Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 4 No. 3 (2020) 531 - 536 ISSN Electronic Media: 2580-0760

Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest

Abduh Salam[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni[3]
[1,2,3]Informatics, School of Computing, Telkom University
[1]abduhsalam@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id, [3]yuliant@telkomuniversity.ac.id

Abstract

Indonesia is a country that is prone to Dengue Fever, this happens because Indonesia is a country with a tropical climate. More than 50 years after Indonesia contracted the dengue virus, dengue fever cases have not been resolved, currently the cases that occur are greatly increased over time this happens because of factors that cause dengue fever. By considering this serious problem, the authors created a system that can predict the vulnerability level in Bandung and looks for the factors that most influence from all factors of Dengue Fever using the KNN Algorithm and Random Forest. The results of the system show the results of the best model is KNN algorithm with RMSE 29,26, and from the model shows the most influencing factors are population density, growth rate population mobility, rainfall, wind speed. by utilizing the results of the study, the government can adjust actions to each level of sub-district vulnerability and pay more attention to the factors that most influence dengue fever according to the results of the study.

Keywords: Dengue Fever, Bandung, KNN, Random Forest.

1. Introduction

Indonesia is a country that is prone to Dengue Fever, this is because Indonesia is a country with a tropical climate. Data at the Bandung health department shows that the highest number of cases occurred in 2013, which was 5,736 cases. This number then dropped in 2014 with 3,132 cases. But then again rise in 2015 which is 3,640. Likewise in 2016 it rises to 3,880. But then back down in 2017 namely 1786 and again rising in 2018 namely 2,826 cases [1]. This shows that cases that occur in dengue fever are difficult to handle seeing the number of cases that do not go down.

In a previous study entitled Research of Dengue Fever Prediction in San Juan, Puerto Rico Based on a KNN Regression Model, it contained a program built using the K-NN algorithm to predict the number of dengue fever cases each week and looked for significant correlations of dengue fever factors [2]. but this study only looks for factors that focus on the temperature factor while there are several other factors that cause dengue fever and this study does not show the vulnerability level in the area so it is difficult for ordinary people to understand.

Other research entitled Mapping dengue risk in Singapore using Random Forest conducted by Janet Ong, Xu Liu, Jayanthi Rajarethinam, conducted a map-making study of DHF risk using Random Forest which has very good accuracy results [3]. But, this study does not mention its accuracy value but only mentions very good, and does not explain why the factors that cause Dengue Fever are chosen and the prediction map made it difficult for the reader to understand.

Next research, built the program using fuzzy algorithm to predicts high levels of Dengue fever vulnerability in provinces in the Philippines [4]. But, in this study, the predicted area is too wide, province, the predicted level of vulnerability is only high and does not indicate the accuracy of the program made.

Next research, Predicts of DHF disease spreading patterns using inverse distances weighted (IDW), ordinary and universal kriging [5]. But, the result of this study did not mention how well the algorithm works based on the case but only compares between the result of methods and didn't mention the factor of Dengue Fever which can help preventive action better.

Based on the discussion researches above with the shortcomings that have been submitted, the authors built a program that compares the K-NN algorithm and Random Forest which predicts the vulnerability level of dengue fever in Bandung and looks for the factors that most influence dengue fever. The program that built showed accuracy, using several factors as attributes including rainfall, humidity, wind, temperature, rainfall,

Accepted by editor: 05-05-2020 | Final Revision: 11-06-2020 | Online Publication: 20-06-2020

population density, and growth rate population mobility, explained why these factors were used, and map output that showed low, medium, and high in every sub-district in the Bandung district.

2. Research Method

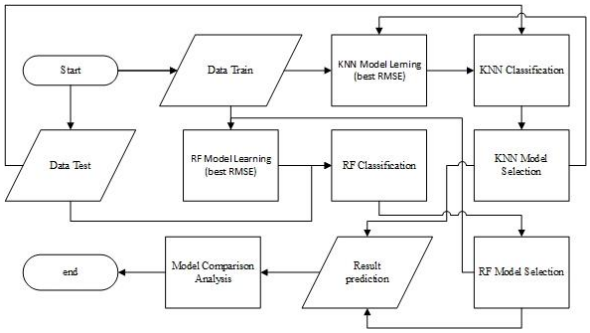

Figure 1 System Architecture

In this research, 2 algorithms are used, K Nearest Neighbor, and Random Forest. The following system architecture is built in Figure 1.

2.1. Data set

The data used are data of dengue fever consisting of 30 districts in Bandung. The data consists of 6 attributes they are humidity (%), wind (knot), temperature (Celcius), rainfall (millimeter), population density (ha), and growth rate population mobility (%) obtained from BMKG and BPS, and 1 attribute of the number of dengue cases (person) obtained from the health office in Bandung regency. The data collected amounted to 210 data from 2012 to 2018.

Table 1 Minimal and maximum each factor

Factor	Min	Max
X1	36	699
X2	149.06	322.4
X3	3	4.6
X4	67	84
X5	0.01	8.2
X6	19.8	28
Y	12	407

2.2. Data Modelling

The initial form of data is X1(n) ... Xk (n), Y (n), and data model for prediction X1 (n) ... Xk (n), Y (n + 2).

Which n = year, k = number of attribute. the following example of data

Table 2 Initial Data

Num of data	X1	X2	X3	X4	X5	X6	Y
1	264	185	4,2	75	0,1	23,5	33
2	186	185	4,2	75	0,4	23,5	93
3	112	199	3	77	0,4	23,4	63

There are 2 data modelling in this study, namely patterned data model and random data model. In the patterned data set, the data used has a pattern that predicts year using data train 2 years before, the predictions do in 2014 - 2018, for example, 2018 prediction, then the test data uses 2018 data (factor X1, X2, X3, X5, and X5), and the data train uses data 2 previous year 2016 (Factor X1, X2, X3, X5, and X5 ,and output Y).

$$X_{x-2}, Y_N$$

Which X is data train and Y is Data test.

In the random data model, the datasets are divided into data train and data test, where the division is 80% data train and 20% data test, the data distribution is taken randomly. Data train is used to make models and test data is used as a measure of how well the performance of the model is based on accuracy. This division of data was carried out 5 times to find the best model.

2.2. Preprocessing

In this process, the original data is in csv file form then will be converted into a multidimensional array. Every data will be checked, if there is data that has a missing value, then the data will not be used in the next process. Data will be divided into 2 types, namely data train and test data. Data train is part of a dataset that is train to build algorithms model and data tests are parts of data sets used to test the built model.

And data will be normalized, the data normalization function is to change/transform into more appropriate/suitable form for the data mining process. The Min-Max method is a normalization method by performing a linear transformation of the original data, with the following formula [6].

$$x^* = \frac{x - \min(x)}{\max(x) - min(x)} \quad (1)$$

Which x is original data.

2.3. K Nearest Neighbor (KNN)

K-NN algorithm is an algorithm that has a function to classify data based on training data taken from k nearest neighbors, where k is the number of nearest neighbors. K-NN classifies with projected learning data in a multi-dimensional space. Each data train is represented as points on multiple dimensional spaces. The Following is KNN algorithm :

To use the K nearest neighbor algorithm, it takes the number of K nearest neighbors used to do the classification, k is a positive integer value.

New classified data is then projected on multiple dimensional space. In the classification process is done by finding the closest point. Here are some distance formulas:

Euclidian distance is the function most often used, when A and B represent vectors A = (x1, x2, ..., xn) and B = (y1, y2, ..., yn), where n is the dimension of the space

feature, To calculate the distance between A and B, here is the euclidean distance formula [7]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (2)$$

Which n is the number of attributes, x is data test and y is data train.

Several other distance functions that can be used by the KNN algorithm, such as Manhattan.

$$d(x,y) = \sum_{i=1}^{n}|xi - yi| \quad (3)$$

Which n is the number of attributes, x is data test and y is data train.

$$d(x,y) = (\sum_{i=1}^{n}|xi - yi|^p)^{1/p} \quad (4)$$

Which n is the number of attributes, x is data test, y is data train, p = 1, manhattan distance P = 2, euclidean distance.

The formulas 2,3, and 4 stated that i is a factor that causes dengue fever, for example if i = 1 is a factor of rainfall and so on as much as n factors.

After calculating the distance, order the data ascending according to the distance calculation results. This stage is used to find the nearest neighbor with a predetermined k value (average k nearest neighbor for regression).

By repeating the algorithm above, determining the value of k is determined based on the best accuracy value model. If a small k value can occur noisy and can have an outlier's effect and if the k value is greater then it has a finer decision limit but reduces variance and increases bias [8].

The following is KNN example:

Data train input example.

Table 3 Data Train KNN example

Num of data	X1	X2	Y
1	5	10	5
2	15	10	15
3	10	10	10

Data test input example.

Table 4 Data Test KNN example

Num of Data	X1	X2
1	5	10

Initialize K=2 and distance function formula 2.

Compute the distance between each data test to data train use formula 2.

Table 5 Computed Distance

Num of data	Distance
1	0
2	10
3	5

Ascending sort by distance value.

Table 6 sorted by distance value

Num of data	Distance
1	0
3	5
2	10

Since K=2, compute average Y of 2 data nearest neighbor.

$$Average = \frac{5 + 10}{2} = 7{,}5$$

The prediction of the data test is 7,5.

2.4. KNN Model Selection

The model built on KNN is based on dataset inputs and random forest parameters (number of trees built and split). The dataset input will be combined 4 and 5 of the 6 factors. For each combination of dataset, try parameters k values with 1 to 20 and distance function parameters will be tested one by one. Models with the lowest RMSE values will be used for model comparison analysis.

2.5. Random forest (RF)

Breiman is the creator of the Random Forest algorithm, which is added to the algorithm is the randomness and bagging method. Also to build each tree use different bootstrap data samples, random forest changes how classification or regression trees are constructed. In the standard tree, each node is divided using the best split among all variables [9].

This algorithm is somewhat counterintuitive, but it works very well compared to other classifiers, including discriminant analysis, machine vector support and neural networks, and strong overfitting. Also, random forest is very friendly for programmers, in the sense that it only has two parameters (the number of variables in a random subset at each node and the number of trees in the forest), and usually these parameters are not very sensitive to their values . The following is Random Forest Algorithm [9]:

Create n bootstrap dataset (each tree in a random forest learns from a random sample of the data points), then for each bootstrap sample, create a unpruned classification tree or regression, at each node, better to take a random sample predictor and choose the best split among the variables rather than choosing the best split among all the predictors.

The process of building a regression tree. Roughly speaking, there are two steps. First divide the data train predictor spaces into different and non-overlapping regions (R_1, R_2, R_3 ... R_J). Then for each test data included in an region of R_j, the predict is the average value of the the data training output included in Rj [10].

Predict data train by aggregating the predictions of the ntree trees (average for regression and majority votes for classification).

The following is RF Algorithm example:

Data train input X and Y

Table 7 Data Train RF Example

Num of data	X	Y
1	25	25
2	40	15
3	50	10

Data test input X1.

Table 8 Data Test RF Example

Num of data	X
1	30

Initialize number of tree = 2 and number of variable = 1.

Build 2 (num of tree) the regression tree bootstrap. Figure 2 shows the regression tree.

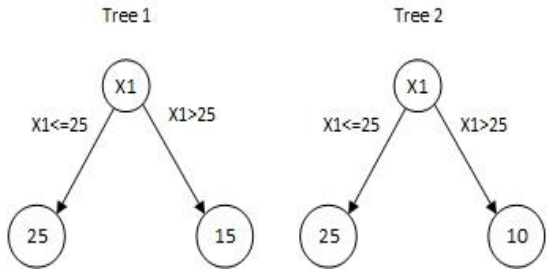

Figure 2 Regression Tree Built Based on Data Train

Insert data test X=30 into each built regression tree the results are 15 (tree 1), and 10 (tree 2).

To get a prediction for data test, Compute average from all regression tree prediction t.

$$Average = \frac{15 + 10}{2} = 12{,}5$$

The prediction for the data test is 12,5

2.6. Random forest Model Selection

The model built on a random forest-based on dataset inputs and random forest parameters (number of trees built and split). The dataset input will be combined 4 and 5 of 6 factors. From each combination of dataset inputs we will try parameters of the number of trees with values from 64 to 128 and the number of variables in the random subset at each node decided by sklearn liblary. Models with the lowest RMSE values will be used for model comparison analysis.

2.7. Model Testing

Model testing requires performance measurement tools to calculate the accuracy of the model, each model calculated for accuracy, because the case is regression, then the prediction measurement tool used is RMSE.

RMSE is an alternative method for evaluating forecasting techniques used to measure the accuracy of the forecast results of a model. RMSE is the average value of the sum of squares error [11].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2}{n}} \quad (5)$$

Which Y is the Original data test, $\hat{Y}$ is prediction data test and n is the sum of the data test.

A good RMSE value is a low value, a low RMSE value indicates that the variation in values produced by a forecast model is close to the variation in its observation value [12].

2.8. Model Comparison Analysis

Comparative analysis of the two algorithms is based on the accuracy value of the model, because the prediction measuring instrument used is RMSE, the lowest RMSE value of the model used to create a vulnerability level map.

3. Result and Discussion

In this study, the KNN algorithm was used in the experiment parameters K 1 to 20 and used 2 distance formulas. and the random forest algorithm with parameter numbers from tree 64 to 128 and split parameters 5,10,15 and 20. From these two methods the parameter experiments that have the best accuracy are used as the best model.

For the determination of the most influential causative factors, a combination of 4 and 5 of the 6 causes of dengue fever is used using the same algorithm as the previous paragraph. The most influential factor is the combination that has the best accuracy.

Following the results of comparison of accuracy in Experiment 1 and Experiment 2 from several scenarios combined with the prediction of DHF with population density (X1), Rainfall (X2), Wind Speed (X3), Humidity (X4), growth rate Population Mobility (X5) and Temperature data (X6).

Table 9 standard deviation for each factor without min-max normalization

Factor	Mean	Standard deviation
X1	207.7	147.4
X5	0.8	0.8
X2	218.4	50.4
X3	3.	0.6
X4	77.2	2.5
X6	23.	0.5
Y	103.4	66.2

Table 9 is the standard deviation for each factor without normalization, which shows the large variance of each factor affecting dengue fever, factor X5 has a small

difference between the mean and standard deviation, which means that the variance of factor X5 has a small variance.

3.2. Patterned Data Model

Table 10 Patterned data set experiment

years to predict	RMSE	
	no min-max	min max
2014	56.71	58,85
2015	29.25	29,83
2016	30.7	54,97
2017	36.22	37,69
2018	45.87	44,67

Based on pattern data model results, the model built in 2015 without min-max normalization has the best RMSE value compared to other years and use min-maxnormalization. Detail of 2015 models built are the following:

Table 11 model built in 2015

sum of factor	combination	Best Model	
		KNN	RF
4	X1, X5, X2, X3	31.95	32.43
4	X1, X5, X2, X4	31.95	33.01
4	X1, X5, X2, X6	31.95	32.55
4	X1, X5, X3, X4	34.06	33.18
4	X1, X5, X3, X6	34.19	32.38
4	X1, X5, X4, X6	34.06	32.69
4	X1, X2, X3, X4	31.95	33.30
4	X1, X2, X3, X6	31.92	32.57
4	X1, X2, X4, X6	31.95	33.17
4	X1, X3, X4, X6	34.02	33.39
4	X5, X2, X3, X4	29.50	32.59
4	X5, X2, X3, X6	29.63	32.05
4	X5, X2, X4, X6	29.50	32.87
4	X5, X3, X4, X6	32.76	32.61
4	X2, X3, X4, X6	29.26	33.35
5	X1, X5, X2, X3, X4	31.95	33.17
5	X1, X5, X2, X3, X6	31.95	32.38
5	X1, X5, X2, X4, X6	31.95	32.84
5	X1, X5, X3, X4, X6	34.06	33.29
5	X1, X2, X3, X4, X6	31.95	33.24
5	X5, X2, X3, X4, X6	29.50	32.82
6	X1, X5, X2, X3, X4, X6	31.95	33.31

Based on the 2015 model details, the best combination is X2, X3, X4, and X6 with RMSE value of 29.26. This combination means the most influencing factors based on the 2015 model.

3.3. Random Data Model

Table 12 Random data set experiment

Loop	Model	RMSE	
		no min-max	min-max
1	RF	46.11	47,81
2	KNN	50.56	45,7
3	KNN	48.04	50,45
4	RF	48.04	45,96
5	RF	45.48	49,3

The results of the random data set have the lowest RMSE value of 45.48 without min-max normalization. But compared to the patterned data set, the value of the RMSE pattern data model has much better RMSE value and the RMSE value model is still not good to implement.

3.4. Discussion.

Based on the results of this study, the results of the pattern data model have much better RMSE value than the random data model, when compared with the average Y, the number of deviation is 28.3% obtained, therefore the results from patterned data sets without data normalization process 2015 models will be used as predictors Vulnerability level of Dengue Fever in Bandung 2020. And also seen from the accuracy of the model without min-max normalization tends to be better than using min-max normalization this happens because min-max normalization cannot handle outliers properly.

Figure 3 is a prediction map of dengue fever vulnerability level in Bandung 2020 based on the results from patterned data sets. Figure 3 is the conversion of prediction results from the number of cases to the level of vulnerability with 3 levels, namely low (incidence rate < 20/100.00 population), medium (incidence rate 20-55/100.00 population), and high (incidence rate > 55/100.00 population) [13]. This conversion is done to make it easier for readers to understand the information provided. It can be seen that most districts have a high vulnerability level, this is a warning to the government to take actions such as socialization (maintaining hygiene, maintaining immunity and increasing

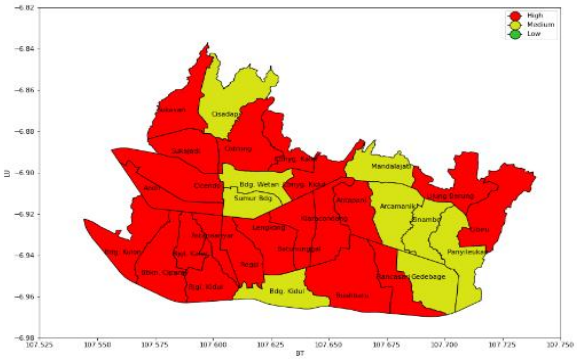

Figure 3 Prediction of Vulnerability level of Dengue Fever based on best model

knowledge about dengue) with frequent and fogging in areas that have high vulnerability level.

4. Conclusion

Based on the results it is can be concluded. Population Density, growth rate population mobility, rainfall, and wind speed are the most influencing factors of the 6 factors, with KNN algorithm RMSE value of 29.26. The results of this study are good enough to be implemented in the real world.

Based on figure 3, the prediction shows Bandung will have an emergency situation where 21 sub-districts have

high vulnerability, and the rest of the sub-districts will have medium vulnerability. with this information the government must have better preventive actions that focus on the factors that most influence in accordance with the results of the study.

References

[1] Z. Istiqomah, "Dinkes Kota Bandung Catat Peningkatan Pasien DBD," *REPUBLIKA.co.id*, 2019. [Online]. Available: https://www.republika.co.id/berita/nasional/daerah/19/01/28/pm17ip335-dinkes-kota-bandung-catat-peningkatan-pasien-dbd. [Accessed: 02-Dec-2019].

[2] Y. Jiang, G. Zhu, and L. Lin, "Research of dengue fever prediction in san juan, puerto rico based on a KNN regression model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[3] J. Ong *et al.*, "Mapping dengue risk in Singapore using Random Forest," *PLoS Negl. Trop. Dis.*, 2018.

[4] A. L. Buczak *et al.*, "Prediction of High Incidence of Dengue in the Philippines," *PLoS Negl. Trop. Dis.*, 2014.

[5] S. S. Prasetiyowati and Y. Sibaroni, "Prediction of DHF disease spreading patterns using inverse distances weighted (IDW), ordinary and universal kriging," in *Journal of Physics: Conference Series*, 2018.

[6] J. Xie, W. Jiang, and L. Ding, "Clustering by Searching Density Peaks via Local," 2017.

[7] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, 2016.

[8] A. Mucherino, P. J. Papajorgji, and P. Pardalos, *Data Mining in Agriculture*. Springer-Verlag New York, 2009.

[9] A. Liaw and M. Wiener, "Classification and Regression with Random Forest," *R News*, 2002.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," in *An Introduction to Statistical Learning with Applications in R*, Springer, 2013, pp. 308–310.

[11] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, 2014.

[12] Indrabayu, N. Harun, M. S. Pallu, A. Andani, and F. C.L, "Prediksi curah hujan dengan jaringan saraf tiruan," *Gr. Tek. Elektro*, vol. 6, pp. 978–979, 2012.

[13] U. F. Achmadi *et al.*, "Buletin Jendela Epidemiologi," *Kementrian Kesehatan Republik Indonesia*, vol. 2, no. Demam Berdarah Dengue, Aug-2010.