K-MEANS CLUSTERING FOR DISEASE SPREAD AREAS DENGUE HEMORRHAGIC FEVER (DHF) IN EAST LOMBOK NTB

Yunita Wilawardani[1], Wiwit Pura Nurmayanti[1*], Sausan Nisrina[1], Ristu Haiban Hirzi[1], Abdul Rahim[2]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Hamzanwadi (INDONESIA)
[2]Department of Pharmacy, Faculty of Pharmacy, Universitas Mulawarman (INDONESIA)

**Corresponding author: wiwit.adiwinata3@gmail.com*

Abstract

K-Means is one algorithm in Data Mining that is used to categorize or cluster data. The advantages of K-Means compared to other cluster methods are that it is easy to implement and can be scalable for large datasets. K-Means can be applied in various fields, one of which is the health sector, namely Dengue Haemorrhagic Fever (DHF) data. DHF is one of the environmental health problems that increases in East Lombok Regency, NTB, so clustering is necessary to see the spread of DHF itself. The purpose of this study was to view the description of DHF data and to classify the areas of DHF distribution in East Lombok. Based on the results of the analysis, information was got that the highest number of cases occurred in Selong District, 85 cases and the lowest cases were in Suela and Sembalun Districts, where there were no dengue cases. For the DHF distribution area, we got three clusters. Cluster-1 with a high category that is 13 sub-districts, and Cluster-2 with a low category that is 8 sub-districts.

Keywords: Data Mining, K-Means, Dengue Haemorrhagic Fever (DHF), Clustering, Lombok

1 INTRODUCTION

Clustering is a tool in data mining that aims to group objects into clusters [1]. The distance between two clusters is the distance between the centroids of the cluster [2]. In data mining, there are two clustering methods that can be used in grouping, namely hierarchical clustering and non-hierarchical clustering [3]. The hierarchical clustering method consists of complete linkage clustering, single linkage clustering, average linkage clustering and ward's method. Meanwhile, in the non-hierarchical clustering method, there is K-Means clustering [4]. K-Means is one of the non-hierarchical data clustering methods that seeks to partition existing data into the form of one or more cluster [5]. This method partitions data into clusters so that data that are into the same cluster and data that have different characteristics are grouped into other groups [3]. There are many different fields that can be used in grouping using K-Means, one of which is the health field.

Dengue Haemorrhagic Fever (DHF) is one of the environmental health problems that tends to increase the number of sufferers and the wider the area of its spread, in line with the increasing mobility and population density caused by the dengue virus and transmitted through the bite of the Aedes Aegypti mosquito [6]. The impact of dengue fever can make the patient's body temperature very high and is generally accompanied by fever, nausea vomiting, headache, abdominal pain, and leucopoenia [1]. Dengue Fever is still one of the problems of public health [7]. The health office has the main task of assisting in the implementation of environmental health activities.

East Lombok Regency is one of the districts located in the east of the island of Lombok, province West Nusa Tenggara. The capital of East Lombok is in Selong District with an area of 1230. 76 km2 and the population in 2020 reached more than one million people spread across 21 districts [8]. The increasing population density in East Lombok Regency and the lack of public awareness of environmental health will have an impact on the environment due to the emergence of various diseases, one of which is Dengue Haemorrhagic Fever (DHF).

There have been several studies related to the K-Means method that have been carried out, including Yunita [9], with the title Application of Data Mining K-Means Clustering Algorithm in New Student Admissions. In this study, three clusters were formed, with the first cluster of 195 items, the second cluster of 271 items and the third cluster of 50 items. The results of this study are used as one of the bases for decision making to determine strategies to promote each existing study program [9]. Furthermore, the related research was also carried out by Fatmawati [10] with the title Application of Rapidminer with K-Means Cluster in Areas Infected with Dengue Hemorrhagic Fever (DHF) Based on Province. The results obtained that there were 4 provinces with high-level clusters (C1), 13 provinces with medium level (C2), and 17 provinces with low-level clusters (C3) [9]. Another relevant research was conducted by Darmansah and Wardani [16] with the title Analysis of the Spread of Corona Virus Transmission in Central Java Province using the K-Means Grouping, with the results found that C0 there are 18 cities/regencies, C1 there is 1 city/regency and C2 there are 16 cities/regencies at the level of the spread of the corona virus in Central Java province [11].

In this study, the K-Means method will group the areas in East Lombok Regency according to the level of dengue disease cases so that they are appropriately and quickly targeted in the prevention and control of dengue disease. Where in this case it will visualize the results of the cluster analysis of dengue disease spread in East Lombok Regency using K-Means with variables in the number of dengue sufferers, the number of health facilities and population density in each district in East Lombok Regency.

2 METHODOLOGY

2.1 Data Mining

Data mining is the process of obtaining useful information from a large database warehouse [12]. Data mining can also be interpreted as the installation of new information taken from free chunks of data that assist in decision making. The term data mining is sometimes also called knowledge discovery [13]. It is worth remembering that the word mining itself means an attempt to obtain a small number of valuables from a large number of basic materials. Therefore, data mining actually has long roots from fields of science such as artificial intelligence, machine learning, statistics and databases [14]. Mining data contains searches for desired trends or patterns in large databases to assist decision making in the time to be come [15].

2.2 Cluster Analysis

Cluster analysis is to find a collection of objects until objects in one group are the same (or have a relationship) with another and are different (or unrelated) to objects in another group [16]. The purpose of the analysis is to minimize the distance within the cluster and maximize the distance between clusters [17]. Cluster analysis is considered as a form of classification that labels objects with their class labels [11]. There are many method methods of clustering developed by experts. Each method has character, advantages, and disadvantages, one of which is the K-Means method [18].

2.3 K-Means Clustering

The K-Means method is one of the commonly used non-hierarchical methods. This method is included in the partitioning technique that divides or separates objects into separate area pok groups [1]. The purpose of the K-means is to divide n observations into group k all observations are part of a cluster that serves as a prototype cluster [18]. The K-Means algorithm uses a process repeatedly to obtain a cluster database [19]. The K-Means clustering algorithm is based on optimizing the similarity scale between each cluster with the lowest value and the highest value for the value in the cluster, in other words K-Means tries to reduce the distance between clusters and increase the similarity in the cluster [20]. The K-Means method will select the k pattern as the starting point of the centroid randomly or randomly. The number of iterations to reach the centroid cluster will be influenced by the initial centroid cluster candidate in random [21]. Here are the steps to use the K-Means method:

a. Specify k as the number of clusters you want to form.
b. Comparing random values for the initial center cluster (centroid) by k.
c. Calculates the distance of each input data to each centroid using the distance formula (Euclidean Distance) until the closest distance of each data is found with the centroid (Equation 1).

$$D(xi,\mu j) = \sqrt{\sum (\mathrm{xi} - \mu \mathrm{j})^2} \qquad (1)$$

d. Classify each data based on its proximity to the centroid (the smallest distance).

Renews the centroid value. The new centroid value is obtained from the average of the cluster concerned using Equation 2, where μj(t+1: new centroid on iteration to (t+1), and Nsj: lots of data on the Sj cluster.

$$\mu j(t+1) = \frac{1}{\mathrm{Nsj}} \sum_{\mathrm{j} \in \mathrm{sj}} \mathrm{xj} \qquad (2)$$

e. Loop from steps 2 to 5, until none of the members of each cluster have changed.

3 RESULT

3.1 Descriptive Statistics

Descriptive statistics is an activity in data collection, structuring, summarizing, and presenting data in the hope that the data is more meaningful and easier to understand. In this section, data exploration will be carried out which aims to describe the specifics of the data. The data used is data on the spread of dengue disease sufferers using the data attributes of the number of dengue sufferers in 2020 and geographical factor attributes in the form of population density, the number of health facilities per district in East Lombok Regency. Here will be shown some graphs that depict the characteristics of the data.

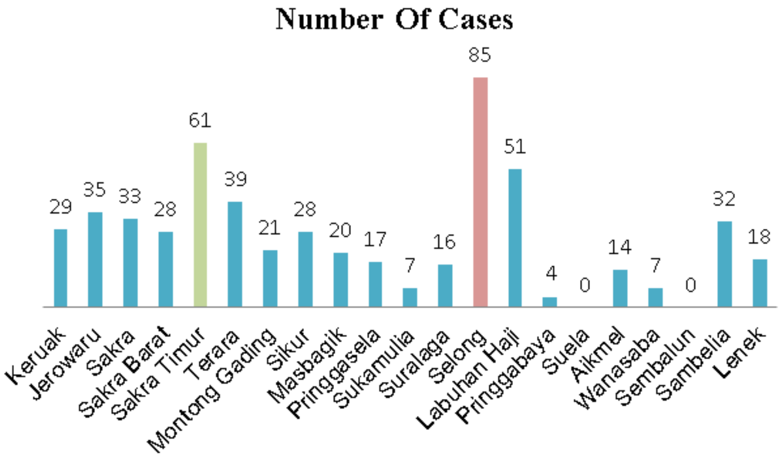

***Figure 1**. Number of Dengue Cases in East Lombok Regency.*

Based on ***Figure. 1,*** it shows a graph of the number of cases of Dengue Hemorrhagic Fever (DHF) patients in 21 districts in East Lombok Regency. The district with the most cases occurred in Selong District with 85 cases, followed by Sakra Timur District with 61 cases. Meanwhile, the districts with the lowest cases are found in Suela and Sembalun Districts, where there are no cases of Dengue Hemorrhagic Fever (DHF) from the districts.

3.2 K-Means Clustering

In this section, the researcher explains a picture of the process of analyzing a problem that occurs and the application of the method used. To assist researchers in analyzing data in the search for knowledge, a transformation of the data obtained from the data results from the East Lombok District Health Office will be carried out in the form of a number processing program file. To perform processing the K-Means Clustering algorithm is processed by using R Studio.

3.3 Determines the number of clusters and the initial centroid value.

At this stage, it is known that the most optimal number of clusters formed for grouping dengue disease spread in East Lombok Regency is 2 clusters. With the initial centroid values that are formed automatically, they are as follows:

Tabel 1. *Early Centroid.*

Cluster	Number of Cases	Population Density	Medical Facility
1	32.12500	2283.6250	92.75000
2	22.15385	821.2308	90.69231

Based on ***Table. 1***, the mean or centroid for each of the variables in the cluster group. This initial centroid is the data that became the center point of the first cluster in determining the grouping of dengue disease spread in East Lombok Regency. For example, the number of cases in cluster 1 centroid is 32.12500, population density in cluster 1 is 2283.6250 and health facilities in cluster 1 are 92.75000 and likewise with other clusters.

3.3.1 Algorithm K-Means Clustering

To determine the analysis of the level of spread of dengue transmission in East Lombok Regency using RStudio, with the results of the analysis as follows:

Table 2. *Centroid Value.*

Cluster	Number of Members	Number of Cases	Population Density	Medical Facility
1	13	32.1	2284	92.8
2	8	22.2	821	90.7

After analyzing with the K-Means algorithm using RStudio, the results were obtained, namely: there are two clusters, namely C1, and C2. Where C1 is the rate of spread of dengue disease in East Lombok Regency with a high category there are 13 districts and C2 with a low category there are 8 districts. The formed graph can be seen in ***Figure. 2***.

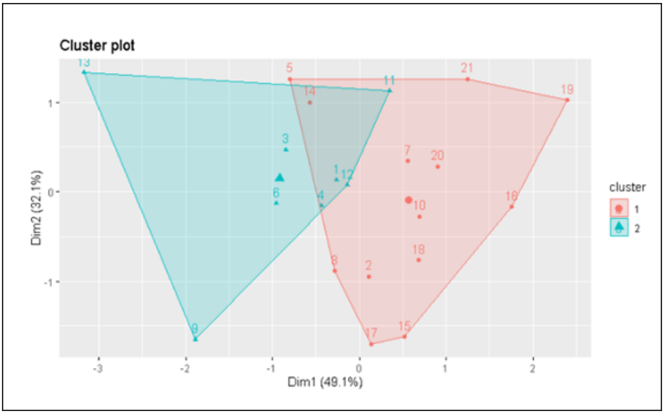

Figure 2. *Cluster Visualization.*

Based on the visualization of the spread of dengue disease in East Lombok Regency with the upper graph shows that there are two colors according to the number of clusters. There are 13 red-colored sub-districts that are members of C1, and 8 sub-districts in blue which are members of C2.

3.3.2 Map Visualization

After the K Means clustering process that has been carried out previously to find out the spread of dengue fever in East Lombok Regency, it will then be formed in the form of a map. In the map, each district has a level of dengue disease spread. Where the sub-districts with red dengue distribution areas are districts with high dengue distribution categories and sub-districts with yellow are low dengue distribution categories.

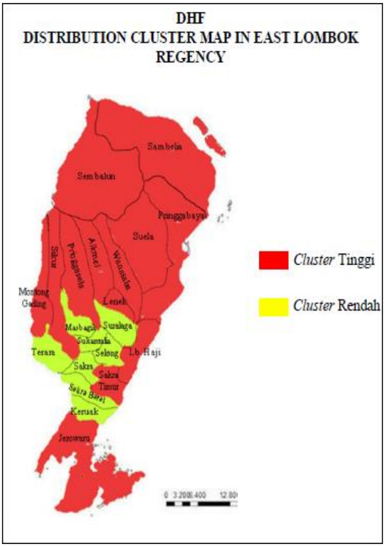

***Figure 3**. DBD Deployment Map.*

4 CONCLUSION

Based on the results of the analysis that has been carried out, it can be concluded that the results of the grouping of two clusters , namely C1 and C2, show that 21 districts in East Lombok Regency, 13 districts are members of C1 with high categories, and 8 districts are members of C2 with fairly low categories, and 4 districts are members of C3 with low categories. Grouping with K-Means using two clusters shows that 21 sub-districts in Lombok Regency The results of visualizing the spread of dengue disease in East Lombok Regency with a graph showed that there were two colors according to the number of clusters. There are 13 red-colored sub-districts that are members of C1, and 8 sub-districts in blue which are members of C2. The suggestion in this study is that this research can be developed using other data mining algorithms, especially algorithms for clustering techniques, or compare with other algorithms to get optimal results.

REFERENCES

[1] Hariyanto, M., & Shita, R. T. (2018). Clustering pada Data Mining untuk Mengetahui Potensi Penyebaran Penyakit DBD Menggunakan Metode Algoritma K-Means dan Metode Perhitungan Jarak Euclidean Distance. Sistem Komputer Dan Teknik Informatika, 1(1), 117–122.

[2] Saky, D. A. L., Jayanti, N. A., & Nurmayanti, W. P. (2020). Clustering Pertani Berdasarkan Dampak Covid-19 Yang Terjadi Pada Sektor Pertanian Studi Kasus di Dusun Lepak Desa Lepak Kecamatan Sakra Timur Kabupaten Lombok Timur NTB (Farmer Clustering Is Based On The Impact Of Covid-19 On The Agricultural Clustering). Seminar Nasional Official Statistics 2019: Pengembangan Official Statistics Dalam Mendukung Implementasi SDG's CLUSTERING, 2020(1), 160–164.

[3] Bastian, A., Sujadi, H., & Febrianto, G. (2018). Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka). 1, 26–32.

[4] Ramadhani, L., Purnamasari, I., & Amijaya, F. D. T. (2018). Penerapan Metode Complete Linkage dan Metode Hierarchical Clustering Multiscale Bootstrap (Studi Kasus: Kemiskinan Di Kalimantan Timur Tahun 2016). Eksponensial, 9(2016), 1–10. https://fmipa.unmul.ac.id/files/docs/[1] Lisda Ramadhani 1307015041_Edit.pdf

[5] Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. Jurnal Media Informatika Budidarma, 4(1), 51. https://doi.org/10.30865/mib.v4i1.1784

[6] Shafarini, A. Y., Moelyaningrum, A. D., & Ellyke. (2018). Pengaruh Penggunaan Serbuk Pare Gajih (Momordica charantia L.) Terhadap Kematian Larva Aedes aegypti. Higiene, 4(1), 11–18.

[7] Saragih, I. D., Fahlefi, R., Pohan, D. J., & Hartati, S. R. (2019). Analisis Indikator Masukan Program Pemberantasan Demam Berdarah Dengue Di Dinas Kesehatan Provinsi Sumatera Utara. Contagion: Scientific Periodical Journal of Public Health

and Coastal Health, 1(01). https://doi.org/10.30829/contagion.v1i01.4821

[8] BPS. (2020). LOMBOK TIMUR DALAM ANGKA 2020.

[9] Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa Baru. Sistemasi: Jurnal Sistem Informasi, 7(3), 238-249.

[10] Fatmawati, K., & Windarto, A. P. (2018). Data Mining: Penerapan rapidminer dengan K-means cluster pada daerah terjangkit demam berdarah dengue (DBD) berdasarkan provinsi. CESS (Journal of Computer Engineering, System and Science), 3(2), 173-178.

[11] Sari, D. N. P., & Sukestiyarno, Y. L. (2021). Analisis Cluster dengan Metode K-Means pada Persebaran Kasus Covid-19 Berdasarkan Provinsi di Indonesia. PRISMA, Prosiding Seminar Nasional Matematika, 4, 602–610. https://journal.unnes.ac.id/sju/index.php/prisma/

[12] Nurul, C., & Ari, W. I. (2018). Implementasi Data Mining Untuk Clustering Daerah Penyebaran Penyakit Demam Berdarah Di Kota Tangerang Selatan Menggunakan Algoritma K-Means. Jurnal Satya Informatika, 3(1), 12–24.

[13] Meilani, B. D., Wahyudiana, S., Putri, A. Y. P., & Pakarbudi, A. (2019). Klasifikasi Identifikasi Faktor Penyebab Ketidaktepatan Masa Lulus Mahasiswa dengan Metode Naïve Bayes Classifier. Seminar Nasional Sains Dan Teknologi Terapan, 297–302. https://ejournal.itats.ac.id/sntekpan/article/view/586

[14] Purwadi, Ramadhan, P. S., & Safitri, N. (2019). Penerapan Data Mining Untuk Mengestimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Deli Serdang. Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer), 18(1), 55. https://doi.org/10.53513/jis.v18i1.104

[15] Syahdan, S. Al, & Sindar, A. (2018). Data Mining Penjualan Produk Dengan Metode Apriori Pada Indomaret Galang Kota. Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI), 1(2). https://doi.org/10.32672/jnkti.v1i2.771

[16] Wardani, S. I., Nurmayanti, W. P., & Malthuf, M. (2020). Analisis Cluster Kecamatan Di Lombok Timur Berdasarkan Banyaknya Perusahaan Dan Cabang Industri. Journal Of Applied Statistics And Data Mining, 1(2).

[17] Akbar, I. (2019). Visualisasi Data Untuk Penyebaran Penyakit Demam Berdarah Di Kabupaten Jember Dengan Menggunakan Metode K-Means. http://repository.unmuhjember.ac.id/7138/

[18] Jahwar, A. (2021). Meta-Heuristic Algorithms for K-means Clustering: A Review. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 7–9. https://archives.palarch.nl/index.php/jae/article/view/4630

[19] Nurdiyansyah, F., & Akbar, I. (2021). Implementasi Algoritma K-Means untuk Menentukan Persediaan Barang pada Poultry Shop. Jurnal Teknologi Dan Manajemen Informatika, 7(2), 86–94. https://doi.org/10.26905/jtmi.v7i2.6377

[20] Zeinalkhani, L., Ali Jamaat, A., & Rostami, K. (2018). Diagnosis of Brain Tumor Using Combination of K-Means Clustering and Genetic Algorithm. Iranian Journal of Medical Informatics, 7(1), 6. https://doi.org/10.24200/ijmi.v7i0.159

[21] Indraputra, R. A., & Fitriana, R. (2020). K-Means Clustering Data COVID-19. 10(3), 275–282.