International Journal of Research and Review
Vol.7; Issue: 4; April 2020
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Implementation of Data Mining Using Clustering Methods for Analysis of Dangerous Disease Data

Rahayu Mayang Sari

Faculty of Science and Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia

ABSTRACT

Method clustering with K-means algorithm used in data mining has the aim to explore information and knowledge from different perspectives but support each other method. Clustering with K-means algorithm gives information about the grouping of types of dangerous diseases, which suffer by patients. Analysis was performed using software Tanagra data mining1.4.38. The results of the data mining process are expected to provide its own benefits for the Islamic Hospital "Ibn Sina" Payakumbuh through new knowledge generated. Data mining is needed because there is a large amount of data that can be used to produce information and knowledge useful. The information and knowledge obtained can be used in many fields, ranging from business management, production control, health, etc.

Keyword: Data Mining, Clustering, K-Means

I. INTRODUCTION

In the era of globalization, the development of increasingly sophisticated technological sophistication is an aspect that is increasingly can be used to achieve facilities, no exception in the flow of information. The sophistication of the technology seems increasingly widespread with the use of computers that are already very wide in various fields of life for example in the fields of education, health, entertainment, especially in businesses that all require the use of computers.

In the Islamic Hospital "Ibnu Sina" Payakumbuh there are still processes that are done manually so that there is an error in the process of recording existing data and also the lack of time efficiency required. From this background a computerized information system is needed that supports the flow of data and information in accordance with the needs of these processes.

Method clustering with K-means algorithm used in data mining has the aim to explore information and knowledge from different points of view but support each other method. Clustering with K-means algorithm gives information about the grouping of types of dangerous diseases, which suffer by patients. Analysis was performed using software Tanagra data mining1.4.38. The results of the data mining process are expected to provide its own benefits for the Islamic Hospital "Ibnu Sina" Payakumbuh through new knowledge generated.

II. LITERATURE REVIEW

2.1 Understanding Data Mining

Data mining is a term used to describe the discovery of knowledge in the database. Data mining is a process that uses statistical techniques, mathematical, artificial intelligence, and machine learning to extract and identify useful information and knowledge that is assembled from various databases.

Broadly speaking, data mining can be grouped into 2 main categories:

1. Descriptive mining
   Process to find the important characteristics of data in one database. Data mining techniques that include descriptive mining are clustering, association, and sequential mining.
2. Predictive

The process of finding patterns from data using several other variables in the future. One of the techniques contained in predictive mining is classification.

2.2 Clustering

Cluster analysis is a technique used to identify similar objects or individuals by taking into account several criteria. Cluster analysis is analysis to group elements that are similar as research objects into distinct (mutually exclusive) clusters.

Cluster analysis is included in the multivariate statistical analysis of the interdependent method. As an interdependent analysis tool, the purpose of cluster analysis is not to link or differentiate with other samples or variables. Cluster analysis is one analysis tool that is useful as a summation of data. In summarizing this data can be done by grouping objects based on the similarity of certain characteristics among the objects to be studied.

2.3 Method of K-Means

K-Means is one method of data clustering non hierarchy that seeks to partition the data into the form of one or more clusters or groups so that the data that have the same characteristics are grouped into the same cluster and data which have characteristics different groups are grouped into other groups.

K-Means is a distance-based clustering method that divides data into a number of clusters and this algorithm only works on numeric attributes. The K-Means algorithm includes partitioning clustering which separates data into separate k regions. The K-Means algorithm is very well known for its ease and ability to cluster large data and outlier data very quickly. In the K-Means algorithm, each data must belong to a particular cluster and can be made possible for every data that belongs to a particular cluster at a stage of the process, in the next stage, move to other clusters.

Basically the use of algorithms in the clustering process depends on the existing data and the conclusions to be achieved. For this reason, the algorithm is used K-Means which makes the following rules: The

1. Number of clusters needs to be inputted
2. Only has numeric attributes

The K-Means algorithm is a non-hierarchical method that initially takes a large portion of the population component to be the center of the initial cluster. At this stage the cluster center is chosen randomly from a collection of data populations. Next K-Means tests each component in the data population and marks the component to one of the cluster centers that have been defined depending on the minimum distance between components with each cluster. The position of the cluster satisfaction will be recalculated until all data components are classified into each -the center of the cluster and finally the new cluster satisfaction position will be formed.

The K-Means algorithm basically carries out two processes, namely the process of detecting the central location of each cluster and the process of finding members from each cluster.

How the K-Means algorithm works:

1. Determine k as the number of clusters you want to form.
2. Awaken the k initial centroid (cluster center point) randomly.
3. Calculate the distance of each data to each centroid.
4. Each data selects the closest centroid.
5. Determine the new centroid position by calculating the average value of data located in the same centroid.
6. Return to step-3 if the new centrroid is not the same as the old centroid.

III. MATERIALS AND METHODS

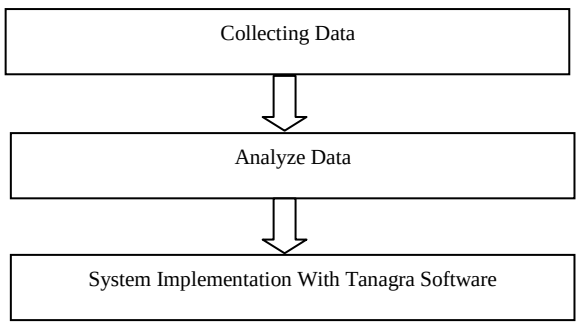

Figure 1 Research Methodology

3.1 Collecting Data

Data collected is income data. The data is studied and grouped, the results of the grouping will be obtained later problems and will be solved, then the solution is sought.

3.2 Analyzing Data

Analysis using the application Microsoft Office, at this stage the income data selection is done, cleaning data that is removing data duplication, correcting data errors so as to produce information that can be processed and evaluated so that the data is easy to understand.

3.3 System Implementation with Tanagra Software

At this stage it is a process of processing data that has been collected using software Tanagra, using the K-Means algorithm.

IV. RESULT AND DISCUSSION

4.1 Test Results

a. Input Dataset

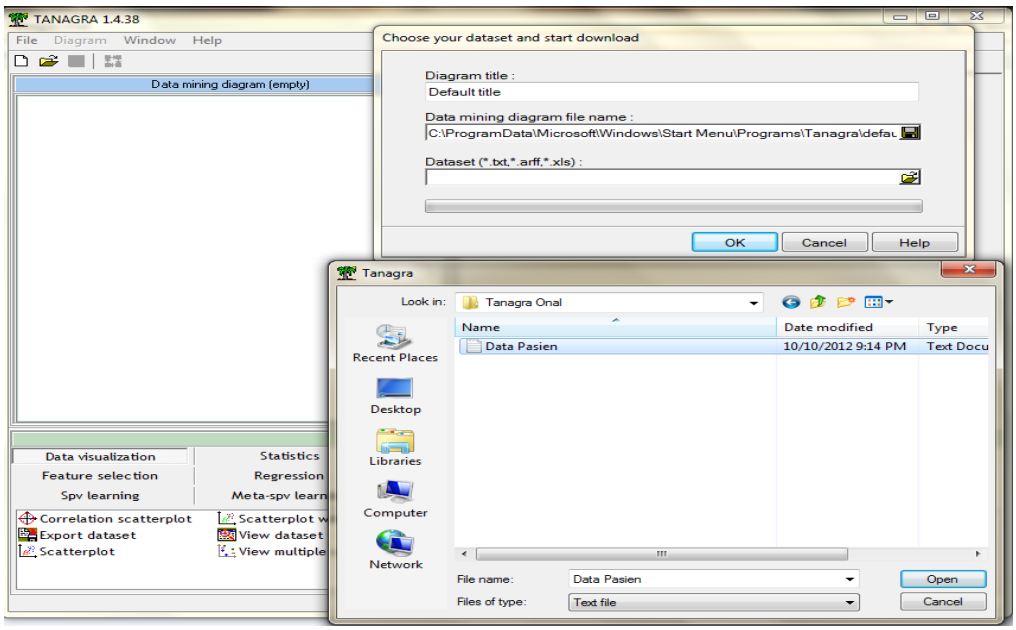

Figure 2 Input Dataset

b. Input Define Status

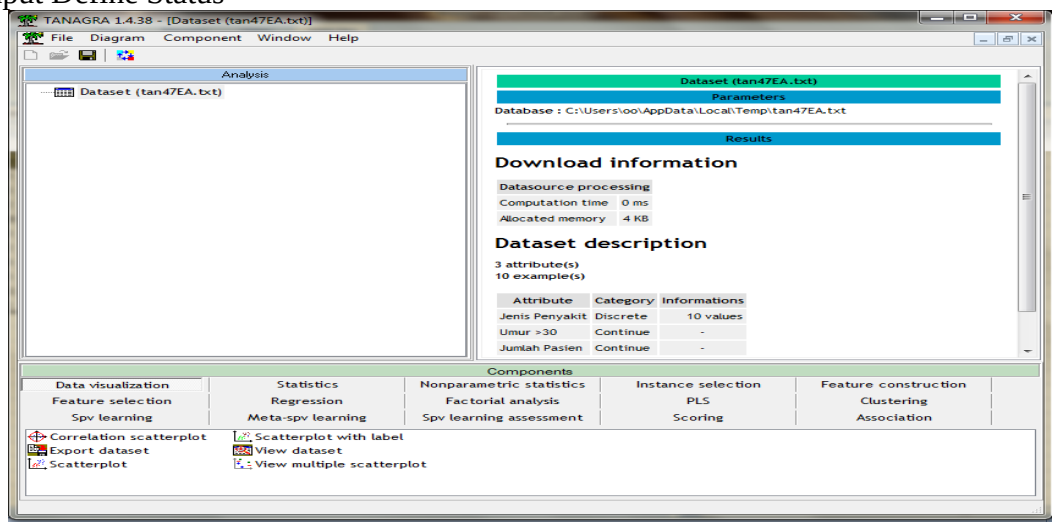

Figure 3 Input Define Status

c. Standardize

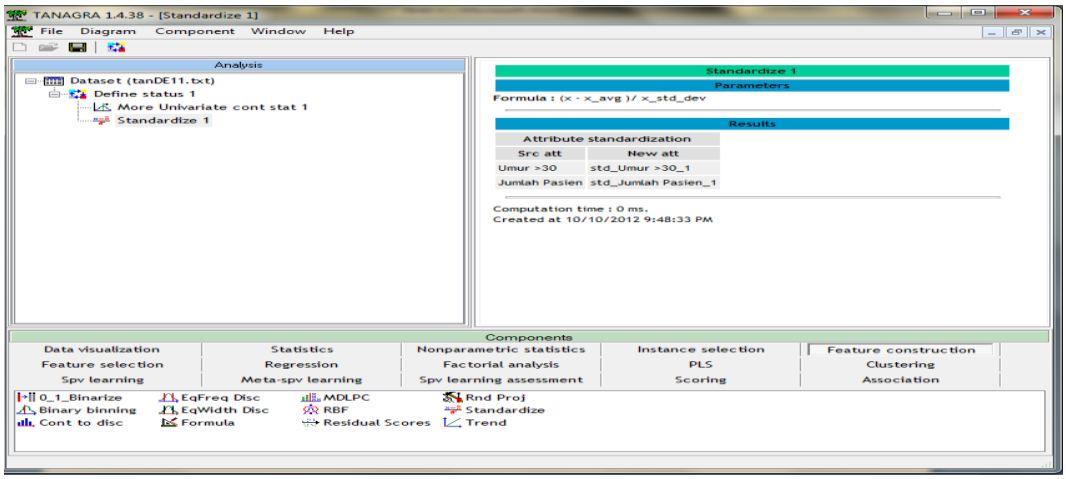

Figure 4 Standardize

d. Add K-Means Components

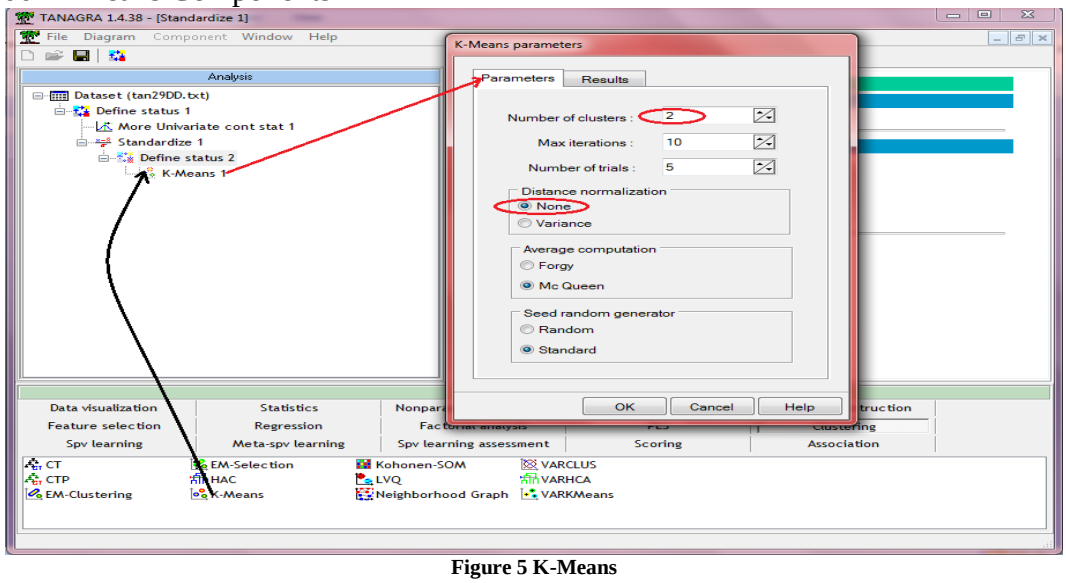

Figure 5 K-Means

e. View Dataset

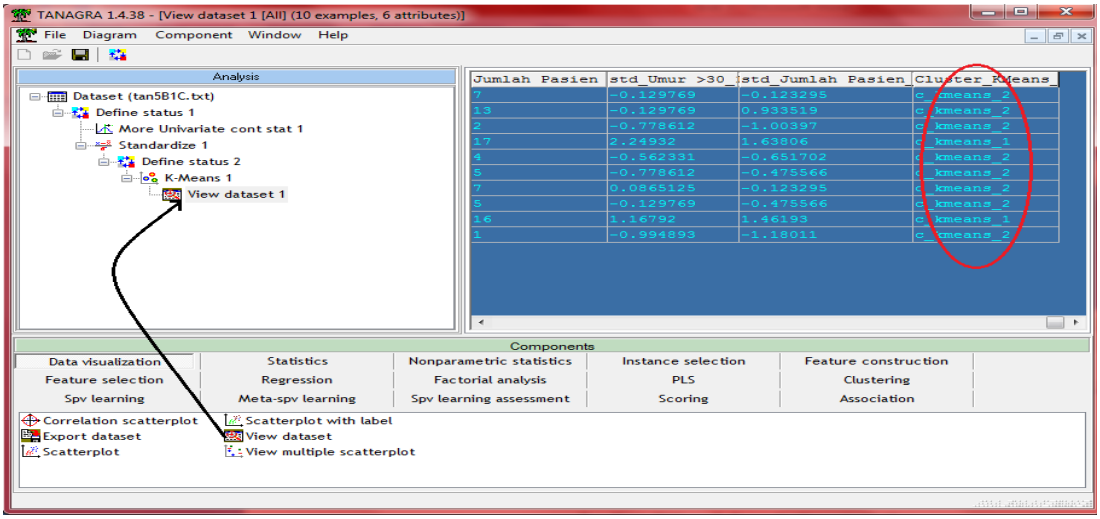

Figure 6 Datasets

f. Add Component Group Characteristics.

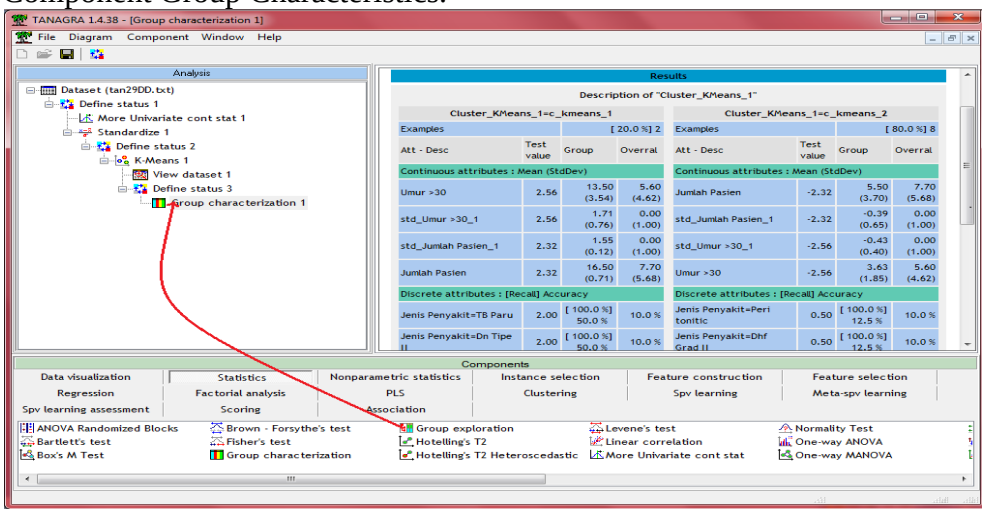

Figure 7 Group Characteristic

g. Add Component scatterplot

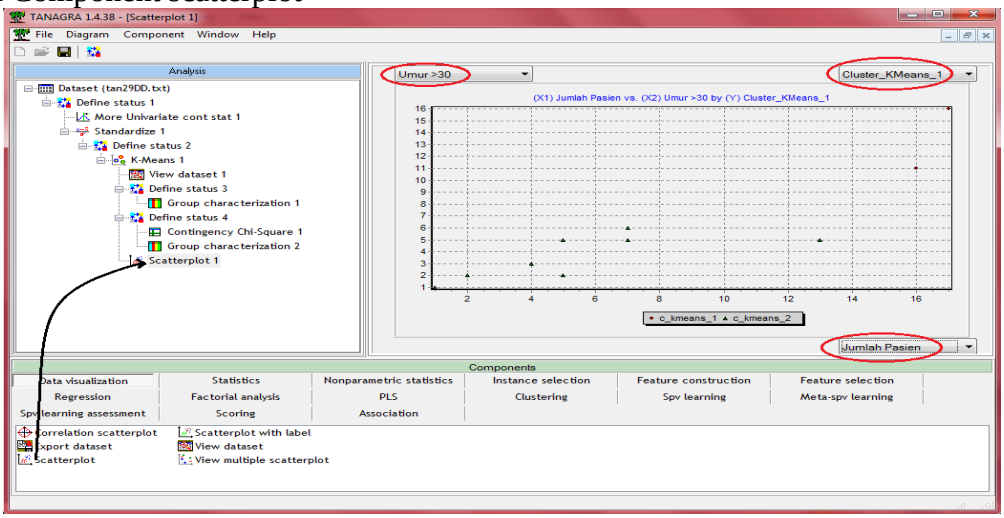

Figure 8 scatterplot

h. Add Component Principal Component Analysis

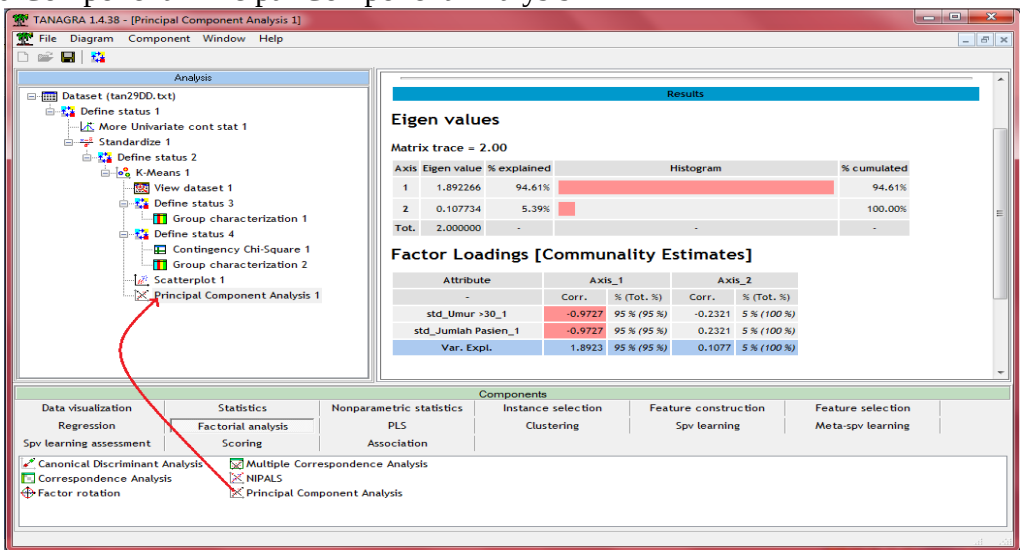

Figure 9 Principal Component Analysis

i. Exsport Dataset

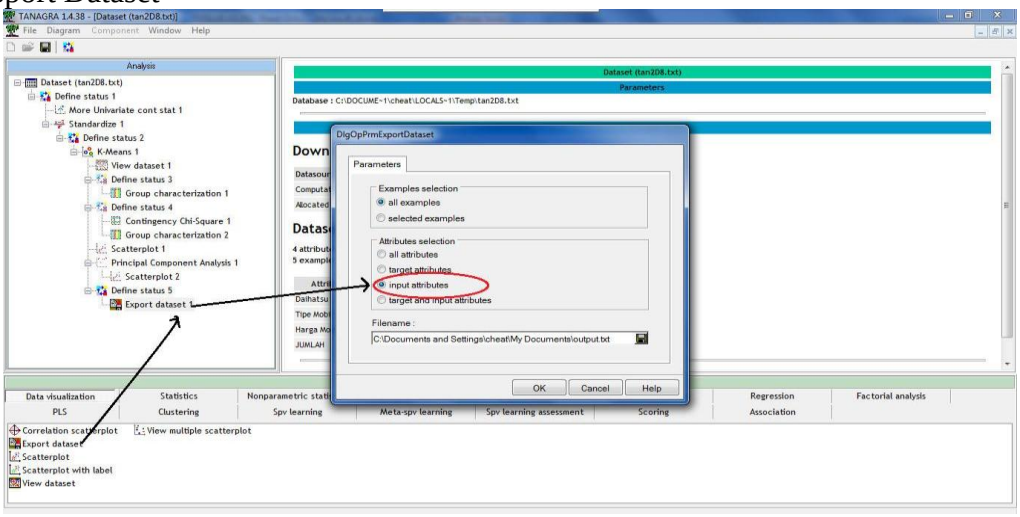

Figure 10 Export Dataset

CONCLUSION

Based on the description that has been discussed before it can be concluded:

1. This study analyzed data malignancies of RSI Avicenna Payakumbuh use clustering with the K-Means algorithm.
2. System clustering Dangerous disease data using K-Means algorithm can classify dangerous diseases that are suffered at the age of >30 years.

REFERENCES

1. N. Sartika, K. Kirmizi, and N. Indrawati, "Analysis of Factors in Regional Budget Structure and Regional Financial Performance that Affects Capital Expenditures in Regencies / Cities in Riau Province," Sorot, vol. 12, no. 2, p. 121, 2017.
2. Y. Apridonal M, "Applying Data Mining Using The Association Rule Method" Jurteksi vol. V, no. 2, pp. 193–198, 2019.
3. RA Asroni, "The Application of the K-Means Method for Student Clustering Based on Academic Values with Weka Interface Case Studies in the Department of Informatics Engineering Magelang UMM," Ilm. Universe Tech., vol. 18, no. 1, pp. 76–82, 2015.
4. A. Bastian, H. Sujadi, and G. Febrianto, "Application of the K-Means Clustering Analysis Algorithm in Human Communicable Diseases (Case Study of Majalengka Regency)," J. Sist. Inf. (Journal of Inf. Syst., Vol. 14, no. 1, pp. 26–32, 2018.
5. NI Febianto and N. Palasara, "Analysis of K-Means Clustering on Poverty Information Data in West Java 2018," J. Sisfokom (Inf. And Computer Systems), vol. 8, no. 2, p. 130, 2019.
6. RM Sari, "Predicting Regional Budget Revenue Using the K-Means Algorithm" SATIN Vol. 1, no. 2, Dec 2015.

How to cite this article: Sari RM. Implementation of data mining using clustering methods for analysis of dangerous disease data. International Journal of Research and Review. 2020; 7(4): 237-242.

******