Hybrid K Means-Multivariate Adaptive Regression Splines For Distribution Of Dengue Fever Risk Mapping In Bojonegoro District

by Alif Yuanita Kartini

Submission date: 02-Oct-2023 02:01AM (UTC+0700)
Submission ID: 2182181952
File name: ribution_Of_Dengue_Fever_Risk_Mapping_In_Bojonegoro_District.pdf (387.65K)
Word count: 6060
Character count: 30495

BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0313–0322
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0313-0322

HYBRID K MEANS-MULTIVARIATE ADAPTIVE REGRESSION SPLINES FOR DISTRIBUTION OF DENGUE FEVER RISK MAPPING IN BOJONEGORO DISTRICT

Alif Yuanita Kartini [1*], Nita Cahyani[2]

[1,2]*Department of Statistics, Faculty of Science and Technology, Universitas Nahdlatul Ulama Sunan Giri*
Jl. A. Yani No. 10, Bojonegoro, Jawa Timur, 62115, Indonesia

*Corresponding author's e-mail: * **alifyuanita@unugiri.ac.id***

Article History:

Received: 8th October 2022
Revised: 27th December 2022
Accepted: 2nd February 2023

Keywords:

Dengue Hemorrhagic Fever;
Clusterization;
Hybrid;
K-Means;
Multivariate Adaptive Regression Splines

ABSTRACT

Dengue Hemorrhagic Fever (DHF) is a dangerous disease transmitted by Aedes aegypti and Aedes albopictus mosquito bites. WHO data shows that almost half of the world's humans are exposed to Dengue Hemorrhagic Fever. The number of mortality caused by dengue disease is around 20,000 every year. In East Java, Bojonegoro District has the highest number of dengue hemorrhagic fever cases (416). To reduce this number, the causative factors need to be known. Additionally, it's important to pinpoint the region or cluster where the variables driving the spread are located so that prevention and treatment efforts are effective. Based on the elements contributing to the transmission of Dengue Hemorrhagic Fever, this study seeks to identify and categorize locations at risk for the spread of the illness. This study uses Hybrid K Means-Multivariate Adaptive Regression Splines (MARS), combining K-Means and MARS methods, to provide better analytical results. The data was divided into simpler parts by considering the Oakley distance. The results obtained from the K Means-MARS hybrid show the relationship between response variables and predictor variables for each cluster. There are three clusters at risk for the spread of dengue hemorrhagic fever in the Bojonegoro district with categories: high-risk cluster, medium-risk cluster, and low-risk cluster. The high-risk cluster consists of 7 sub-districts (Baureno, Kepohbaru, Balen, Sumberrejo, Kedungadem, Bojonegoro, and Dander). The variables affecting the DHF Sufferer in the high-risk cluster were population density (X_2), Altitude (X_3) and Health Worker (X_6). Meanwhile, the medium risk cluster consists of 10 sub-districts (Kalitidu, Kanor, Kapas, Ngasem, Ngraho, Padangan, Sugihwaras, Sukosewu, Tambakrejo, and Trucuk). The variables that affect the DHF Sufferer in the medium cluster are the Number of Dead (X_1), Population Density (X_2) and Health Facility (X_5). The low-risk cluster comprised 11 sub-districts (Bubulan, Gayam, Gondang, Kasiman, Kedewan, Malo, Margomulyo, Ngambon, Purwosari, Sek, and Temayang). The variables affecting the DHF Sufferer rate in the low-risk cluster were the number of dead (X_1) and population density (X_2).

How to cite this article:

A. Y. Kartini and N. Cahyani., "HYBRID K MEANS-MULTIVARIATE ADAPTIVE REGRESSION SPLINES FOR DISTRIBUTION OF DENGUE FEVER RISK MAPPING IN BOJONEGORO DISTRICT," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0313-0322, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article • Open Access

1. INTRODUCTION

Dengue Hemorrhagic Fever (DHF) is a dangerous disease that is still a significant concern related to public health in Indonesia. DHF is caused by the bite of the Aedes aegypti and Aedes albopictus mosquitoes which then spread the dengue virus. Indonesia, with its tropical climate, is the most potential area for the growth of the Aedes aegypti mosquito [1].

East Java is an area with high endemic potential year to year. Data issued by the Ministry of Health of the Republic of Indonesia in 2021 revealed that the highest dengue cases in Indonesia occurred in East Java. Data from the East Java Provincial Health Office shows that Bojonegoro District had the greatest number (416) of dengue cases in East Java in 2018 [2]. It is important to take dengue fever seriously. Therefore, for more effective prevention, a prediction and mapping of the DHF distribution region are required.

The spread of mosquitoes is influenced by meteorological variables such as rainfall, temperature, and humidity, all of which are highly sensitive to DHF [3]. At various periods and places, these climatic conditions affect the occurrence of dengue hemorrhagic fever [4]. For modeling, predicting, and mitigating DHF risk, the majority of research used Bayesian techniques like conditional autoregressive [5], Poisson [6][7], and negative binomial [6], which could only be used to study the interaction of two risk variables that cause dengue.

Other artificial intelligence models such as Analysis Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS), Least Square Support Vector Machine (LS-SVM), Gene Expression Programming (GEP), Group Method of Data Handling (GMDH), Dynamic Evolving Neural-fuzzy Inference System (DENFIS), and M5 Tree have been widely used as effective tools for modeling complex nonlinear systems and predicting different variables. Several studies have shown that the MARS model has a better level of accuracy and prediction when compared to other artificial intelligence models [8][9][10][11]. The MARS model has become popular because it does not require assumptions and does not require special relationships such as linear, quadratic, and cubic relationships between the response variables and predictor variables [12]. MARS is a form of multivariate regression with a nonparametric approach developed by Friedman in 1991. MARS model is used on data that has high dimensions, large number of variables, and data with large sample sizes. MARS is the development of the Recursive Partition Regression (RPR) model, where the model still has a weakness because the model obtained is not continuous at knots.

In addition to modeling and prediction, research on regional clustering for cases of dengue fever is also often carried out. Clustering is a method that organizes data with many attributes into clusters with similar behavior. The most frequently used clustering method is the K-Means method because of its convenience and efficiency. The K-Means method is often used to determine clusters of tuberculosis [13], dengue fever [14][15], and schistosomiasis [16]. For cases of Dengue Hemorrhagic Fever, including mapping the area of DHF distribution based on the number of sufferers [17], mapping the spread of Aedes aegypti mosquito larvae [18], and mapping the area of DHF distribution based on geography and demography [19]. However, apart from these risk factors, DHF is also influenced by other risk factors, including climate [20][21] [4], socio-economic [20][22], environment [22], and demographics [22].

In this study, modeling and mapping the area of risk factors for the spread of dengue cases, especially in Bojonegoro district, was carried out using the Hybrid K Means-MARS method, a combination of the K Means and MARS methods. The use of this method is expected to provide better analytical results. With this method, the data is divided into simpler parts by considering the Oakley distance. The results obtained from the K Means-MARS hybrid will make it easier to show the relationship between response variables and predictor variables for each cluster. This study aimed to obtain the area of risk factors mapping for the spread of dengue cases in Bojonegoro district along with the variables thought to affect the DHF Sufferer for each region. This research is expected to help the Bojonegoro District Health Office in handling cases of DHF more effectively and efficiently.

2. RESEARCH METHODS

2.1 Data Source

Secondary data was taken from several related agencies, including the Bojonegoro District Health Office, the Bojonegoro District Central Bureau of Statistics, and the Meteorology, Climatology, and Geophysics Agency of the Juanda Sidoarjo Meteorological Station. The data from 2021 was used with an area of 28 sub-districts (Balen, Baureno, Bubulan, Bojonegoro, Kalitidu, Gondang, Dander, Kanor, Kapas, Kasiman, Kedewan, Kedungadem, Kepohbaru, Malo, Margomulyo, Ngambon, Ngasem, Ngraho, Padangan sub-districts Purwosari, Temayang, Trucuk, Tambakrejo, Gayam, Sumberrejo, Sukosewu, Sugihwaras, and Sekar) in Bojonegoro district.

2.2 Research Variables

The variables in this are response variables (Y) which included DHF Sufferer with predictor variables number of dead (X_1), population density (X_2), altitude (X_3), rainfall (X_4), health facility (X_5), health workers (X_6) with all ratio/interval scale variables.

2.3 Research Procedure

The procedure carried out in this study is as follows.

1) Processing data.
2) Determining the value of descriptive statistics for each variable.
3) Determining the cluster with the K-Means algorithm.
   a) Determining the number of clusters (3 clusters according to the DHF distribution area) including high clusters, medium clusters and low clusters [23].
   b) Determining the initial centroid value randomly, then calculate the cluster center value (average) from each cluster using the equation

$$v_{ij} = \frac{1}{N}\sum_{k=0}^{N_i} x_{kj} \quad (1)$$

   Where v_{ij} is the average of the i-th cluster and the j-th variable, N_i is the number of data for the i-cluster, i and k are the indexes on the cluster, j is the index on the variable and x_{kj} is the value for the k-th data in the k-cluster j [21][24].
   c) Calculating the Euclidean distance (distance between the centroid point and the point of each object) with the formula

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2)$$

   With D_e is Euclidean distance, i is the number of objects, (x, y) are the coordinates of the object and (s, t) is the centroid of the object [25].
   d) Grouping objects to determine the number of clusters with the shortest distance between objects [26].
   e) Counting data with the new cluster center. After getting a new center point from each cluster, the iteration is carried out until a fixed centroid value is produced and no member of the cluster moves to another cluster [25][27].
   f) Obtaining the results of the cluster and the members of each cluster.
4) Performing MARS modeling for each cluster.
   a) Determining the main components in the MARS model which included the Basis of Function (BF), knot points and spline functions. The basis of the function is a function that describes the relationship between the response variable and the predictor variable. The basis of the function used has a polynomial form that is continuous at every knot point [28]. A knot point is a point that describes the end of a regression line and the beginning of another regression line. At each knot point, there is continuity between regions with a functional basis [29].
   b) Obtaining the MARS model with the following formula

$$Y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m BF_m(X) \quad (3)$$

Where Y is the response variable, β_0 is the constant or parent of the base function, β_m is the coefficient of the basis of the m-th function and $BF_m(X)$ is the basis of the m-th function whose value is $\max(0, c - x)$ or $\max(0, x - c)$, where x is the predictor variable and c is the knot of the response variable [28][29].

c) Obtaining the best MARS model through the forward stepwise and backward stepwise stages. Forward stepwise is used to obtain the number of basis functions and backward stepwise is used to obtain a simple model by eliminating the basis functions that have a small contribution to the response variable [11]. The MARS model chosen is the one that has a minimum GCV with the following formula [28][29].

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}_M(x_i)\right)^2}{\left(1 - \frac{C(M)}{N}\right)^2} \quad (4)$$

$$C(M) = (d + 1) \times M \quad (5)$$

Where N is the number of observations, y_i is the response variable, $\hat{f}_M(x_i)$ is the estimated value on the basis of the function, $C(M)$ is the complexity cost function and d is the value when each basis function reaches optimization$(2 \leq d \leq 4)$.

d) Obtaining the variables that contribute to the MARS model for each cluster through the forward stage, namely by first selecting the variable that has the greatest GCV value.

5) Obtaining a map of the distribution area of Dengue Hemorrhagic Fever in Bojonegoro District with variables that affect each region in Bojonegoro District.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

Descriptive statistics is a summary of the data used to obtain the characteristics of the research variables. The results of descriptive statistics are shown in Table 1.

Table 1. Descriptive Statistics of Research Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
DHF Sufferer	1	54	12.14	11.96
Number of Dead	0	2	0.18	0.48
Population Density	11879	87563	47902.11	22501.97
Altitude	9.38	506.25	105.51	120.31
Rainfall	600	1880	1085	316.64
Health Facility	25	145	76.50	32.18
Health Worker	18	92	47.21	20.83

Table 1 shows that the lowest number of DHF Sufferers was 1, found in the Gondang, Margomulyo, Ngambon, Ngraho and Sekar sub-districts, while the highest number of DHF Sufferers was in the Sumberrejo sub-district (54). The average number of DHF Sufferers is 12.14, with a standard deviation of 11.96. The number of dead in Bojonegoro district is low and the highest number of dead is 2 in Temayang sub-district. The average number of dead is 0.18 with a standard deviation of 0.48. The average altitude in Bojonegoro district is 105.51 masl and the standard deviation was 120.31. The Gondang sub-district has the highest altitude (506.25 masl), while Kanor sub-district has the lowest altitude (9.38 masl). Gondang sub-district in the Bojonegoro district received the most annual rainfall (1880 mm), while Kalitidu sub-district receives the least (600 mm). In the Bojonegoro sub-district, annual precipitation averages 1085 mm with a standard deviation of 316.64 millimeters. With an average of 76.50 health facilities and a standard deviation of 32.18, the greatest number of health facilities is 145 in the Kepohbaru sub-district and at least 25 in the Ngambon sub-district. In Kedungadem sub-district, there are 92 health staff, while Kedewan sub-district has only 18. In the Bojonegoro district, there are 47.21 health professionals on average, with a standard deviation of 20.83.

3.2 Clustering the Risk of Dengue Hemorrhagic Fever Spread Using K-Means

One of the data mining algorithms used to group or cluster data is K-Means. Data mining techniques include clustering, which involves creating a rule that groups clusters according to their constituent parts' similarity. Additionally, clustering can be carried out by developing a collection of functions to assess the type of clustering for a number of clustering factors [21]. The number of dead, population density, altitude, rainfall, health facility, and health worker variables are used with the K-Means approach to group the risk locations in the Bojonegoro district for spreading dengue hemorrhagic fever. The number of dead, population density, altitude, rainfall, health facility, and health worker variables are used in conjunction with the K-Means approach to group the risk locations in the Bojonegoro district for the spread of dengue hemorrhagic fever. By dividing the cluster of dengue hemorrhagic fever distribution in the Bojonegoro area into three, the K-Means algorithm's first stage in clustering is to establish the number of clusters. The following stage entails computing the centroid value, Euclidean distance, object grouping, and data calculation using the new centroid. Repeat this process until neither the centroid value nor the cluster members move to a different cluster. Members are obtained for each cluster based on these stages, as demonstrated in Table 2.

Table 2. Members of Each Cluster

Cluster	Cluster Members
Cluster_0	Kalitidu, Kanor, Kapas, Ngasem, Ngraho, Padangan, Sugihwaras, Sukosewu, Tambakrejo, Trucuk
Cluster_1	Balen, Boureno, Bojonegoro, Dander, Kedungadem, Kepohbaru, Sumberrejo
Cluster_2	Bubulan, Gayam, Gondang, Kasiman, Kedewan, Malo, Margomulyo, Ngambon, Purwosari, Sekar, Temayang

Each cluster's members are listed in Table 2, with Cluster_0 having 10 members, Cluster_1 having 7 members, and Cluster_2 having 11 members. Additionally, the plot shown in Figure 1 below can be used to gauge the degree of risk associated with the spread of dengue hemorrhagic fever from the cluster that has formed.

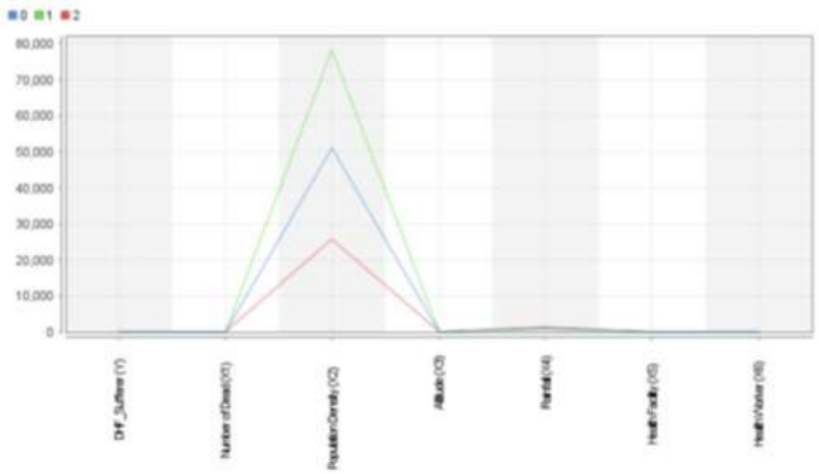

Figure 1. Visualization of the Spread of Dengue Hemorrhagic Fever

According to Figure 1 1's line plot, Cluster_1 has the highest line position and a high risk of transmitting dengue hemorrhagic fever. As a result, Cluster_1 is a cluster that poses a significant danger. In the meantime, Cluster_0's position in the middle of the line indicates that it has a moderate chance of spreading dengue disease. Thus, Cluster_0 is a cluster with a medium level of risk. Cluster_2 is less likely to spread because it has the lowest line position. Thus, Cluster 2 is a low-risk cluster.

According to Table 2, the subdistricts of Balen, Boureno, Bojonegoro, Dander, Kedungadem, Kepohbaru, and Sumberrejo are clusters with a high risk. The Kalitidu, Kanor, Kapas, Ngasem, Ngraho, Padangan, Sugihwaras, Sukosewu, Tambakrejo, and Trucuk subdistricts are among the clusters having a medium risk. The sub-districts of Bubulan, Gayam, Gondang, Kasiman, Kedewan, Malo, Margomulyo, Ngambon, Purwosari, Sekar, and Temayang are included in the cluster with low risk. Table 3 below lists the features of each cluster that carries a high risk, a medium risk, and a low risk.

Table 3. Characteristics of Each Cluster

Variabels	Low Risk Cluster	Medium Risk Cluster	High Risk Cluster
Number of Dead	0.273	0.100	0.143
Population Density	25843.091	50922.100	78252.000
Altitude	180.966	63.438	47.054
Rainfall	1284.545	970.000	935.714
Health Facility	46.182	80.400	118.571
Health Worker	29.273	49.600	72.000

Table 3 shows that clusters with high risk have a moderate average number of dead of 0.143, a high average population density of 78252, an average low altitude of 47.054, a low average rainfall of 935.714, an average high average health facility is 118.571 and a high average health worker is 72. Meanwhile, clusters with medium risk have a low average number of dead of 0.1, a moderate average population density of 50922.1, an average moderate altitude is 63.438, an average moderate rainfall is 970, an average health facility is 80.4 and an average health worker is 49.6. And for low-risk clusters, a high average number of dead of 0.273, a low average population density of 25843.091, an average high altitude of 180.966, a high average rainfall of 1284.545, the average health facility low at 46.182 and an average health worker is low at 29.273.

3.3 MARS Modeling in Each Cluster Formed

DHF Sufferer (Y), Number of Dead (X_1), Population Density (X_2), Altitude (X_3), Rainfall (X_4), Health Facility (X_5), and Health Worker (X_6) are the variables utilized in MARS modeling. The next step is to do MARS modeling in each cluster after getting the findings of clustering the risk areas for the spread of DHF in the Bojonegoro district. Finding the most basis functions (BF), most interactions (MI), and least observations between knots (MO) is the first step in MARS modeling [11][29]. The advised BF ranges from 2 to 4 times the number of predictor variables, which are 12, 18, and 24. In the meantime, there are 1, 2, and 3 MI and 0, 1, 2, and 3 MO [13][14]. BF, MI, and MO were combined to create the MARS model through trial and error. Table 4 displays the outcomes of the Low Risk Cluster's MARS modeling experiment using the BF, MI, and MO combination.

Table 4. Trial and Error Results for MARS Modeling in Low Risk Clusters

No	BF	MI	MO	GCV	R^2	No	BF	MI	MO	GCV	R^2
1	12	1	0	5.39954	0.914	7	18	2	2	2.18791	0.941
2	**12**	**1**	**1**	**1.14752**	**0.989**	8	18	2	3	3.76800	0.913
3	12	1	2	1.92297	0.941	9	24	3	0	6.28091	0.743
4	12	1	3	5.39954	0.914	10	24	3	1	2.18791	0.941
5	18	2	0	6.28091	0.743	11	24	3	2	2.18791	0.941
6	18	2	1	2.18791	0.941	12	24	3	3	3.76800	0.913

Based on the results of Trial and Error as shown in Table 4, it can be seen that the best MARS model is the model with a combination of BF=12, MI=1, and MO=1, with a GCV value of 1.14752 and an R^2 value of 0.989. So the MARS model for low risk clusters is as follows.

$$Y = 1.33334 + 0.00213185BF_1 + 2.45703BF_3 + 0.00355197BF_4 + 0.00214463BF_6$$

Where:

$BF_1 = \max(0, X_2 - 28544)$

$BF_3 = \max(0, X_1 - 0)$

$BF_4 = \max(0, X_2 - 30733)$

$BF_6 = \max(0, X_2 - 31898)$

According to the MARS model's interpretation, the DHF Sufferer number in the low risk cluster is 1.33334 if no affecting variables exist. Additionally, if the population density is greater than 28544 individuals, increasing one unit in basis function 1 will only result in a 0.00213185 rise in the DHF Sufferer number. Only if a significant number of deceased will an increase of one unit in basis function 3 result in an

increase of 2.45703 in the DHF Sufferer number. Only if the population density is more than 30733 people would the DHF Sufferer number increase by 0.00355197 with an increase of one unit of basis function 4.

Additionally, the MARS model revealed that Population Density (X2) and Number of Dead (X1) are the factors that influence the number of DHF Sufferers in the low risk cluster. Table 5 below illustrates the size of these factors' contributions.

Table 5. Low-Risk Clusters' Contributions and Influencing Factors

Variable	Amount of Contribution (%)
Population Density (X_2)	100
Number of Dead (X_1)	73.96694

Meanwhile, the results of Trial and Error for MARS modeling in medium risk clusters are shown in Table 6 below.

Table 6. Trial and Error Results for MARS Modeling in Medium Risk Cluster

No	BF	MI	MO	GCV	R^2	No	BF	MI	MO	GCV	R^2
1	12	1	0	30.56309	0.832	7	18	2	2	41.41562	0.678
2	**12**	**1**	**1**	**8.37638**	**0.997**	8	18	2	3	44.93882	0.678
3	12	1	2	24.71784	0.897	9	24	3	0	44.93882	0.678
4	12	1	3	30.56309	0.832	10	24	3	1	21.93008	0.958
5	18	2	0	44.93882	0.678	11	24	3	2	41.41562	0.678
6	18	2	1	21.93008	0.958	12	24	3	3	44.93882	0.678

According to Table 6 6's Trial and Error findings for modeling MARS in the medium risk cluster, the model with the combination of BF=12, MI=1, and MO=1 with a GCV value of 8.37638 and R^2 of 0.997 is the best MARS model. As a result, the MARS model for clusters at medium risk looks like this.

$$Y = 13.8199 - 12.221BF_1 - 1.27199BF_2 + 0.886023BF_3 + 0.00329648BF_4 - 0.00137998BF_5 + 0.00648552BF_6 - 1.7748BF_8$$

With:

$BF_1 = \max(0, X_1 - 1.49012e^{-8})$
$BF_2 = \max(0, X_5 - 71)$
$BF_3 = \max(0, 71 - X_5)$
$BF_4 = \max(0, X_2 - 51371)$
$BF_5 = \max(0, 51371 - X_2)$
$BF_6 = \max(0, X_2 - 56093)$
$BF_8 = \max(0, X_5 - 87)$

The MARS model can be interpreted if there are no influencing variables, then the number of DHF sufferers is 13.8199. Next, if there is an increase of one unit of basis function 1, it will decrease the DHF Sufferer number by 12.221 only if there is no number of dead. If there is an increase of one unit of basis function 2, it will decrease the DHF Sufferer number by 1.27199 only if the number of health facilities is more than 71. And if there is an increase of one unit of basis function 3 will increase the DHF Sufferer number by 0.886023 only if the number of health facility is less than 71. If there is an increase of one unit of basis function 4, the DHF Sufferer number will increase by 0.00329648 only if the population density is more than 51371 people. And if there is an increase of one unit of basis function 5, the DHF Sufferer number will decrease by 0.00137998 only if the population density is less than 51371 people. Meanwhile, if there is an increase of one basis function unit 6, it will increase the DHF Sufferer number by 0.00648552 only if the population density is more than 56093 people. And if there is an increase of one basis function unit 8, it will reduce the DHF Sufferer number by 1.7748 only if the number of health facilities is more than 87.

In this medium cluster, the variables that affect the number of DHF Sufferers are the Number of Dead (X_1), Population Density (X_2), and Health Facility (X_5). The amount of the contribution of each variable is shown in Table 7 below.

Table 7. Influential Variables and Contributions to Clusters with Medium Risk

Variable	Amount of Contribution (%)
Number of Dead (X_1)	100
Population Density (X_2)	48.97673
Health Facility (X_5)	37.80817

For the results of Trial and Error on MARS modeling in the High Risk Cluster as shown in Table 8 below.

Table 8. Trial and Error Results for MARS Modeling in High Risk Clusters

No	BF	MI	MO	GCV	R^2	No	BF	MI	MO	GCV	R^2
1	12	1	0	196.88072	0.716	7	18	2	2	159.28701	0.687
2	**12**	**1**	**1**	**4.86832**	**0.976**	8	18	2	3	159.28701	0.687
3	12	1	2	196.88072	0.716	9	24	3	0	159.28701	0.687
4	12	1	3	196.88072	0.716	10	24	3	1	8.03326	0.961
5	18	2	0	159.28701	0.687	11	24	3	2	144.93967	0.715
6	18	2	1	8.03326	0.961	12	24	3	3	144.93967	0.715

The model with the combination of BF=12, MI=1 and MO=1 with a GCV value of 4.86832 and an R^2 value of 0.976 is the best MARS model for high risk clusters based on Table 8. With BF = 12, MI = 1, and MO = 1, the optimal MARS model is as follows.

$$Y = 33.2654 + 0.00934912BF_1 + 0.000220473BF_2 + 0.00666449BF_3 + 0.0710497BF_5 + 1.48931BF_6 - 0.124065BF_7 + 2.27082BF_8$$

where :

$BF_1 = \max(0, X_2 - 72251)$
$BF_2 = \max(0, 72251 - X_2)$
$BF_3 = \max(0, X_2 - 67390)$
$BF_5 = \max(0, X_3 - 28.125)$
$BF_6 = \max(0, 28.125 - X_3)$
$BF_7 = \max(0, X_6 - 48)$
$BF_8 = \max(0, 48 - X_6)$

From the MARS model, it can be interpreted that if there are no influential variables, the DHF Sufferer number in the high risk cluster is 33.2654. Furthermore, if there is an increase of one unit of basis function 1, the DHF Sufferer number will increase by 0.00934912 only if the total population density is more than 72251 people. If there is an increase of one unit of basis function 2, it will increase the incidence of DHF by 0.000220473 only if the population density is less than 72251 people. If there is an increase of one unit of basis function 3, the DHF Sufferer number will increase by 0.00666449 only if the population density is more than 67390 people. If there is an increase of one basis function unit 5, the DHF Sufferer number will increase by 0.0710497 only if the altitude is more than 28.125 masl. And if there is an increase of one basis function unit 6, the DHF Sufferer number will increase by 1.48931 only if the altitude is less than 28.125 masl. An increase of one unit base function 7 will reduce the number of DHF Sufferers only if the health worker is more than 48 people. An increase of one base function unit 8 will increase the DHF Sufferer number by 2.27082 only if the health worker is lower than 48 people.

In the high risk cluster, the variables that affected the DHF Sufferer were Population Density (X_2), Altitude (X_3), and Health Worker (X_6). The amount of the contribution of each of these variables is shown in Table 9.

Table 9. Influential Variables and Contributions to High-Risk Clusters

Variable	Amount of Contribution (%)
Population Density (X_2)	100
Altitude (X_3)	51.28326
Health Worker (X_6)	36.63727

Below are the results of risk areas mapping for the spread of dengue hemorrhagic fever in Bojonegoro district along with the variables that affect the DHF Sufferer for each region as shown in Figure 1.

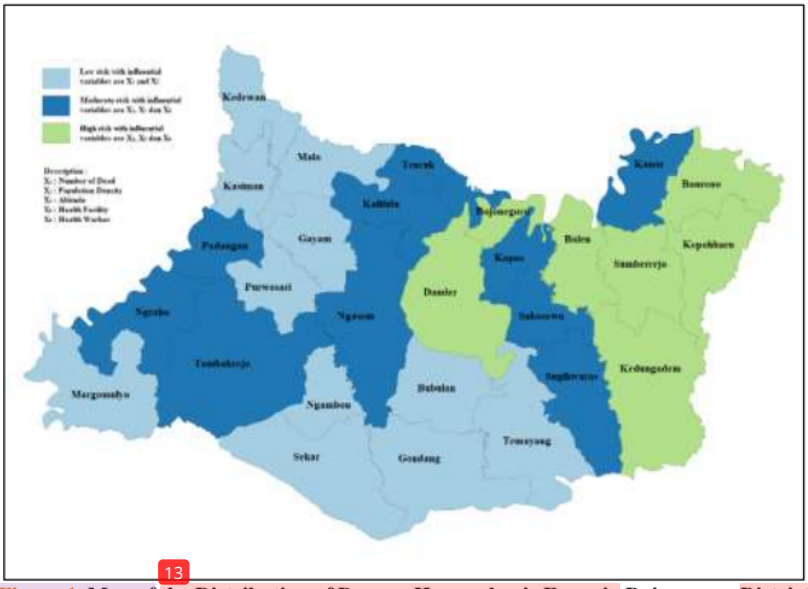

Figure 1. Map of the Distribution of Dengue Hemorrhagic Fever in Bojonegoro District

4. CONCLUSIONS

According to this study's findings, 3 clusters of risk high, medium and low were discovered in the municipality of Bojonegoro. Cluster with high risk included seven sub-districts (Baureno, Kepohbaru, Balen, Sumberrejo, Kedungadem, Bojonegoro, and Dander). The three variables that affected DHF sufferers in high risk clusters were population density (X_2), altitude (X_3), and health worker (X_6). In addition, the cluster with medium risk is currently made up of 10 sub-districts (Kalitidu, Kanor, Kapas, Ngasem, Ngraho, Padangan, Sugihwaras, Sukosewu, Tambakrejo, and Trucuk). The number of dead (X_1), Population Density (X_2), and Health Facility (X_5) are the three variables that most significantly affected DHF Sufferers in the current cluster. There were 11 sub-districts in the low-risk cluster (Bubulan, Gayam, Gondang, Kasiman, Kedewan, Malo, Margomulyo, Ngambon, Purwosari, Sekar, and Temayang). Number of Dead (X_1) and population density (X_2) were the factors that influence the DHF Sufferer rate in the low risk cluster.

REFERENCES

[1] M. U. G. Kraemer *et al*., "The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus," *Elife*, vol. 4, p. e08347, 2015, doi: 10.7554/eLife.08347.

[2] Pusat Data dan Informasi Kementerian Kesehatan RI, "Situasi Demam Berdarah Dengue," *InfoDATIN*. 2018. [Online]. Available: https://pusdatin.kemkes.go.id/

[3] I. G. N. M. Jaya and H. Folmer, "Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia," *J. Geogr. Syst*., vol. 22, no. 1, pp. 105–142, 2020, doi: 10.1007/s10109-019-00311-4.

[4] L. Xu *et al*., "Climate variation drives dengue dynamics," *Proc. Natl. Acad. Sci*., vol. 114, no. 1, pp. 113–118, 2017, doi: 10.1073/pnas.1618558114.

[5] L.-C. Chien and H.-L. Yu, "Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence," *Environ. Int*., vol. 73, pp. 46–56, 2014, doi: https://doi.org/10.1016/j.envint.2014.06.018.

[6] S. A. Thamrin, Aswi, Ansariadi, A. K. Jaya, and K. Mengersen, "Bayesian spatial survival modelling for dengue fever in Makassar, Indonesia," *Gac. Sanit*., vol. 35, pp. S59–S63, 2021, doi: https://doi.org/10.1016/j.gaceta.2020.12.017.

[7] F. Kristiani, Y. Claudia, B. Yong, and A.-M. Hilsdon, "A comparative analysis of frequentist and Bayesian approaches to estimate dengue disease transmission in Bandung-Indonesia," *J. Stat. Manag. Syst*., vol. 23, no. 8, pp. 1543–1559, Nov. 2020, doi: 10.1080/09720510.2020.1756049.

[8] W. Zhang and A. T. C. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile drivability," *Geosci. Front*., vol. 7, no. 1, pp. 45–52, 2016.

[9] R. M. Adnan, Z. Liang, S. Heddam, M. Zounemat-Kermani, O. Kisi, and B. Li, "Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs," *J. Hydrol*., vol. 586, p. 124371, 2020, doi: https://doi.org/10.1016/j.jhydrol.2019.124371.

[10] O. Kisi, P. Khosravinia, M. R. Nikpour, and H. Sanikhani, "Hydrodynamics of river-channel confluence: toward modeling

separation zone using GEP, MARS, M5 Tree and DENFIS techniques," *Stoch. Environ. Res. Risk Assess.*, vol. 33, no. 4, pp. 1089–1107, 2019, doi: 10.1007/s00477-019-01684-0.

[11] A. Ghasemzadeh and M. M. Ahmed, "Utilizing naturalistic driving data for in-depth analysis of driver lane-keeping behavior in rain: Non-parametric MARS and parametric logistic regression modeling approaches," *Transp. Res. Part C Emerg. Technol.*, vol. 90, pp. 379–392, 2018, doi: https://doi.org/10.1016/j.trc.2018.03.018.

[12] X. Ju, J. M. Rosenberger, V. C. P. Chen, and F. Liu, "Global optimization on non-convex two-way interaction truncated linear multivariate adaptive regression splines using mixed integer quadratic programming," *Inf. Sci. (Ny).*, vol. 597, pp. 38–52, 2022.

[13] R. S. Wardani, Purwanto, Sayono, and A. Paramananda, "Clustering tuberculosis in children using K-Means based on geographic information system," *AIP Conf. Proc.*, vol. 2114, no. June, 2019, doi: 10.1063/1.5112483.

[14] K. S. Ahmed Bin and S. Kamran Jabbar, "Dengue Fever in Perspective of Clustering Algorithms," *J. Data Mining Genomics Proteomics*, vol. 06, no. 03, 2015, doi: 10.4172/2153-0602.1000176.

[15] N. Mathur, V. S. Asirvadam, S. C. Dass, and B. S. Gill, "Visualization of dengue incidences for vulnerability using K-means," in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2015, pp. 569–573. doi: 10.1109/ICSIPA.2015.7412255.

[16] S. Li *et al.*, "The spatial-temporal trend analysis of schistosomiasis from 1997 to 2010 in Anhui Province, Eastern China," *Am. J. Trop. Med. Hyg.*, vol. 98, no. 4, pp. 1145–1151, 2018, doi: 10.4269/ajtmh.17-0475.

[17] M. A. Sembiring, "Penerapan Metode Algoritma K-Means Clustering Untuk Pemetaan Penyebaran Penyakit Demam Berdarah Dengue (Dbd)," *J. Sci. Soc. Res.*, vol. 4, no. 3, p. 336, 2021, doi: 10.54314/jssr.v4i3.712.

[18] E. A. Pratama and C. M. Hellyana, "Penerapan Algoritma K-Means Untuk Clustering Data Tempat Penyebaran Jentik Nyamuk Aedes Aegepty Pada Kelurahan Sumampir, Banyumas," *MEANS (Media Inf. Anal. dan Sist.*, vol. 6, no. 1, pp. 13–18, 2021, doi: 10.54367/means.v6i1.1169.

[19] A. Fariza, Mu'Arifin, and D. W. Astuti, "Spatial-Temporal Visualization of Dengue Haemorrhagic Fever Vulnerability in Kediri District, Indonesia, Using K-means Algorithm," no. November 2021, pp. 1–6, 2021, doi: 10.1109/icodse53690.2021.9648487.

[20] P. W. Dhewantara *et al.*, "Spatial and temporal variation of dengue incidence in the island of Bali, Indonesia: An ecological study," *Travel Med. Infect. Dis.*, vol. 32, p. 101437, 2019, doi: https://doi.org/10.1016/j.tmaid.2019.06.008.

[21] Sanusi and J. Husna, "Utilization of Rapidminer using the K-Means Clustering Algorithm for Classification of Dengue Hemorrhagic Fever (DHF) Spread in Banda Aceh City," *J. Inotera*, vol. 5, no. 2 SE-Articles, pp. 146–151, Oct. 2020, doi: 10.31572/inotera.Vol5.Iss2.2020.ID119.

[22] S. P. M. Wijayanti *et al.*, "The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia," *PLoS Negl. Trop. Dis.*, vol. 10, no. 9, p. e0004964, Sep. 2016, [Online]. Available: https://doi.org/10.1371/journal.pntd.0004964

[23] A. I. Widyatami and D. A. Suryawan, "Pengelompokan Daerah Rawan Demam Berdarah Dengue di Provinsi DKI Jakarta," *Indones. Heal. Inf. Manag. J.*, vol. 9, no. 1, pp. 73–82, 2021.

[24] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022.

[25] S. Rahayu and A. Y. Kartini, "ALGORITMA K-MEANS DAN K-MEDOIDS UNTUK PENGELOMPOKAN KECAMATAN PENERIMA BANTUAN SOSIAL DI KABUPATEN BOJONEGORO," *MEDIA BINA Ilm.*, vol. 16, no. 5, pp. 6815–6822, 2021.

[26] A. Bigdeli, A. Maghsoudi, and R. Ghezelbash, "Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran," *J. Geochemical Explor.*, vol. 233, p. 106923, 2022.

[27] C. Barile, C. Casavola, G. Pappalettera, and V. P. Kannan, "Laplacian score and K-means data clustering for damage characterization of adhesively bonded CFRP composites by means of acoustic emission technique," *Appl. Acoust.*, vol. 185, p. 108425, 2022.

[28] G.-W. Weber, İ. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya-Özkurt, "CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization," *Inverse Probl. Sci. Eng.*, vol. 20, no. 3, pp. 371–400, Apr. 2012, doi: 10.1080/17415977.2011.624770.

[29] A. Y. K. Kartini and L. N. Ummah, "Pemodelan Kejadian Balita Stunting di Kabupaten Bojonegoro dengan Metode Geographically Weighted Regression dan Multivariate Adaptive Regression Splines," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 1, 2022.

Hybrid K Means-Multivariate Adaptive Regression Splines For Distribution Of Dengue Fever Risk Mapping In Bojonegoro District

ORIGINALITY REPORT

9%	4%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Sitti Wetenriajeng Sidehabi, Ansar Suyuti, Intan Sari Areni, Ingrid Nurtanio. "Classification on passion fruit's ripeness using K-means clustering and artificial neural network", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018 Publication	1%
2	Nur Mahmudah, Fetrika Anggraini. "ON COMPUTATIONAL BAYESIAN ORDINAL LOGISTIC REGRESSION LINK FUNCTION IN CASES OF CERVICAL CANCER IN TUBAN", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2022 Publication	1%
3	Marta Sundari, Khairil Anwar Notodiputro, Bagus Sartono. "Modeling the influence of climatic factors on the number of dengue hemorrahagic fever (DHF) patients in DKI	1%

Jakarta 2017-2020 using generalized linear mixed model", AIP Publishing, 2023
Publication

---

4 Rana Muhammad Adnan, Payam Khosravinia, Bakhtiar Karimi, Ozgur Kisi. "Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline", Applied Soft Computing, 2020 — 1%
Publication

---

5 Rizki Fitri Ananda, Lisa Harsyiah, Muhammad Rijal Alfian. "Classification Of Perceptions Of The Covid-19 Vaccine Using Multivariate Adaptive Regression Spline", Jurnal Varian, 2023 — 1%
Publication

---

6 Sorour Alotaibi, Mohammad Ali Amooie, Mohammad Hossein Ahmadi, Narjes Nabipour, Kwok-wing Chau. "Modeling thermal conductivity of ethylene glycol-based nanofluids using multivariate adaptive regression splines and group method of data handling artificial neural network", Engineering Applications of Computational Fluid Mechanics, 2020 — 1%
Publication

---

7 Riry Sriningsih, Bambang Widjanarko Otok, Sutikno. "Determination of the best — 1%

multivariate adaptive geographically weighted generalized Poisson regression splines model employing generalized cross-validation in dengue fever cases", MethodsX, 2023
Publication

8	repository.unugiri.ac.id Internet Source	1 %
9	Septia Devi Prihastuti Yasmirullah, Bambang Widjanarko Otok, Jerry Dwi Trijoyo Purnomo, Dedy Dwi Prastyo. "Modification of Multivariate Adaptive Regression Spline (MARS)", Journal of Physics: Conference Series, 2021 Publication	1 %
10	archive.lstmed.ac.uk Internet Source	1 %
11	Handbook of Genetic Programming Applications, 2015. Publication	1 %
12	digibuo.uniovi.es Internet Source	1 %
13	ijphs.iaescore.com Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		