Volume 8 No 1 June (2025)

e-ISSN: 2615-3270
p-ISSN: 2615-3599

Eigen Mathematics Journal

homepage: http://eigen.unram.ac.id/index.php/eigen

Implementation of Random Forest Algorithm to Classify Earthquake in Indonesia

Alda Putri Pratiwi[1], **Prizka Rismawati Arum**[1*]

[1] *Department of Statistics, Universitas Muhammadiyah Semarang, Indonesia*
[*] *Corresponding author: prizka.rismawatiarum@unimus.ac.id*

A B S T R A C T

Earthquakes are shocks that occur on the surface of the earth due to shifts in the earth's plates. Geographically, Indonesia is located in the Pacific Ring of Fire (King of Fire) region, this makes Indonesia prone to earthquakes. Earthquakes can cause environmental damage and tsunami disasters, but not all earthquakes can cause tsunamis. Classifying earthquakes that have the potential for a tsunami is very important to mitigate the damage caused. One classification method that has a high level of accuracy is random forest. The advantage of random forest is that this algorithm tends to be resistant to overfitting and can handle large data. This research uses real-time earthquake data from July to August 2023, sourced from the website of the Meteorology Climatology and Geophysics Agency (BMKG). The training data and test data used in this research are 70% and 30%. Confution Matrix is used as model evaluation, to measure the accuracy of the classification model. The results of this research obtained a high accuracy, equal 0.97 or 97%.

Keywords: random forest algorithm; earthquake; classification; tsunami

Received : 31-08-2023;
Revised : 24-01-2025;
Accepted : 06-02-2025;
Published : 10-04-2025;

DOI: https://doi.org/10.29303/emj.v8i1.185

1. Introduction

Indonesia is located in a very active tectonic zone because it has three large plates that meet each other. Therefore, this research will use the random forest algorithm to classify earthquake data to determine the potential strength of earthquakes in the Indonesian region. Earthquakes are divided into three types, namely earthquakes with shallow, medium and deep depths. Shallow earthquakes occur at depths of less than 70 km below sea level, medium earthquakes occur at depths between 70 km - 300 km, while deep earthquakes occur at depths of more than 300 km below sea level. Indonesia often experiences earthquakes because it is located in the Pacific Ring of Fire and has complex plate tectonics. Therefore, it is necessary to make preventive efforts, such as improving the quality of buildings, improving community skills in overcoming earthquakes, and increasing public awareness of natural disasters [1]. Earthquakes are one of the natural disasters that often occur in Indonesia, as the country is located on the Pacific Ring of Fire [2]. Earthquakes can be classified into several types, such

as tectonic earthquakes, volcanic earthquakes, and collapse earthquakes. The energy generated from an earthquake is transmitted in all directions as earthquake waves, which can be felt on the earth's surface. If an earthquake occurs at a shallow depth (0-70 km), there is the potential for a tsunami.

According to data from the Meteorology, Climatology and Geophysics Agency, earthquakes in Indonesia occur regularly within a few months of each other. Classification of earthquake depths is necessary to determine the potential strength of earthquakes in the Indonesian region, whether they cause tsunamis or not. Classification of earthquake depths is necessary to determine the potential strength of earthquakes in the Indonesian region, whether they cause tsunamis or not [3]. The results of classification often suffer from the problem of classification inaccuracy. Overcoming classification inaccuracy can be done by using machine learning that is able to handle unbalanced datasets, one of which is the random forest method. Random Forest is able to capture this nonlinear relationship well, thanks to the decision tree structure that can handle complex patterns. In addition, it can handle complex datasets well, allowing the use of various features to make more accurate predictions.

Previous research has applied the Random Forest method. In research conducted by [4] with the title "Implementation of the Random Forest Method for Student Majoring at Madrasah Aliyah Negeri Sintang" resulted in an accuracy value of 0.9438 or 94.38% indicating that the random forest algorithm is well used for the classification of three majors, namely Science, Social Studies and Religion. In addition [5] with the research "Classification of Earthquake Areas Using the Random Forest Algorithm" produces an accuracy value of 0.9997 or 99.97%, indicating that the random forest algorithm is good for classifying areas where earthquakes occur in the world.

Based on the explanation above, it is important to classify earthquakes that have the potential to cause a tsunami in order to mitigate the consequences that will be caused. Based on previous research, the Random Forest algorithm is very effective for classifying data with high accuracy results. Therefore, this study will use the Random Forest algorithm for the classification of earthquake data processed based on its depth.

Random Forest algorithm is a classification technique that has a fairly high level of accuracy, and random forest is based on decision tree techniques so that it can overcome nonlinear problems [6]. The advantages of the random forest algorithm in research are that this algorithm tends to be more resistant to overfitting, can cope with various types of data including numerical and categorical data, and can provide information about the importance of each feature in making predictions.

2. Research Methods

In this study, the data that the author uses is realtime data of earthquakes in Indonesia that occurred during July 2023 to August 2023, from the catalog of the Meteorology Climatology and Geophysics Agency. Indonesian territory with coordinates 6^0 LU - 11^0 LS and 95^0 East - 141^0 East, with magnitudes of 1.0 to 10.0, and depths of 1 to 1000km. The research variables consist of Latitude, Longitude, Magnitude, and Depth.

The variable observed in this research is depth (km), which is used to classify whether an earthquake has the potential to cause a tsunami or not. The data is divided into two parts with 70% training data and 30% testing data. The method used in this research is the random forest algorithm, which is a data classification method consisting of a number of decision trees in order to obtain more stable and accurate final results. This algorithm can be used for the classification of large amounts of data, so it can be used in various fields.

2.1. Research Steps

Some of the research stages carried out are as follows:

1. Data Collection
   The initial stage of this research aims to collect relevant and valid data to achieve the research objectives. The data that the author uses is realtime earthquake data in Indonesia from July 2023 to August 2023, obtained from the Meteorology Climatology and Geophysics Agency.
2. Pre-processing
   This stage is carried out by processing the data that has been obtained with a series of steps or processes carried out to clean, transform, and prepare raw data to be ready for use in analysis. The main purpose of data pre-processing is to improve data quality, reduce interference, and ensure data is ready for processing. This is done to overcome problems that may be present in the data.
3. Visualization
   After pre-processing the data, the next stage is data visualization, namely determining the random forest algorithm that will be used to classify the data. Data visualization is very important in the exploration, analysis, and communication of information from various types of data.
4. Algorithm Evaluation
   This stage is carried out by reviewing the research process, such as the methodology used, data analysis and results. The random forest classification method that has been optimized on earthquake depth data is evaluated using the Confusion Matrix algorithm. The evaluation aims to help identify weaknesses and to improve for future research. The classification is divided into three classes, so that accuracy, precision, and recall can be measured by calculating the average of the accuracy, precision, and recall values in each class. In addition, further analysis can also be done by evaluating the Confusion Matrix of each class to determine the error rate in classification.

2.2. Analysis of Research Data

This research uses data from the Meteorology, Climatology, and Geophysics Agency (BMKG) website. The amount of data used is around 708 data, using realtime data of earthquakes in Indonesia that occurred in July 2023 to August 2023. The data was used because during that month there was an increase in earthquake activity. The raw data that already exists, needs to be processed or preprocessed again.

Table 1. Data Structure

No	**Tsunami Event** (Y)	**Latitude** ($X1$)	**Longitude** ($X2$)	**Magnitude** ($X3$)	**Depth** ($X4$)
1	Y_1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$
2	Y_2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$
3	⋮	⋮	⋮	⋮	⋮
4	Y_{708}	$X_{1,708}$	$X_{2,708}$	$X_{3,708}$	$X_{4,708}$

Table 1. Data structure based on the parameters in this study, which consists of:

Y = Tsunami Event, 0 no tsunami potential and 1 tsunami potential
X_1 = Latitude (degree)
X_2 = Longitude (degree)
X_3 = Magnitude (mm)
X_4 = Depth (km)

The parameters contained in the website of the Meteorology, Climatology and Geophysics Agency (BMKG) are date time, latitude, longitude, magnitude, mag type, depth, phase count, azimuthgap, location and agency. While the variable observed in this study is depth (km), this variable is observed

to classify the type of earthquake that has the potential to cause a tsunami or does not have the potential to cause a tsunami. The data is divided into two parts with 70% training data and 30% testing data. The method used in this research is the random forest algorithm, which is a data classification method consisting of a number of decision trees to obtain more stable and accurate final results. This algorithm can be used for data classification in large quantities, so it can be used in various fields.

The following is the random forest formula [7]:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

where;

S = Data set
C = Number of classes
p_i = Class probability

To see the performance of a classification, including whether it is good or not, a confusion matrix calculation is needed. The following is the confusion matrix formula [8]:

$$Accuracy = \frac{TP + TN}{All} \times 100\% \tag{2}$$

$$Presicion = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

In this research using realtime earthquake data so that the amount of data is quite large, the data will be studied in accordance with the stages of the research.

3. Result and Discussion

After obtaining data and conducting research, the following results were obtained:

3.1. Preprocessing

Data preprocessing is done by selecting data fields that will be used to process the raw data into research-ready data. Researchers need descriptive statistics to summarize central tendencies, the spread of the dataset and help get a quick overview of the data set.

Table 2. Descriptive Statistics

Variable	Min	Max	Mean	Std. Dev
Magnitude	1.389	5.9	3.444	0.763
Depth (km)	5	648	49.533	72.061
Phase Count	6	402	38.307	37.868

From the Table 2 above, it is known that the average depth of earthquakes is around 49 km. For the depth of the most basic earthquake, 648 km occurred on August 3, 2023, located in Minahassa

Peninsula, Sulawesi. While the shallowest earthquake depth is 5 km which occurred on July 30, 2023 located in North Sumatra. It can be concluded that there are earthquakes that have the potential to cause a tsunami in Indonesia from July 2023 to August 2023.

3.2. Data visualization

Data visualization is a way to present data using graphical elements, such as charts, diagrams and maps. Data visualization aims to make it easier to understand patterns, trends and variable relationships in data. This is done to make it easier to draw conclusions and make decisions.

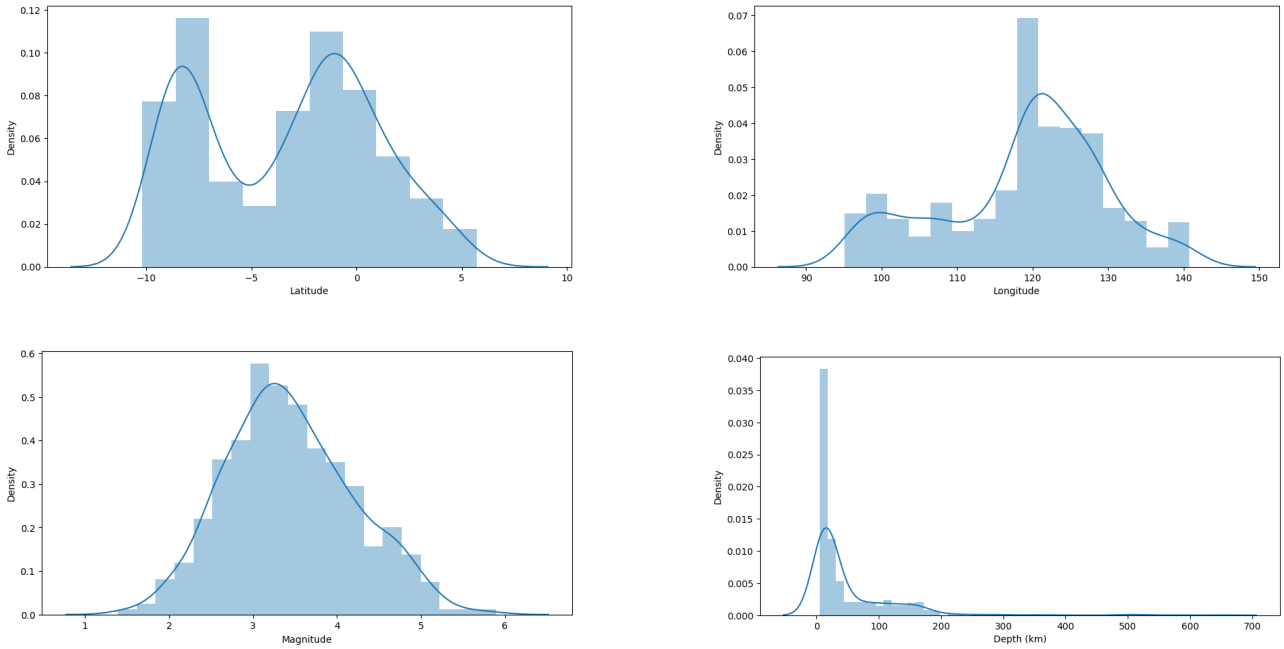

Figure 1. Data Visualization

Figure 1 show that data Visualization of latitude, longitude, magnitude and depth (km) variables. The bar structure shows the surface in blue while the line structure is marked in dark blue describing the change of variables over time.

3.3. Data Classification Results

The results of the earthquake classification are classified based on their depth, namely shallow earthquakes with a hypocenter of less than 70 km, medium earthquakes with a hypocenter between 70km and 300km, and deep earthquakes with a hypocenter of more than 300 km. In this study, researchers classified the types of earthquakes into two, namely earthquakes that have the potential to cause a tsunami and those that do not have the potential to cause a tsunami. Earthquakes that have the potential to cause a tsunami have a depth (hypocenter depth) that is located at a shallow depth, in the range of 0-70 km below sea level.

Table 3. Data Classification Results with Random Forest

	Latitude	Longitude	Magnitude	Dept (km)	Tsunami Category
0	2,324,009	126,793,556	3,916,701	10	1
1	-6,067,009	130,813,995	4,721,311	10	1
2	-4,680,160	104,608,658	3,032,803	169	0
3	0,859564	125,391,731	3,923,735	30	1
⋮	⋮	⋮	⋮	⋮	⋮
706	-0,064081	122,847,672	3,401,604	156	0
707	-4,516,113	102,331,856	3,353,651	30	1

The Table 3 above shows the grouping (classification) of earthquake data that had the potential to cause a tsunami and those that did not. Earthquake data is divided into two categories, namely, earthquakes with no tsunami potential indicated by number 0 and earthquakes with tsunami potential indicated by number 1. Classification is calculated using the random forest algorithm, based on the depth of the hypocenter (depth).

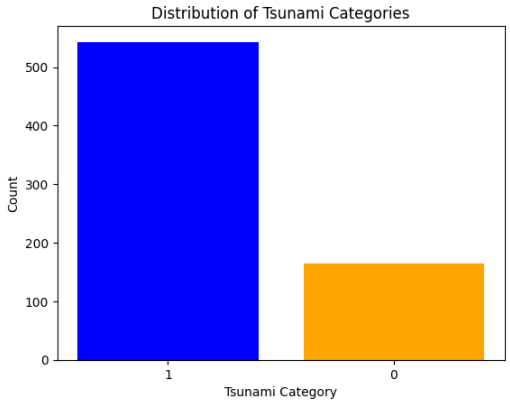

Figure 2. Earthquake Classification

Based on the bar chart on Figure 2 above, it can be concluded that there were more earthquakes with the potential to cause a tsunami than those without the potential to cause a tsunami. As a result of this study, the results for the category of earthquakes with no tsunami potential were 165, while for the category of earthquakes with tsunami potential, the results were 543.

The dataset from the Meteorology Climatology and Geophysics Agency (BMKG), will be divided into train data and test data before modeling with Random Forest, with a proportion of 70% for training data and 30% for test data. After that, feature scaling will be carried out, because there are some data with different scales.

Table 4. Training Data

Random Forest Training Result				
Classification Report				
	precision	recall	f1-score	Support
0	1	1	1	113
1	1	1	1	382
accuracy			1	495
macro avg	1	1	1	495

The Table 4 above shows the value of data using training data, training data or training data used to train the model. Based on the table above, the precision, recall, and f1-score values are 1.00 with an accuracy value of 1.00. Recall is a matrix that measures the model's ability to identify all positive instances, meaning that the model has recognized all positive instances in the dataset perfectly. If precision has a value of 1.00, then all positive predictions made by the model are correct. If recall and f1-score have a value of 1.00, it indicates that the model has perfect recall and precision. The model is able to recognize all positive instances and make near-perfect positive predictions with predictions that have minimum error.

Table 5. Testing Data

Random Forest Testing Result				
Classification Report				
	Precision	recall	f1-score	Support
0	1	0.88	0.94	52
1	0.96	1	0.98	161
accuracy			0.97	213
macro avg	0.98	0.94	0.96	213

The Table 5 above shows the value of data using testing data, testing data is used to test the performance of the model on the data. Based on the table above, the results are 1.00 for precision, 0.88 for recall, and 0.94 for f1-score with an accuracy of 0.97. If precision has a value of 1.00, then all positive predictions made by the model are correct. If the recall f1-score has a value of 0.88 and 0.94 or close to 1.00, it indicates that the model has perfect recall and precision. The model is able to recognize all positive instances and make near-perfect positive predictions with predictions that have minimum error.

Model evaluation is carried out to determine the accuracy of the model in classifying and predicting data, in this study using Confusion Matrix. Confusion Matrix is a table that shows the number of instances that are correctly or incorrectly classified based on the actual label and the predicted label of the model [9]. In evaluating this model, the data used is testing data [10].

Table 6. Confusion Matrix

Confusion Matrix	
True Positive	False Positive
46	6
21.60%	2.82%
False Negative	True Negative
0	161
0.00%	75.59%

Based on the Confusion Matrix above, using testing data with a total of 213 data. True Positive shows the amount of earthquake data that is predicted to have no tsunami potential is correctly classified as much as 46 data. The True Negative shows that the earthquake data predicted to have the potential for a tsunami was correctly classified as much as 161 data. Based on the confusion matrix above, it can be concluded that the amount of earthquake data that has the potential to cause a tsunami is more than the earthquake that does not have the potential to cause a tsunami, and the random forest method is quite well used for the classification of large datasets.

4. Conclusions

Based on the research conducted, the data used in the research is realtime data of earthquakes in Indonesia that occurred from July 2023 to August 2023. Researchers classified the earthquakes into two categories, namely earthquakes with tsunami potential and those without. These categories were calculated by looking at the depth of the hypocenter (depth), which has the potential to cause a tsunami if the depth is located at a shallow depth in the range of 0-70 km below sea level. Based on the results of the study, it was found that the number of earthquakes that did not have the potential to cause a tsunami was 165, while those with the potential to cause a tsunami were 543.

Classification of research data was carried out using the random forest algorithm, obtained an accuracy value of 0.97 or 97%. Based on the classification results, it can be concluded that the random forest algorithm can be used to classify earthquake data in Indonesia quite well, because it is able to handle large datasets. This research aims to find out the potential strength of earthquakes in the Indonesian region, so that the government and the general public are more aware of disasters. This algorithm has weaknesses, namely difficult interpretation and requires tuning the right model for the data. Therefore, for further research it is necessary to use other algorithms to compare the results.

References

[1] T. Duha, M. Laia, A. K. Huda, and A. Jasuma, "Klasifikasi data gempa bumi di pulau sumatera menggunakan algoritma naïve bayes," *Jurnal Informatika*, vol. 2, no. 1, pp. 23–27, 2023. https://doi.org/10.57094/ji.v2i1.840.

[2] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi potensial gempa bumi indonesia menggunakan metode random forest dan feature selection," *IDEALIS: Indonesia Journal Information System*, vol. 6, no. 2, pp. 83–89, 2023.

[3] D. Deswita, S. Yuliharni, and N. N. Efniyati, "Studi kasus: Gambaran kesiapsiagaan remaja menghadapi gempa bumi dan tsunami," *Jurnal'Aisyiyah Medika*, vol. 8, no. 2, 2023. https://doi.org/10.36729/jam.v8i2.1112.

[4] F. Mu'alim and R. Hidayati, "Implementasi metode random forest untuk penjurusan siswa di madrasah aliyah negeri sintang," *JUPITER: Jurnal Penelitian Ilmu dan Teknologi Komputer*, vol. 14, no. 1, pp. 116–125, 2022. https://doi.org/10.5281./4588/5.jupiter.2022.04.

[5] I. Ismail, "Klasifikasi area gempa bumi menggunakan algoritma random forest," *Jurnal Ilmiah Informatika Komputer*, vol. 26, no. 1, pp. 56–64, 2021. https://doi.org/10.35760/ik.2021.v26i1.3853.

[6] T. Tambunan, M. Yohanna, and A. P. Silalahi, "Penerapan metode random forest dalam mendeteksi berita hoax," *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, vol. 7, no. 2, pp. 301–306, 2023. https://doi.org/10.46880/jmika.Vol7No2.pp301-306.

[7] F. Fauzi, W. Setiayani, T. W. Utami, E. Yuliyanto, and I. W. Harmoko, "Comparison of random forest and naïve bayes classifier methods in sentiment analysis on climate change issue," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1439–1448, 2023. https://doi.org/10.30598/barekengvol17iss3pp1439-1448.

[8] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi random forest untuk klasifikasi motif songket palembang berdasarkan sift," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, pp. 310–320, 2020. https://doi.org/10.35957/jatisi.v7i2.289.

[9] O. W. Yuda, D. Tuti, *et al.*, "Penerapan penerapan data mining untuk klasifikasi kelulusan mahasiswa tepat waktu menggunakan metode random forest," *SATIN-Sains Dan Teknologi Informasi*, vol. 8, no. 2, pp. 122–131, 2022. https://doi.org/10.33372/stn.v8i2.885.

[10] S. Amaliah, M. Nusrang, and A. Aswi, "Penerapan metode random forest untuk klasifikasi varian minuman kopi di kedai kopi konijiwa bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, 2022. https://doi.org/10.35580/variansiunm31.