of medical records that follow technology provides benefits for doctors and health workers in accessing patient information which ultimately helps in clinical decision making.

Recording of medical records is mandatory for hospital doctors who perform medical procedures on patients according to the rules so that there is no reason for doctors not to make medical records. Likewise with medical services. Medical services are one of the things that must be improved. With the increase in medical services, the science of medical records is very important to learn, while many people still do not understand about medical records [1] [2]. Medical record is a file that contains patient data, patient data at the hospital contains quite large data because every patient visiting or checking their health at the hospital will be used as a medical record which makes the data accumulate which causes various problems that often occur such as inadequate storage space, size, loss of medical records, expenditure of the required data, and others.

Sylvani public hospital is one of the hospitals in Binjai City. Sylvani Hospital has a lot of patients, every month the patients at Sylvani Hospital always come, so patient history data accumulates in the medical record data, and there is no follow-up benefit from the available data. Yet these data have great potential to provide valuable insights to evidence-based health professionals, researchers, and policy makers. However, managing and analyzing large and complex data is a significant challenge. One of the emerging approaches to overcome this challenge is data mining using the k-means clustering method. In addition, with the k-means clustering method using medical record data from Sylvani Binjai Public Hospital patients, the authors can see patterns of disease spread in Binjai District.

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 [3]. Data mining is a fairly new branch of computer science that is widely used and studied by computer scientists and programmers. Data mining is a concept that is intended to find valuable knowledge or information hidden in large databases or very large amounts of data. Data mining has several techniques in generating important information from piles and different techniques, such as description, estimation, prediction, classification, clustering and association. The application of data mining is widely used in
various fields, such as education, government, business or commerce and health (hospitals). The data processing technique in this study is using data mining techniques by applying the K-Means Clustering algorithm.

Previous researchers have used the K-Means Clustering method, including clustering patient medical record data using K-Means Clustering at Anwar Medika Bolong Bendo Hospital, Sidoarjo. The conclusion obtained is to produce 4 clusters consisting of sub-districts, disease diagnoses, age and sex. Cluster 1 produced a large number of patients consisting of 79 (15%) female patients, cluster 2 produced a large number of patients consisting of 214 (50%) male patients, cluster 3 produced a large number of patients consisting of 89 (17%) female patients, cluster 4 produced a large number of patients consisting of 152 (28%) male patients. This grouping of patient medical record data generates new information about the pattern of grouping the spread of disease in each sub-district based on patient medical record data from Anwar Medika Hospital as many as 534 data with a completion time of 0.06 seconds [4].

Another researcher is about clustering medical record data of patients who use BPJS health services using the K-Means method. The resulting conclusion is that the clustering process using the K-Means algorithm has succeeded in grouping BPJS patient medical record data into 3 clusters with different criteria. Disease with code A09.9 (Diarrhoea/Dysentery) dominates in cluster 0, the distribution of disease types is more diverse in cluster 1 and the spread of disease types with code K30 (Dyspepsia) is more dominated in cluster 2 [5].

Another researcher is about the application of data mining for grouping patient medical record data based on the type of disease with the K-Means Algorithm. The resulting conclusion is that the data mining method of the K-Means Algorithm can help classify the medical record data of PT polyclinic patients. Inecda by region, sex and age. The number of patients based on age, the first is adults with a total of 4912 patients, the second is the category of children with a total of 1262 patients and the third is the toddler category with a total of 144 patients. The number of diseases with the highest number of patients was ISPA with a total of 1985 patients due to the housing environment of PT. Inecda which is an oil palm plantation and PKS (Palm Oil Factory) and also the most common diseases with other diseases such as falling from motorbikes, cholesterol, pregnancy control checking blood pressure with a total of 2142 patients [6].

Another researcher is about the use of medical records in determining disease clusters through data mining at Aisiyiah Siti Fatimah Hospital, Tulangan Sidoarjo. With the conclusion of the RapidMiner software clustering process with 5 clusters, C0 is dominated by diagnoses I11.0, C1 is dominated by diagnoses E11.8, C2 is dominated by diagnoses A15.3, C3 is dominated by diagnoses E05.9, C4 is dominated by diagnoses E11.8. Cluster validity test using the Davies Bouldin Index (DBI) method with 5 clusters scored 0.469 [7].

2. METHOD

In solving a problem in research, of course the researcher must have a way or a method that will be applied in solving the problem so that the research carried out can be completed properly and in accordance with the expected results. The research method is carried out to look for something systematically by using the scientific method and applicable sources. On the basis of the research methodology used in this study, a flow of research work method activities can be made as follows:
2.1 Flowchart

In using the clustering method, the initial process for forming clusters is to transform the data into numeric form with predetermined codes, then determine the number of groups (K), calculate the centroids, calculate the object distance to the centroids and then group them based on the closest distance. If no objects are moved or grouped then the iteration is finished. The system process or workflow for clustering patient medical records at RSU Sylvani using the clustering method can be seen in the form of a flowchart as shown in the image below:
3. RESULT AND DISCUSSION

In discussing this interface, we will explain the results of program design that uses the GUI (Graphical User Interface) in the MATLAB R2014a software, and can be seen as follows:

1. Home Menu

Main Menu / Home Menu here displays the initial interface which contains home, cluster, and exit. When clicked on the cluster, it will immediately go to the cluster page. The display of the home menu page is as follows:

![Figure 3. Display of the Home Menu](image)

2. Cluster

Cluster menu, here 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 using the k-means algorithm. For more details as follows:

![Figure 4. Display of the Cluster Menu](image)

Based on the processed data, clustering results will appear for each centroid and a clustering result graph will also appear. The results of the clustering graph can be seen in the following figure:
4. CONCLUSION

It is known that the results of the 2 cluster process show the distribution of data into two groups based on the distance of the data to the centroid / cluster center. Where cluster 1 has 346 data and cluster 2 has 547 data. From these data it can be seen that in cluster 1 there are 346 data with a diagnosis center for Myalgia (Muscle Pain) and male gender at the Pepper Garden address and in cluster 2 there are 547 data with a diagnosis center for Dengue Hemorrhagic Fever (Dengue Hemorrhagic Fever) and type female gender in Hero's address. Then it is known that the results of the 3 cluster process show the distribution of data into three groups based on the distance of the data to the centroid / cluster center. Where cluster 1 has 277 data, cluster 2 has 319 data and cluster 3 has 297 data. From these data it is known that in cluster 1 there are 277 data with a hydronephrosis disease diagnosis center and male gender at the Pepper Garden address. Whereas in cluster 2 there are 319 data with a diagnosis center for Asthma (Asthma) and male gender at the Hero's address. And in cluster 3 there are 297 data with a diagnosis center for Congestive Heart Failure (Heart Failure) and male gender at Kebun Lada address. Next, it is known that the results of the 4 cluster process show the distribution of data into four groups based on the distance of the data to the centroid / cluster center. Where in cluster 1 there are 268 data, cluster 2 is 289 data, cluster 3 is 185 data and cluster 4 is 151 data. In the data cluster 1 there are 268 data with a diagnosis center for Hypertension (High Blood Pressure) and female gender at the Pepper Garden address. In cluster 2 there are 289 data with a diagnosis center for Asthma (Asthma) and female gender at the Hero's address. In cluster 3 there are 185 data with a diagnosis center for Gastro Esophageal Reflux Diases (GERD) and male gender at Kebun Lada and in cluster 4 there are 151 data with a diagnosis center for Enlarged Prostate disease (enlarged prostate) and male gender man at the address of Kebun Pepper.

REFERENCE


