Level alert classification of dengue hemorrhagic fever cases in DKI Jakarta with the implementation of the random forest algorithm

Ikhwanul Ghazy Dzakwan, Gatot Fatwanto Hertono*, Dipo Aldila, and Bevina Desjwiandra Handari
Deparment of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Corresponding author : *gatot-fl@ui.ac.id

Abstract. In 2007, the capital district, DKI Jakarta had one of the worst floods that submerged nearly 60% of the area. One of the causes was heavy rainfall. Two months after the incident, the Governor of DKI Jakarta stated that the capital district was affected by an outbreak of dengue fever. From these two incidents, there are some indications of dengue hemorrhagic fever cases related to weather. In this research, a random forest classification algorithm was used to build a model that could classify the level alert of dengue fever for each district in DKI Jakarta into three categories: safe, moderately safe, and unsafe. The results and the accuracy of the classification model for the level alert of cases of dengue hemorrhagic fever for East Jakarta, North Jakarta, South Jakarta, Central Jakarta, and West Jakarta are 94.51%, 86.81%, 81.32%, 80.22%, and 80.22% respectively

1. Introduction
Dengue hemorrhagic fever is an infection caused by the dengue virus, which is transmitted through infected Aedes spp. mosquitoes [1]. The incidence rate is the number of new cases of a disease in a population at risk during a certain period of time, and can be calculated by counting the number of new cases suffering from illness in a certain time period divided by the number of individuals in the population who at risk in a certain time period [2]. In DKI Jakarta, BMKG uses monthly incidence figures to form the basis for categorizing the level of alertness for dengue hemorrhagic fever cases for each district which is categorized into three categories: safe, moderately safe, and unsafe.

Weather can affect cases of dengue hemorrhagic fever. Rainfall, temperature and humidity are the climatic factors that influence dengue hemorrhagic fever cases the most [3]. Heavy rainfall will increase the number of breeding places for mosquitoes to breed [4]. The average optimum temperature for mosquito growth is 25° - 35° C and the Aedes aegypti mosquito can survive low temperatures [4]. Humidity of more than 60% extends the life of the Aedes aegypti mosquito and makes mosquito breeding more likely [4].

One of the machine learning algorithms, the random forest classification algorithm, has been used to map the area of dengue fever cases in Singapore [5] and the City of Shenzhen, China [6]. In this study, a random forest classification algorithm will be used to build a model that can classify the level of alertness in cases of dengue hemorrhagic fever with data. Then the accuracy value for each model that has been built will be calculated using a confusion matrix.
2. Materials and methods

2.1. Materials
The data that will be used in this research is daily data from rainfall, temperature, and humidity (2008 to 2016) and DHF incidents (2009 to 2016) in several districts in DKI Jakarta (except for Kepulauan Seribu), such as West Jakarta, East Jakarta, South Jakarta, North Jakarta, and Central Jakarta, that was obtained from Meteorological, Climatological, and Geophysical Agency (BMKG) and DKI Jakarta Health Office. There is a missing value in the rainfall, temperature, and humidity data, so an imputation process is carried out using the average value for each data in each district. There were no missing values in DHF incidents data, so the imputation process for DHF incidents is not necessary. The daily data was then transformed into 467 weekly rainfall, temperature, and humidity data and 453 weekly DHF incidents data. Visualization for each data in each district can be seen in Figures 1-4.

![Rainfall](image1)

**Figure 1.** Rainfall data plot in each district

![Temperature](image2)

**Figure 2.** Temperature data plot in each district
In this research, the dataset will be divided into training data and testing data. The composition used 80% for training data and 20% for testing data. The population data for each district (2010) in DKI Jakarta (except for Kepulauan Seribu) from Central Bureau of Statistics will be used to determine the total population at risk in the calculation of the incidence rate.

2.2. Random forest

Random forest is a classifier that consists of a collection of classification trees \( \{ h(x, \Theta_k), k = 1, \ldots \} \) where \( \Theta_k \) independently distributed randomly and each tree provides a vote for the most popular class on the input \( x \) [1]. Random forest is an ensemble learning method, proposed by Breiman (2001) by adding a randomness to bagging concept, then building \( n_{\text{tree}} \) decision tree from each sub-dataset obtained through random sampling of the original dataset with returns. Besides building each tree using
a different bootstrap sample, random forest algorithm changes the way the decision tree is constructed. On the decision tree, each node is divided using the best separator among all variables. However, on random forest, each node is divided using the best separator among subset predictor randomly selected on the node. The strategy works well for defending against overfitting [7]. Random forest is user-friendly because it only has 2 parameters: the number of variables in the random subset at each node and the number of decision trees in the random forest [8].

In this research, random forest algorithm, which is a form supervised learning will be used. The process of random forest algorithm is: [8].

a. Choose \(n_{tree}\) bootstrap sample from every dataset,
b. For each bootstrap sample, build a decision tree with a modification on each node, and instead of choosing the best separator among all predictors, attempt random sampling on \(m_{try}\) from predictors and choose the best separator from each variables,
c. Predict a new dataset by joining \(n_{tree}\) decision tree. For a classification problem, majority votes will be used. For a regression problem, the mean value will be used.

2.3. Confusion matrix
The correctness of classification can be evaluated by counting the number of positive classes that are predicted to be positive (true positive), the number of negative classes that are predicted to be negative (true negative), the number of classes that are predicted to be negative but belong to positive classes (false positive), and the number of classes predicted to be positive but belong to negative classes (false negative). These four calculations form a confusion matrix [9] which is shown in Table 1.

| Actual | Predicted |
|--------|-----------|
|        | True Positive (TP) | False Negative (FN) |
|        | False Positive (FP) | True Negative (TN) |

From table 1, the accuracy value can be calculated to measure the accuracy of the model in classifying with the following equation [9]

\[
\text{accuracy} = \frac{\sum_{l} TP_{l} + TN_{l}}{TP_{l} + TN_{l} + FP_{l} + FN_{l}} \times 100\%
\]

2.4. Proposed methods
In this research, dengue hemorrhagic fever case classification models will be built for each district in DKI Jakarta (except for Kepulauan Seribu) using a random forest algorithm and weekly data on rainfall, temperature, humidity, and dengue incidence. The first step in data processing is to carry out the imputation process for the missing value for each daily data variable in each district using the average value of each variable in each district. Then the conversion will be carried out from daily data to weekly data. There are 3 categories of data on the classification result variables, namely: safe, moderately safe, and unsafe. Data on the variable classification results are obtained by calculating the monthly incidence rate which is then categorized based on BMKG's categorization. an incidence rate of less than 3 will be safe, between 3 and 10 will be moderately safe, and more than 10 will be unsafe. Because the data used is weekly, to adjust to the monthly data, the assumption is that in one month there are four weeks, so that the weekly incidence rate can be multiplied by 4 to adjust BMKG's categorization.
There is a time lag required by the variables in affecting the classification result variables. Therefore, feature selection will be carried out based on the level of correlation between the incident variables at a certain time, with previous incident data, previous rainfall, previous temperature, and previous humidity. Then, after feature selection, the data will be divided by the proportion of 80% for training data and 20% for testing data.

The random forest algorithm has a parameter of the number of trees that can be adjusted to get the best model. The out of bag (OOB) error rate is one of the features of the random forest which will be used to validate the selection of the best tree number parameters in model building using training data. The smaller the number of trees built, the lighter the computational burden. After the best parameters are obtained, the model will be built and then it will be evaluated using testing data which is measured using confusion matrix.

3. Experimental results
The variable selection that will be used in building each model focusing on the correlation value of each variable with the incident variable with a simulated lag. The results of the correlation value between variables can be seen in Tables 2-6.

Table 2. The correlation value of each variable with the incidence variable in East Jakarta

| Lag | Temperature | Rainfall | Humidity | Incidence |
|-----|-------------|----------|----------|-----------|
| 1   | 0.02282     | 0.05226  | 0.12741  | 0.86119   |
| 2   | -0.00243    | 0.09272  | 0.16529  | 0.8318    |
| 3   | -0.03119    | 0.12816  | 0.17761  | 0.79219   |
| 4   | -0.0552     | 0.14109  | 0.18535  | 0.78031   |
| 5   | -0.07181    | 0.15841  | 0.20647  | 0.74455   |
| 6   | -0.09875    | 0.15848  | 0.22848  | 0.68294   |
| 7   | -0.10691    | 0.20805  | 0.24044  | 0.65091   |
| 8   | -0.11932    | 0.2072   | 0.2618   | 0.62022   |
| 9   | -0.13474    | 0.19213  | 0.2392   | 0.59849   |
| 10  | -0.14636    | 0.21035  | 0.24383  | 0.53347   |
| 11  | -0.16064    | 0.19756  | 0.22238  | 0.49596   |
| 12  | -0.12185    | 0.188    | 0.18262  | 0.44681   |

Table 3. The correlation value of each variable with the incidence variable in Central Jakarta

| Lag | Temperature | Rainfall | Humidity | Incidence |
|-----|-------------|----------|----------|-----------|
| 1   | -0.03636    | 0.07472  | 0.26822  | 0.73583   |
| 2   | -0.08125    | 0.13905  | 0.32281  | 0.68652   |
| 3   | -0.15931    | 0.17207  | 0.38619  | 0.63525   |
| 4   | -0.17641    | 0.21222  | 0.40434  | 0.63281   |
| 5   | -0.20391    | 0.2552   | 0.41804  | 0.56879   |
| 6   | -0.22995    | 0.26692  | 0.437    | 0.51319   |
| 7   | -0.21762    | 0.27901  | 0.43416  | 0.42486   |
| 8   | -0.23969    | 0.28573  | 0.4322   | 0.3782    |
| 9   | -0.25886    | 0.2849   | 0.43936  | 0.33898   |
| 10  | -0.227      | 0.22873  | 0.39787  | 0.27552   |
| 11  | -0.17511    | 0.23916  | 0.36082  | 0.26068   |
| 12  | -0.13793    | 0.21142  | 0.32187  | 0.20657   |
Table 4. The correlation value of each variable with the incidence variable in South Jakarta

| Lag | Temperature | Rainfall | Humidity | Incidence |
|-----|-------------|----------|----------|-----------|
| 1   | -0.00497    | 0.06021  | 0.30771  | 0.85655   |
| 2   | -0.03864    | 0.05918  | 0.32533  | 0.84911   |
| 3   | -0.03447    | 0.0609   | 0.32004  | 0.80503   |
| 4   | -0.01682    | 0.06453  | 0.33607  | 0.80503   |
| 5   | -0.03981    | 0.06704  | 0.35846  | 0.77601   |
| 6   | -0.03214    | 0.06383  | 0.3443   | 0.7382    |
| 7   | -0.02257    | 0.07541  | 0.34007  | 0.71478   |
| 8   | -0.01795    | 0.0657   | 0.329    | 0.66268   |
| 9   | -0.01644    | 0.05304  | 0.31099  | 0.64241   |
| 10  | 0.00268     | 0.04386  | 0.30374  | 0.59334   |
| 11  | 0.05194     | 0.04268  | 0.26388  | 0.5463    |
| 12  | 0.07574     | 0.02824  | 0.23272  | 0.51216   |

Table 5. The correlation value of each variable with the incidence variable in West Jakarta

| Lag | Temperature | Rainfall | Humidity | Incidence |
|-----|-------------|----------|----------|-----------|
| 1   | 0.04472     | 0.0599   | 0.30712  | 0.74732   |
| 2   | 0.02185     | 0.10823  | 0.35949  | 0.69789   |
| 3   | -0.01827    | 0.16122  | 0.37797  | 0.62514   |
| 4   | -0.05866    | 0.19607  | 0.40654  | 0.6152    |
| 5   | -0.04837    | 0.23847  | 0.417    | 0.53752   |
| 6   | -0.07312    | 0.30337  | 0.43535  | 0.45615   |
| 7   | -0.06691    | 0.32137  | 0.43476  | 0.4001    |
| 8   | -0.1001     | 0.2784   | 0.43605  | 0.3521    |
| 9   | -0.13848    | 0.34209  | 0.47397  | 0.3252    |
| 10  | -0.07155    | 0.27679  | 0.42471  | 0.21439   |
| 11  | -0.02411    | 0.26523  | 0.39489  | 0.20979   |
| 12  | -0.05704    | 0.18652  | 0.31079  | 0.14873   |

Table 6. The correlation value of each variable with the incident variable in North Jakarta

| Lag | Temperature | Rainfall | Humidity | Incidence |
|-----|-------------|----------|----------|-----------|
| 1   | 0.06111     | 0.00965  | 0.14807  | 0.74765   |
| 2   | 0.04213     | 0.00704  | 0.19639  | 0.7866    |
| 3   | 0.02688     | 0.06619  | 0.21858  | 0.70769   |
| 4   | 0.01407     | 0.08766  | 0.24117  | 0.71185   |
| 5   | -0.01421    | 0.09987  | 0.27221  | 0.66935   |
| 6   | -0.02597    | 0.14553  | 0.28302  | 0.59739   |
| 7   | -0.00798    | 0.09064  | 0.26794  | 0.58856   |
| 8   | -0.03938    | 0.10237  | 0.2954   | 0.52283   |
The results of lag selection for each variable are summarized in table 7.

| Variable | East Jakarta | Central Jakarta | South Jakarta | West Jakarta | North Jakarta |
|----------|--------------|----------------|---------------|--------------|---------------|
| Temperature | 11 | 9 | 5 | 9 | 8 |
| Rainfall | 10 | 8 | 7 | 9 | 10 |
| Humidity | 8 | 9 | 5 | 9 | 8 |
| Incidence | 1 | 1 | 1 | 1 | 2 |

Afterwards, each data in the dataset was labeled into 3 categories represented by numbers 1 for safe, 2 for nearly safe, and 3 for not safe. The labeling uses dengue hemorrhagic fever incidence data which has been converted into a weekly incidence rate, then adjusted to a weekly incidence rate which is subsequently adjusted for classification from BMKG.

After the data preparation has been completed, the dataset is split into the proportion of 80% for training data and 20% for testing data. The parameter to be simulated is the n_estimators parameter or the number of trees built. To validate the value to be selected from the simulation results using training data, the best parameter selection will be based on the lowest OOB error rate value. The following are the simulation results for each district visualized in Figures 5-9.

![Figure 5. Simulation of the number of trees with the OOB error rate in East Jakarta](image-url)
Figure 6. Simulation of the number of trees with the OOB error rate in Central Jakarta

Figure 7. Simulation of the number of trees with the OOB error rate in South Jakarta
After the simulation is complete, the selection of the value of n_estimators or the number of trees built is based on the smallest error rate OOB value and the lowest number of n_estimators. The selection of parameter values is summarized in Table 8.
Table 8. The value of the n_estimators parameter that is used for each variable in each district.

| District          | East Jakarta | Central Jakarta | South Jakarta | West Jakarta | North Jakarta |
|-------------------|--------------|-----------------|---------------|--------------|---------------|
| n_estimators      | 100          | 150             | 100           | 150          | 650           |

The model for each district that has been built using training data and the n_estimators parameter in Table 8 will be evaluated using confusion matrix to see the accuracy of the model in classifying the level of alertness for dengue hemorrhagic fever cases for each district in DKI Jakarta against testing data. The results of confusion matrix evaluation for each model can be seen in Tables 9-13.

Table 9. Confusion matrix model in East Jakarta

|       | Actual |          |          |          |          |
|-------|--------|----------|----------|----------|----------|
| Predicted | 74     | 4        | 0        |          |          |
|        | 1      | 11       | 0        |          |          |
|        | 0      | 0        | 1        |          |          |

Table 10. Confusion matrix model in Central Jakarta

|       | Actual |          |          |          |          |
|-------|--------|----------|----------|----------|----------|
| Predicted | 31     | 5        | 0        |          |          |
|        | 8      | 38       | 5        |          |          |
|        | 0      | 0        | 4        |          |          |

Table 11. Confusion matrix model in South Jakarta

|       | Actual |          |          |          |          |
|-------|--------|----------|----------|----------|----------|
| Predicted | 42     | 9        | 0        |          |          |
|        | 6      | 28       | 1        |          |          |
|        | 0      | 1        | 4        |          |          |

Table 12. Confusion matrix model in West Jakarta

|       | Actual |          |          |          |          |
|-------|--------|----------|----------|----------|----------|
| Predicted | 46     | 7        | 0        |          |          |
|        | 7      | 26       | 3        |          |          |
|        | 0      | 1        | 1        |          |          |

Table 13. Confusion matrix model in North Jakarta

|       | Actual |          |          |          |          |
|-------|--------|----------|----------|----------|----------|
| Predicted | 72     | 7        | 0        |          |          |
|        | 4      | 6        | 0        |          |          |
|        | 0      | 1        | 1        |          |          |

From Tables 9-13, using equation (1), the calculated value of the model's accuracy in classifying the testing data is summarized in Table 14.
Table 14. Calculation results of model accuracy for each district

|                | East Jakarta | Central Jakarta | South Jakarta | West Jakarta | North Jakarta |
|----------------|--------------|-----------------|---------------|--------------|--------------|
| Accuracy       | 94.51%       | 80.22%          | 81.32%        | 80.22%       | 86.81%       |

From the results shown in Table 14, each model has an accuracy value of over 80%. The East Jakarta model has the highest accuracy value with an accuracy value of 94.51%, while the lowest accuracy value is obtained by the Central Jakarta and West Jakarta models with an accuracy value of 80.22%.

4. Conclusion

From the analysis and results, it can be concluded that the classification of the alertness level for dengue hemorrhagic fever cases by implementing the random forest algorithm provides an accuracy value for each district in DKI Jakarta of 94.51%, 86.81%, 81.32%, 80.22%, and 80.22% for East, North, South, Central and West Jakarta respectively. The results of this accuracy are obtained with the parameters in Table 8 and the lag value in Table 7. This research can only predict the level of alertness in the next week, because in Table 7, almost all models require incident data from the previous week. However, for the North Jakarta model, it uses incident data from the previous 2 weeks, so that it can predict the next 2 weeks.

Acknowledgement

This research is supported by Publikasi Terindeks Internasional (PUTI) 2020 Research Grant No: NKB-1011/UN2.RST/HKP.05.00/2020 from Universitas Indonesia. The data in this paper are provided by Jakarta Health Office, Central Bureau of Statistics, and Meteorological, Climatological, and Geophysical Agency.

References

[1] Indrayani Y A, & Wahyudi T 2018 Situasi Penyakit Demam Berdarah di Indonesia Tahun 2017 InfoDATIN Pusat Data dan Informasi Kementerian Kesehatan RI
[2] Rothman K J, Greenland S, & Lash T L 2008 Modern epidemiology Lippincott Williams & Wilkins
[3] Sintorini M M 2007 Pengaruh iklim terhadap kasus demam berdarah dengue Kesmas: National Public Health Journal, 2(1), 11-18
[4] Nirwana T, Rakasanagara A, & Afriand I 2013 Pengaruh Curah Hujan, Temperatur dan Kelembaban terhadap Kejadian Penyakit DBD, ISPA dan Diare Fakultas Kedokteran Universitas Padjadjaran
[5] Ong J, Liu X, Rajarethinam J, Kok S Y, Liang S, Tang C S, & Yap G 2018 Mapping dengue risk in Singapore using Random Forest PLoS neglected tropical diseases, 12(6), e0006587
[6] Mao L, Yin L, Song X, & Mei S 2016 Mapping intra-urban transmission risk of dengue fever with big hourly cellphone data Acta tropica, 162, 188-195
[7] Breiman L 2001 Random forests Machine learning, 45(1), 5-32
[8] Liaw A, & Wiener M 2002 Classification and regression by randomForest R news, 2(3), 18-22
[9] Sokolova M, & Lapalme G 2009 A systematic analysis of performance measures for classification tasks Information processing & management, 45(4), 427-437