ISSN: 2581-1274

Utilization of Rapidminer using the K-Means Clustering Algorithm for Classification of Dengue Hemorrhagic Fever (DHF) Spread in Banda Aceh City

Sanusi[a,1,*], Juniana Husna[b,2]

[a] *Abulyatama University, Blangbintang Lama Street, Lampoh Keude, Kuta Baro, 24415, Indonesia*
[b] *Abulyatama University, Blangbintang Lama Street, Lampoh Keude, Kuta Baro, 24415, Indonesia*
[1] *sanusi@abulyatama.ac.id;* [2] *juniana@abulyatama.ac.id*

ARTICLE INFO

Article history:
Accepted

Keywords:
Dengue Hemorrhagic Fever (DHF)
K-Means *Clustering*
Kota Banda Aceh

ABSTRACT

Dengue Hemorrhagic Fever (DHF) is still a serious problem in Banda Aceh City. The grouping of dengue disease distribution areas can use data mining techniques through the K-Means Clustering Algorithm by involving several factors that influence it, such as population density, rainfall, air humidity, and temperature. The purpose of this study is to try to create a distribution cluster group that is included in the high category (C1), medium (C2), and low (C3) in 9 sub-districts in the city of Banda Aceh. The data used in this study are secondary data during the period 2010 to 2017, which includes population density data obtained from the Banda Aceh City BPS office, rainfall, humidity, temperature were obtained from the BMKG Indrapuri Aceh Besar office, and data on dengue cases were obtained from the Banda Aceh City Health Office. The results showed that up to 4 iterations of K-Means Clustering were good enough to classify dengue case data. The high cluster group (C1) is Baiturrahman, Kuta Alam, and Syiah Kuala sub-districts. The medium cluster group (C2) is Jaya Baru, Banda Raya, and Ulee Kareng sub-districts. The low cluster group (C3) is Meuraxa and Kuta Raja.

I. Introduction

Dengue Hemorrhagic Fever (DHF) is a disease caused by the dengue virus transmitted through the bites of Aedes aegypti and Aedes albopictus mosquitoes, which can cause disturbances in capillary blood vessels and the blood clotting system so that it can cause bleeding. [1], [2]. According to WHO data, Asia Pacific accounted for 75% of the world's dengue burden between 2004 and 2010. Indonesia is reported as the 2nd country with the largest DHF cases among 30 countries in endemic areas. The number of cases of Dengue Fever in Indonesia tends to increase over the next year. The increasing number of dengue fever in various cities in Indonesia is due to the difficulty of controlling the Aedes aegypti mosquito's disease. [3], [4], [5].

Dengue Hemorrhagic Fever (DHF) is still a severe problem in Aceh Province, one of which is Banda Aceh City, and this is a significant threat to the people of Banda Aceh City after the aftermath of the tsunami where the number of sufferers and the area of the environment transmitted by the mosquito Ae. Aegypti is still relatively high. Most cases were found in the Baiturrahman Community Health Center's work area, with 64 subjects and the least in the Lampaseh Health Center working area as many as 9 cases. In 2013 there were three deaths due to dengue cases that occurred in Batoh, Kuta Alam, and Lampaseh Health Centers, one each. [6].

Various factors that cause the incidence of DHF are often overlooked and not comprehensively implemented by the related institutions. One way to analyze and group the area of

dengue fever in the city of Banda Aceh can use data mining techniques through the K-Means Clustering Algorithm by involving several influencing factors such as population density, rainfall, humidity, and temperature. The purpose of this study was to create high (C1), medium (C2), and low (C3) cluster groups from 9 sub-districts in Banda Aceh city.

II. Method

The first step of this research is field observation and literature study. The data used in this study are secondary data for a period of 8 years from 2010 to 2017 consisting of population density data obtained from the Central Statistics Agency of Banda Aceh City and rainfall data, air humidity, temperature obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) Idrapuri Aceh Besar, and the data for DBD cases recovered from city health service Banda Aceh. After the preprocessing stage is complete, the next step is to classify using the K-Means clustering algorithm.

a. Research Location Map

Banda Aceh city is located in the province of Aceh, Indonesia. Longitude at position 05 16 '15 "- 05 36' 16" North Latitude and 95 16 '15 "- 95 22' 35" East Longitude with a mean height of 0.80 meters above sea level. The map of the research location can be view in Figure 1.

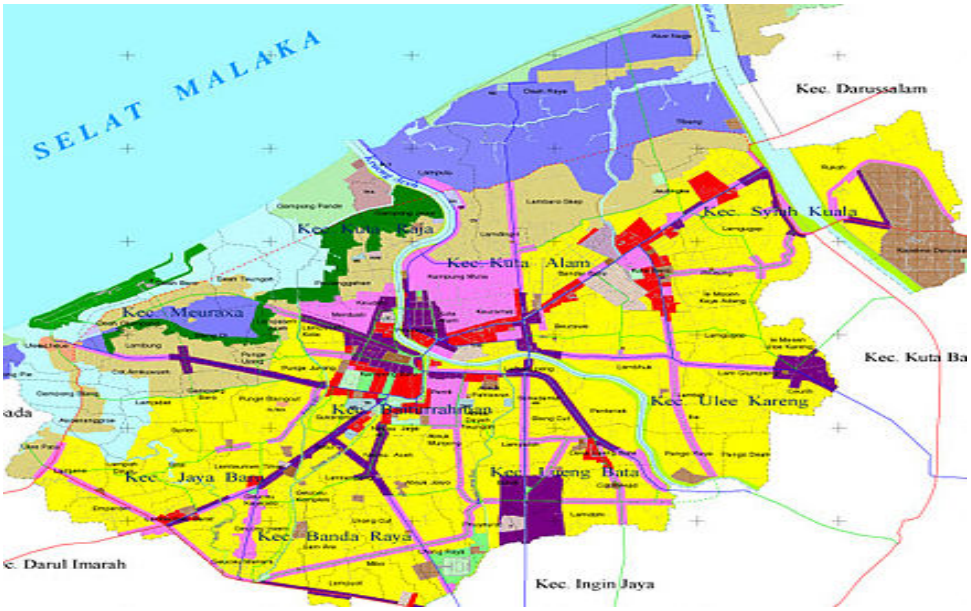

Figure 1. Map of Banda Aceh City

Figure 1 above is a map of the clustering of dengue fever in 9 sub-districts in Banda Aceh, consisting of Baiturrahman, Banda Raya, Jaya Baru, Kuta Alam, Kuta Raja, Lueng Bata, Meuraxa, Syiah Kuala, and Ulee Kareng districts.

b. Tools and Materials

Hardware specifications: AMD A9-9420 RADEON R5 3.00GHz processor, 4 GB memory (RAM), 1 TB hard drive. The required software is the Windows 10 Pro 64-bit operating system. Microsoft Office 2019 operating system, Matlab R2018, Rapidminer Studio Version 5.1.

c. The k-Means Clustering Algorithm

K-Means Clustering is one of the algorithms in data mining that can be used to cluster data. Several methods are using for data clusters. One of them is making rules whose members are collected based on the same group from the level of similarity among its members. Another approach can be made by creating a set of functions that can measure several properties of a grouping as a function of several clustering parameters. [7], [8]. K-Means Clustering Algorithm to

partition data, where each cluster center is represented according to the average value of the objects in the cluster. Generally, the k-means clustering algorithm can be summarized as follows [9], [10], [11].

1. Initialize: The data set in the first step determines the number of *clusters* and *centroids* for each cluster.
2. Classification: The distance is calculated for each data point from the *centroid*, and the data point, which has the minimum distance from the center of the cluster is assigned to a particular cluster. At this stage, it is necessary to calculate the distance of each data to each cluster center. To calculate the distance, a formula can be used, *Euclidean Distance Space*.

$$\mathrm{D} = \|X - Z\| = \sqrt{\sum_{i=1}^{n}(X - Z)^2} \quad (1)$$

   Information:

   Z = data object
   X = *centroid*/the closest average
   n = data dimension
   D = The distance is the smaller the distance D, then the more similar X and Z (D is the distance X and Z in n-dimensional space)

3. Recalculation Centroid: Previously generated *clusters*, the centroid is again counted repeatedly, which means recounting the centroid.
4. Convergence Conditions: Several conditions of convergence are given as follows:
   a. Stop when it reaches the specified or defined number of iterations.
   b. Stop when there is no data point exchange between clusters.
   c. Stop when the threshold value is reached.
5. If all of these conditions are not met, proceed to step 2, and the entire process is repeated until the state gives are met.

III. Results and Discussion

This study tries to classify the distribution pattern of dengue hemorrhagic fever in the city of Banda Aceh. The data used for eight periods from 2010 to 2017 consisted of climatic influence factors, namely temperature, temperature, and rainfall, obtained from the Meteorology and Geophysics Agency (BMKG) Indrapuri Aceh Besar. Population density data were obtained from the Central Bureau of Statistics (BPS) Banda Aceh, and data on the number of cases of Dengue Hemorrhagic Fever (DHF) were obtained from the Banda Aceh City Health Office.

a. Early Centroid

The process of finding the initial centroid value is done by calculating the largest (maximum) value for the high-level cluster (C1), the average value for a group of moderate levels (C2), and the smallest (minimum) value for the low-level group (C3). The early centroid point cluster center can be seen in table 1.

Table 1. Early Centroid Data

Sub-district	Population Density	Rainfall	Air Humidity	Temperature	DBD	Early Centroid
Kuta Alam	381.502	791.1	217.7	224	425	C1
Jaya Baru	192.126	972.7	216.7	224	233	C2
Kuta Raja	97.248	981.2	216.6	222	137	C3

Based on table 1, It can be explained that the determination of the centroid at the beginning of the first iteration is by finding the maximum value, average, and a minimum of all data from 2010 to 2017, so that the first cluster group is obtained, namely Kuta Alam sub-district (C1), Jaya Baru (C2), and Kuta raja (C3).

b. Input Data

The determination of the initial cluster in table 1 is used to carry out the next iteration process. The proximity of a location to the cluster is calculated by finding the distance between the data and the cluster center. As long as the iteration process, the results of the clusters are still moving. Iteration will continue so that the position of the clusters does not change. The data entered into the clusters are shown in table 2.

Tabel 2. Data

Sub-distric	Year: 2010-2017				
	KP	CH	KU	C	DBD
Meuraxa	147.098	1180.5	218.1	222	185
Jaya Baru	192.126	972.7	216.7	224	233
Banda Raya	180.899	963.3	216.4	225	238
Baiturrahman	272.477	878.6	218.8	222	419
Lueng Bata	197.457	843.5	218	224	307
Kuta Alam	381.502	791.1	217.7	224	425
Kuta Raja	97.248	981.2	216.6	222	137
Syiah Kuala	288.601	878.9	216	223	348
Ulee Kareng	197.285	952	217.6	227	237

The table above is the data input into k-mean clustering. Previously, the data had been carried out in the pre-post process by taking an average of each of the factors that could cause the outbreak of dengue cases in 9 districts in the city of Banda Aceh.

c. Rapidminer implementation

The process of grouping the affected areas of DHF cases in 9 sub-districts in the city of Banda Aceh starting from the structural design of the k-means clustering process, read xlsx formatted data, determined the number of clusters used in this study as many as k = 3, determine the number of iterations as many as 4, numeric measurement type, the distance between the data against the cluster using euclidean distance, the maximum optimization standard is 100. Cluster result data can be seen in Table 3.

Tabel 3. Cluster Results

Sub-distric	Cluster	KP	CH	KU	C	DBD
Meuraxa	Cluster_0	147.098	1180.5	218.1	222.0	185.0
Jaya Baru	Cluster_2	192.126	972.7	216.7	224.0	233.0
Banda Raya	Cluster_2	180.899	963.3	216.4	225.0	238.0
Baiturrahman	Cluster_1	272.477	878.6	218.8	222.0	419.0
Lueng Bata	Cluster_2	197.457	843.5	218.0	224.0	307.0
Kuta Alam	Cluster_1	381.502	791.1	217.7	224.0	425.0
Kuta Raja	Cluster_0	97.248	981.2	216.6	222.0	137.0
Syiah Kuala	Cluster_1	288.601	878.9	216.0	223.0	348.0
Ulee Kareng	Cluster_2	197.285	952.0	217.6	227.0	237.0

Table 3 shows that after four iterations of the classification process were carried out, the results showed that the cluster results' position did not change and was close to the initial centroid of the previously determined cluster. So that the results of the high category group cluster are obtained (C1) which is Baiturrahman, Kuta Alam dan Syiah Kuala district, then followed by the medium cluster group (C2) which is Jaya Baru, Banda Raya dan Ulee Kareng districts, then included in the low category (C3) which is Meuraxa dan Kuta Raja district.

Cluster 1 is a sub-district that is included in the high category in this case, which is the Baiturrahman, Kuta Alam dan Syiah Kuala district, judging by population density, rainfall, moisture, and the number of dengue cases in these areas is prone to dengue cases. From the entire cluster shown in graphical form, it can be used as anticipation for an outbreak of cases that can be seen in Figure 2.

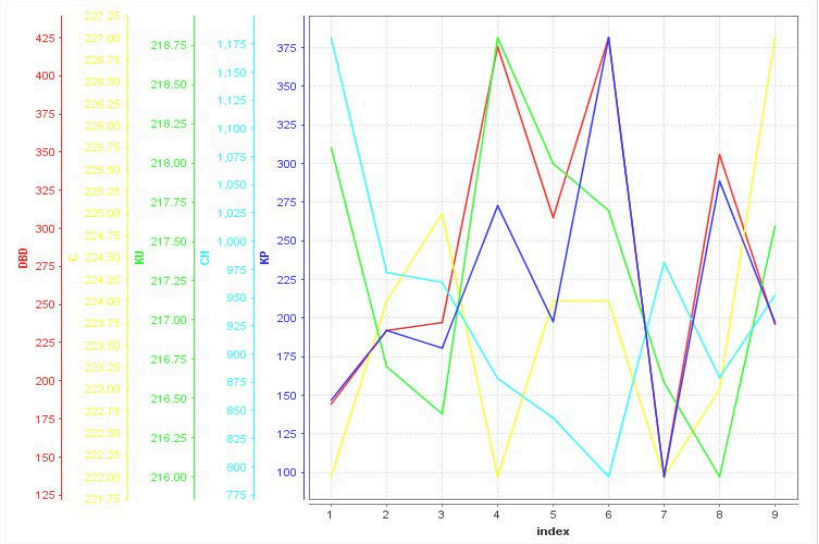

Figure 2. Graph of all factors on DHF cases

In Figure 2, it can be seen that from the climatic factor, The population density for the contagious of DHF cases in 9 sub-districts in Banda Aceh city can be seen from the direction of the graph that follows the pattern of DHF cases such as population density (blue line) and humidity (green line) which are very influential factors and can follow the case pattern. DHF during the period 2010 to 2017, especially in index 4, which is Baiturrahman sub-district, index 6 for Kuta Alam sub-district, and index 8 for Syiah Kuala sub-district.

IV. Conclusion

From the results of the research that has been done, the authors can draw some conclusions as follows:

1. K-means clustering can be used as an algorithm for data classification on cases of dengue cases involving several influencing factors.
2. Climate factors and population density to determine the distribution pattern of DHF cases in 9 sub-districts in the city of Banda Aceh can be clustered correctly for four iterations.
3. The high cluster group (C1) is Baiturrahman, Kuta Alam and Syiah Kuala sub-districts, the medium cluster group (C2) is Jaya Baru, Banda Raya, and Ulee Kareng districts. The low cluster group (C3) is Meuraxa and Kuta Raja.
4. Population density and humidity significantly affect dengue cases in several districts, namely Baiturrahman, Kuta Alam, and Syiah Kuala.

Some suggestions that can be given by the author are as follows:

1. The concerned party must always periodically evaluate the influence of climatic factors and population density because these factors can trigger dengue contagious in an area.

2. It can involve the presence of mosquito larvae in water reservoirs (TPA) in the house or around the house because of the existence of the vector larva Ae. Aegypti can trigger dengue fever cases.
3. Able to design Geographic Information System Web applications (WebGIS) by using the multilayer perceptron algorithm, Support Vector Machine, or comparing several other algorithms to classify cases of dengue contagious in a region.

Acknowledgment

We express our gratitude to the Directorate of Research and Community Service (DRPM) of the Ministry of Research and Technology of Higher Education that provided funding to this research through a program called " *Skim hibah penelitian dosen pemula tahun 2019 didanai tahun 2020*"

References

[1] Nugroho, GS., Nugroho, D., Hasbi, M. 2013. Geographic Information System Penyebaran DBD Berbasis Web di Wilayah Kota Solo, ISSN: 2338-4018.

[2] Wahyuningsih, NE., Suhartono., Sufia. 2014. Hubungan Kondisi Lingkungan Rumah dan Perilaku Keluarga dengan Kejadian Demam Berdarah *Dengue* Di Kabupaten Aceh Besar. Jurnal Kesehatan Lingkungan Indonesia Vol. 13 No. 1

[3] Nainggolan, F. 2007. Epidemiology and Clinical Pathogenesis of Dengue in Indonesia; presented at Seminar on Management of Dengue Outbreaks; University of Indonesia; Jakarta; November 22

[4] Depkes – Departemen Kesehatan, Republic of Indonesia. 2007. CDC and EH Yearly Report; Jakarta

[5] InfoDatin. 2018. Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia. April 22. ISSN. 2442-7659

[6] Elvin, SD., Mulyadi., Kamil, H. 2016. Tugas Kesehatan Keluarga Dalam Pencegahan Demam Berdarah Dengue Dengan Pendekatan Health Belief Model The Family Health Task In Prevention Of Dengue Hemorrhagic Fever With Health Belief Model Approach

[7] Neha., Chaudhary, N., Singh, T. 2014. A Comprehensive Review on k-Means Clustering Algorithm in Neural Networks. International Journal of electronIcs & communication technology (IJECT) Vol. 5, issue 3 spl - 1, ISSN : 2230-7109 (Online) | ISSN : 2230-9543 (Print).

[8] Kou, G., Peng, Y., Wang, G. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences. 275: p. 1-12.

[9] Bamer, M. 2007. Principles of Data Mining. ISSN 1863-7310. Springer-Verlag London Limited 2007.

[10] Kaur N, Sahiwal KJ, kaur N. 2012. Efficient K-Means Clustering Algorithm Using Ranking method in Data Mining. International Journal of Advanced Research in Computer Engineering & Technology. Volume 1, Issue 3. ISSN: 2278-1323

[11] Li, Y., Wu, H. 2012. A Clustering Method Based on K-Means Algorithm. International Conference on Solid State Devices and Materials Science. Sciverse ScienceDirect. Physics Procedia 25. 1104 – 1109.