An Approach for Earthquake Magnitude Prediction in Sumatra Using Linear Regression and Random Forest Algorithms

Iktri Madrinovella*, Clara Agnatasia Br Matondang, Maria Octaviana Moi

Department of Geophysical Engineering, Universitas Pertamina, Jakarta 12220, Indonesia
*Corresponding Author: iktri.madrinovella@universitaspertamina.ac.id
Received: 8 Juli 2024
Accepted: 24 Maret 2025

Abstracts

Sumatra is one of the largest islands in Indonesia and is highly susceptible to seismic activity due to its location between the Indo-Australian and Eurasian tectonic plates. Many significant earthquakes have occurred in Sumatra, resulting in substantial losses and damage. Despite advances in seismic monitoring technology and methods, achieving high accuracy in earthquake prediction remains a major challenge. This study presents a machine learning approach to predicting earthquake magnitudes in Sumatra using Random Forest and Linear Regression algorithms. Utilizing a comprehensive dataset from the United States Geological Survey (USGS) covering seismic events in Sumatra from July 1960 to April 2024, this research aims to improve the accuracy of magnitude predictions. This dataset includes variables such as time, location, depth, latitude, longitude, and magnitude. The study employs Python libraries for data visualization and correlation analysis. Visualizations show a strong correlation between predicted and actual earthquake magnitudes, mapping seismic events along well-known fault lines such as the Sunda Megathrust and Sumatran Fault. The results of this study demonstrate the potential of machine learning models to enhance seismic hazard assessment strategies, significantly contributing to more effective disaster management and mitigation practices in earthquake-prone areas.

Keywords: Earthquake, magnitude predictions, random forest, linear regression

Introduction

Sumatra is one of the largest islands in Indonesia, being known as the island that borders the Indo-Australian plate to the west and the Eurasian plate to the east. The two plates meet on the ocean floor, which is called the subducted slab or the subduction zone of the oceanic Indo-Australian plate to the continental Eurasian plate. This boundary expands to the Java island, which is known as the Sumatra or Sunda megathrust, or the largest thrust fault mechanism in Indonesia. Many large earthquakes ever occurred in Sumatra, for example the Sumatra-Andaman earthquake (Desember 26th, 2004, Mw = 9.1), the Nias-Simeulue earthquake (March 3rd, 2005, Mw = 8.7), the twin Bengkulu earthquakes (September 12th, 2007, Mw = 7.9 and 8.4). and the Enggano earthquake (June 4th, 2000, Mw = 7.9). The historical data also shows the large earthquakes in the center part of Sumatra (1797, 1833, 1861, 1935). Those earthquakes mostly were triggered by the activities of the megathrust [1][14].

Along with the subduction zone, the volcanic arc (a belt of volcanoes) is formed above the subducting oceanic plate. The area between the subduction zone and the volcanic arc is known as the forearc basin, and the rear area of the volcanic arc is called the backarc basin. The fault is formed between the forearc basin and the volcanic arc because of the friction in the continental plate. The fault in Sumatra is 1900 km-length from Aceh to Lampung (the Sumatran Fault System) and very segmented. Many earthquakes were also triggered by the

activities of the Sumatran Fault System, for example the Padang earthquake (September 30th, 2009, Mw = 7.9), the double Lake Singkarak earthquakes (March 6th, 2007, Mw = 6.4), and the Pasaman earthquake (February 25th, 2022, Mw = 6.2)

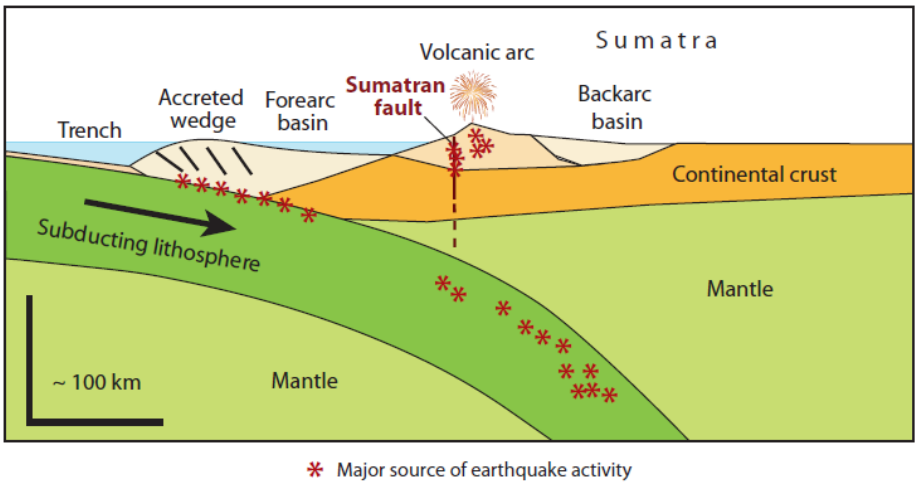

Figure 1. The schematic cross section of the Sumatra plate boundary [2]

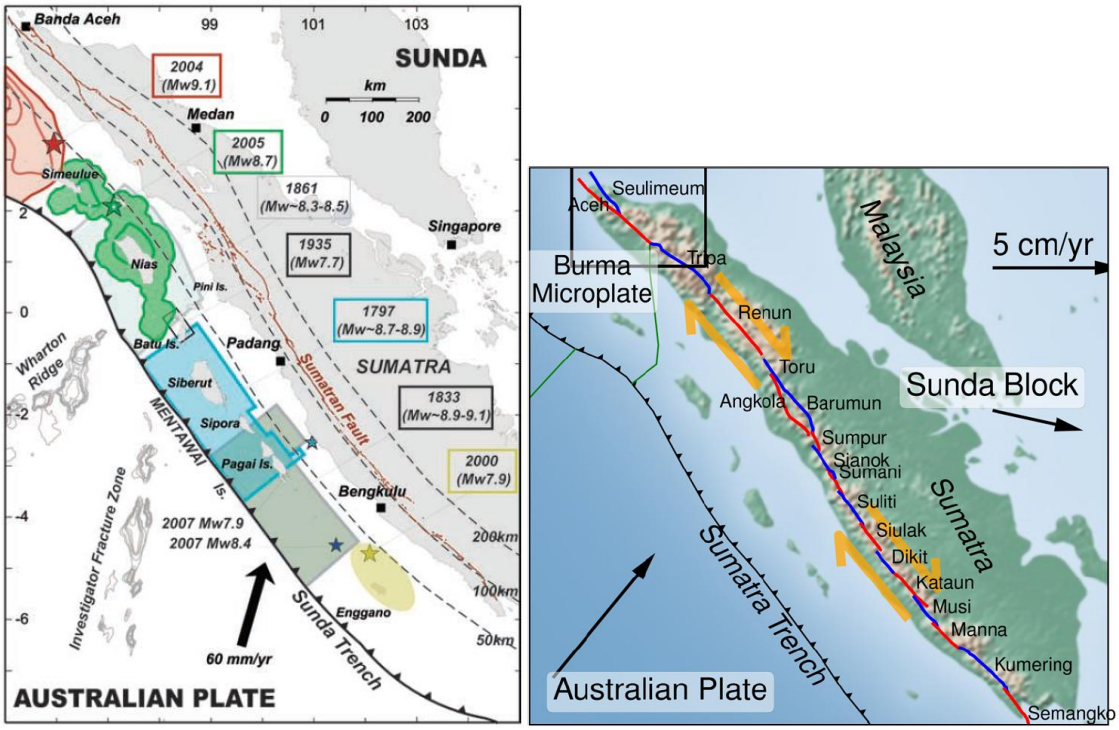

Figure 2. (Left) The active structural elements and historical great earthquakes in Sumatra [1].
The colored boxes are the rupture (released energy) area affected by the nearest large earthquakes [3].
(Right) The major segments of the Sumatran Fault System or SFS. The SFS is a large right-lateral strike-slip fault

Magnitude is the size of the earthquake to represent how strong the earthquake is. It is calculated using the distance between the earthquake and the station (the epicentral distance), the seismogram amplitude shown in the recording station [4]. People usually think that the earthquake magnitude unit is Richters scale. However, Richter calculated the magnitude based on the earthquake data in California (the local magnitude or M_L), but the formula is still relatable to be used in other regions. The best way to calculate the magnitude is using the seismic moment, based on the displacement and the rupture area

affected by the earthquake. The magnitude derived by the seismic moment is called the moment magnitude (M_w).

$$M_L = \log A - \log A_o\,(\delta) \qquad (1)$$

The formula of the Richter (local) magnitude (M_L), where A is the maximum excursion of the seismograph, and the empirical function A_o depends on the epicentral distance of the station δ.

$$M_W = \frac{2}{3}\log M_o - 10.7 \qquad (2)$$

The formula of the moment magnitude (M_W), where M_o is the seismic moment (in dyne.cm or 10^{-7} N.m) [5].

The recent challenge to compute the earthquake magnitude in a region is a "what-if" scenario of the limited data. What if we do not have the data of the station location or the rupture area? Can we predict the earthquake magnitude at one spot if we use some known data? Machine learning can be used to answer those questions. In this study, we will use some of the simple machine learning techniques to predict the earthquake magnitude in Sumatra by using the earthquake catalog from 1960 – 2024. We use the linear regression and the random forest method in this study.

Linear regression method predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values. Random forest is a combination of tree predictors that each tree depends on the values for all trees in the forest [6].

Material and Methods

In this study, we use the dataset from the United States Geological Survey (USGS) data in Sumatra (longitude between 95°E to 108°E and latitude between 6°N to 8°S) in the period between July 1960 – April 2024, total 2558 events with magnitude more than 5. The data was generated using the open-source Obspy in Python. The visualization was also plotted using PyGMT (Generic Mapping Tools) coded in Jupyter Notebook.

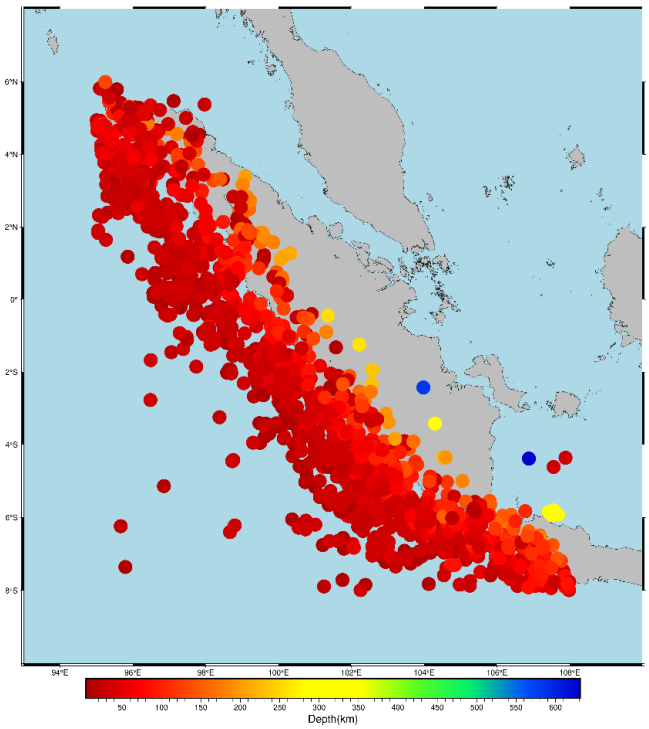

Figure 3. The USGS dataset of Sumatra earthquakes in 1960 – 2024 (2558 events)

	ID	year	month	day	hour	minute	second	lon	lat	depth	mag
0	0	2024	4	7	17	46	18.442	103.9286	-6.5495	38.777	5.00
1	1	2024	3	23	13	20	28.536	102.7307	-4.5597	62.692	5.30
2	2	2024	3	20	8	24	57.931	100.4628	-2.0189	67.702	5.00
3	3	2024	3	4	2	48	21.113	102.1790	-5.8279	10.000	5.20
4	4	2024	3	3	15	7	43.246	103.0276	-5.0063	63.387	5.10
...	...	...	...	...	...	...	...	...	...	...	...
2553	2553	1962	3	28	4	5	19.780	97.3000	1.2700	25.000	5.58
2554	2554	1961	10	26	15	27	7.470	98.7040	-0.3970	28.900	6.18
2555	2555	1961	6	24	9	35	54.830	97.3380	3.6710	20.000	5.67
2556	2556	1960	10	3	19	50	49.470	103.0690	-5.7140	30.000	5.84
2557	2557	1960	7	10	0	5	26.770	98.0980	0.4040	30.000	6.28

2558 rows × 11 columns

Figure 4. The USGS catalogue of Sumatra earthquakes in 1960 – 2024

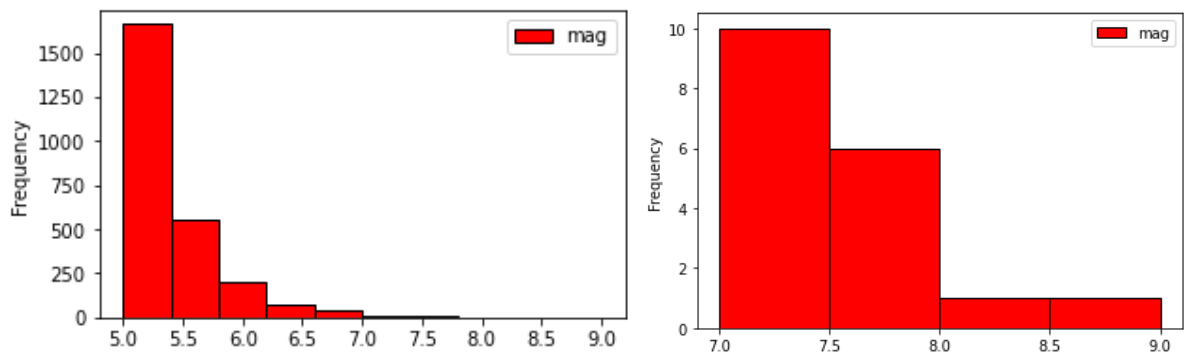

Figure 5. The USGS catalogue of Sumatra earthquakes in 1960 – 2024 has 2558 events with magnitude ≥ 5 (left), particularly very few events with magnitude ≥ 7 (right)

Machine learning is a method of data analysis that automates analytical model building. A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems. The data to be subjected to a pattern in the learning phase can be arrays composed by a single value per element, or by a multivariate value [7]. These values are often referred to as features or attributes. This study employs machine learning techniques to predict earthquake magnitudes in the Sumatra region, aiming to enhance seismic hazard assessment and risk mitigation.

The dataset used in this study is derived from the Sumatra earthquake catalogue (USGS, 1960 -2024), which includes seismic event records with geographic coordinates (longitude and latitude), depth, and magnitude. The data were divided into training and testing sets, with 70% of the data used for training the model and 30% for testing its performance. This split ensures that the model is trained on a significant portion of the data while being evaluated on a separate subset to test its predictive capabilities [8][9].

Linear regression is a method that predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values. Linear regression analysis can be divided into simple linear regression and multiple linear regression. Simple Linear Regression is a statistical method used to model the relationship between two variables by fitting

a linear equation to observed data. This method estimates the slope and intercept of the line of best fit, which represents the relationship between the independent variable $\bar{x}$ and the dependent variable $\bar{y}$. The slope m indicates the change in the dependent variable for each unit change in the independent variable, while the intercept c represents the predicted value of the dependent variable when the independent variable is zero.

$$\bar{y} = m\bar{x} + c \quad (3)$$

A simple linear regression model can also be expressed as follows

$$y = a_o + a_1 x + e \quad (4)$$

y the dependent variable and x the independent variable. a_0 the intercept of the regression line on the vertical axis and a_1 the slope of the regression line, e the random error which will be used to express the effect of random factors on dependent variable [10].

However, in this study, we used multiple linear regression by incorporating three key parameters, namely longitude, latitude, and depth, to predict the magnitude of a phenomenon. Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. The assumptions in multiple linear regression include a linear relationship between the dependent variable and each of the independent variables, which can be checked using a scatterplot, and the requirement that the independent variables are not highly correlated with each other.

$$\bar{y} = c + m_1\bar{x}_1 + m_2\bar{x}_2 + \ldots + m_n\bar{x}_n \quad (5)$$

The random forest method, pioneered by Leo Breiman and Adele Cutler, is an ensemble learning technique that constructs multiple decision trees during training. It combines the predictions of these trees to output the mode of the classes for classification tasks or the mean prediction for regression tasks. This approach mitigates overfitting by averaging predictions from diverse trees trained on different subsets of the dataset. Key methods used to ensure tree diversity in random forests include bootstrap aggregating (bagging), where subsets of the training data are sampled with replacement to train individual trees, and feature randomness, which limits the number of features each tree can consider for splitting. Random forests typically involve tuning parameters such as node size, number of trees, and number of features sampled. In this study, the random forest algorithm was applied to predict earthquake magnitudes, achieving robust predictions by training on 70% of the data and testing on the remaining 30%. The methods ability to generate uncorrelated trees enhances prediction reliability and robustness [11][12]. A random forest combines predictor trees in such a way that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The random forest algorithm has demonstrated strong utility in many fields of study, including predicting earthquakes and big data application [7][11].

We use Python and the Scikit-Learn library for implementing the machine learning models, which provides efficient tools for data mining and data analysis, including a range of supervised and unsupervised learning algorithms [13]. Key steps in the implementation process include data preprocessing (normalizing features and handling missing data), model training (splitting data into training and testing sets and training the models), and model evaluation (assessing performance using metrics such as R-squared (R^2) and Root Mean Squared Error (RMSE) to understand accuracy and prediction errors).

$$R^2 = \frac{1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (7)$$

y_i is the actual value, $\hat{y}$ is the predicted value, $\bar{y}$ is the average of the actual values and n is the number of observations.

Scikit-Learns robust and flexible framework facilitates the implementation and evaluation of both linear regression and random forest models, enabling accurate earthquake magnitude predictions and effective performance assessment. By employing multiple linear regression and random forest methods, this study aims to predict earthquake magnitudes in the Sumatran region with high accuracy. The combination of these machine learning techniques and comprehensive data analysis contributes to a better understanding of seismic activities and aids in disaster preparedness and mitigation efforts. The use of Scikit-Learn in Python provides a robust and flexible framework for model implementation and evaluation, highlighting the importance of integrating detailed geological and geophysical data to enhance predictive models [7][15].

The earthquake magnitude prediction using machine learning, especially random forest algorithms has been successfully tested using 256 Padang earthquakes with magnitude around 4 (2008-2003) [16]. Another research was also conducted using The Flores Sea Earthquake, and it concluded that random forest methods increase accuracy and overcome the overfitting problem that occurs in machine learning models [17]. Some other machine learning techniques were also applied to predict the North Zagros earthquake, and the results of random forest algorithm showed good accuracy (83.25% - 93.48%) at certain magnitude threshold (5 – 5.5) [18]. This algorithm was also implemented to predict the earthquake magnitude in the Matano fault, and the prediction model has outstanding performance, with a high degree of accuracy in predicting values that quite match the actual values [19].

Not only in predicting earthquake magnitude, this algorithm was also conducted to classify earthquake area using 23,412 world earthquake data (1965-2016), and concluded the satisfying accuracy 99.97% [20].

Results and Discussion

Feature Correlation Analysis

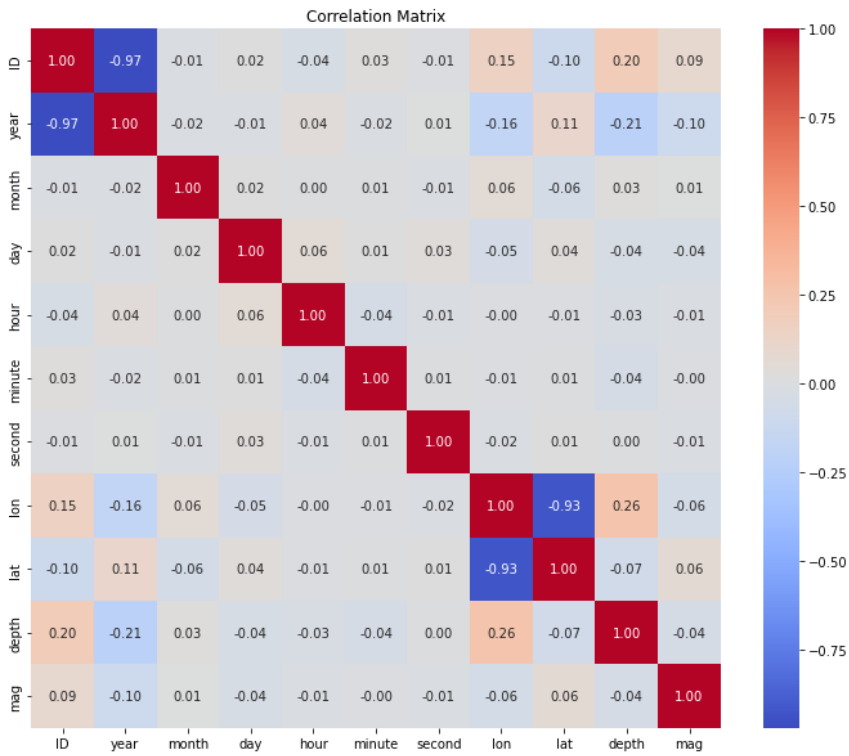

Figure 6. The correlation matrix

The correlation matrix reveals key insights into the dataset used for earthquake magnitude prediction using a Random Forest machine learning model. Strong positive correlation is

found between latitude and longitude (0.97) due to the megathrust location or seismic activities in the trend of Northwest – Southeast (NW-SE) in Sumatra, and it does not necessarily impact the prediction. From the correlation matrix, it can be concluded that the earthquake data with date, time, location features are not correlated to magnitude. These correlations emphasize that it is statistically difficult to predict earthquake magnitudes based on limited features. However, to improve the prediction using geological data is a complicated challenge that must be conducted by advanced modelling. The linear regression and random forest algorithms are expected to give modest but robust earthquake prediction.

Data Visualization

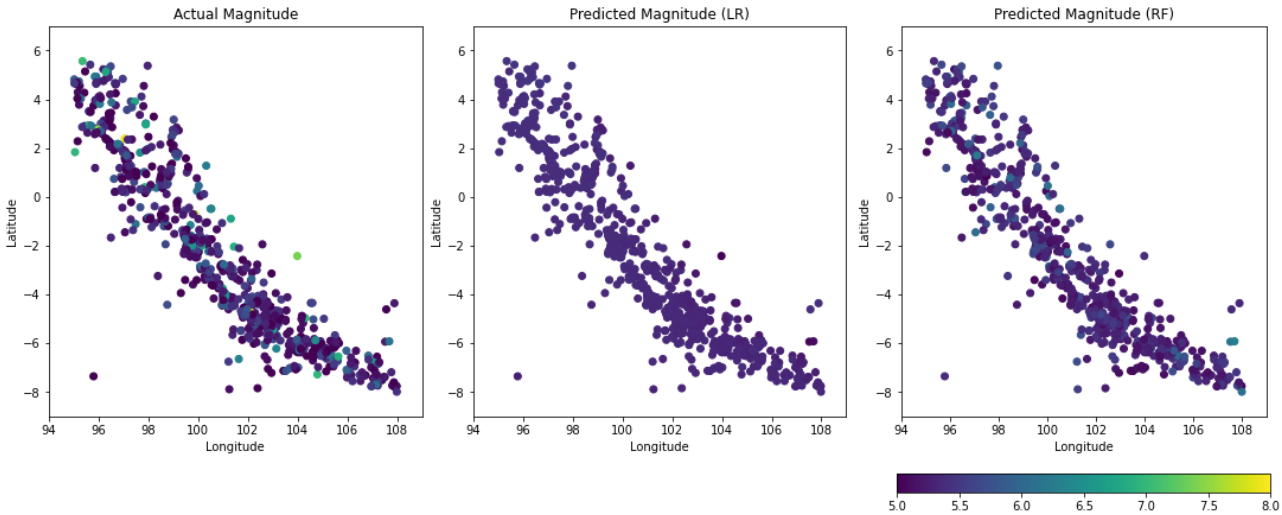

Figure 7. The actual magnitude (left) vs predicted magnitude using Linear Regression (middle) vs predicted magnitude using Random Forest (right)

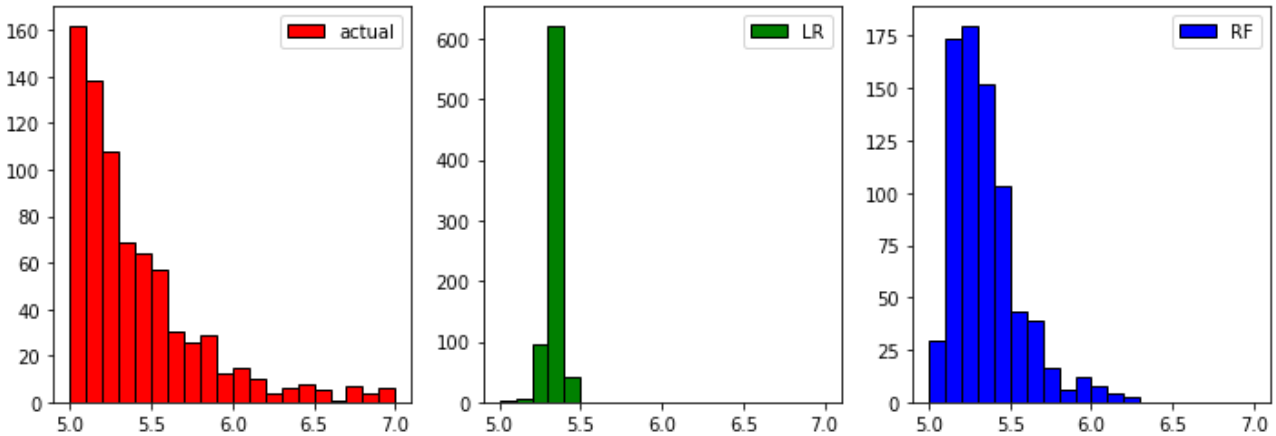

Figure 8. The histogram of magnitude and the frequency of earthquakes from actual data (left), predicted data using linear regression (middle) and predicted data using random forest (right) in the magnitude range between 5 to 7

Figure 7 shows the actual magnitude of earthquakes in Sumatra (USGS, 1960-2024), and the predicted magnitude using Linear Regression (middle) and Random Forest (right) algorithms. Figure 8 shows the relationship between magnitude and the frequency of earthquakes in actual and predicted data. The actual magnitude ranges between 5.0 – 8.0, but the predicted magnitude ranges between 5.0 – 5.5 (LR) and 5.0 – 6.3 (RF). This pattern aligns with the tectonic activity along the Sunda Megathrust, the Sumatran Fault, and the

Mentawai Fault system. The linear regression algorithm is only stable for smaller magnitudes, but it tends to underestimate the value of greater earthquakes. The random forest algorithm has slightly different result, it can predict the larger magnitudes than the prediction using linear regression. Both algorithms are unstable to predict the earthquakes with magnitude larger than 7.

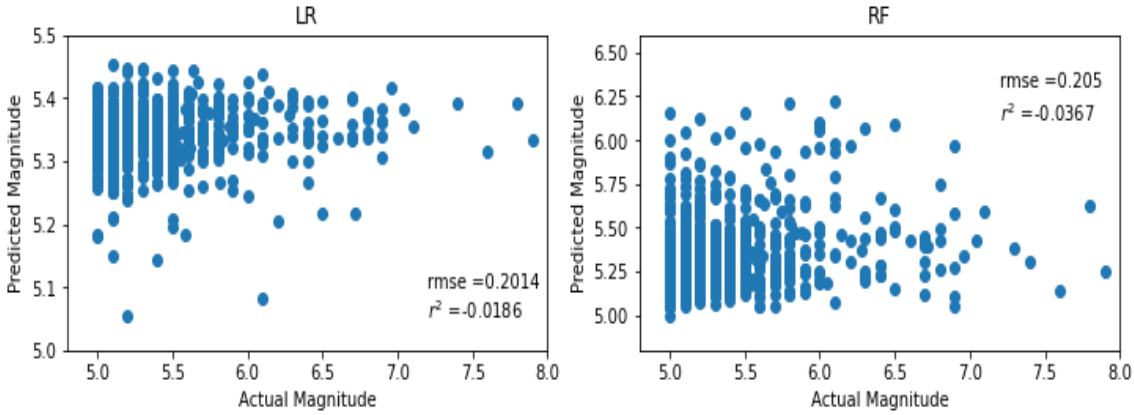

Figure 9. The relationship between actual and predicted magnitudes

Figure 9 depicts the relationship between actual and predicted earthquake magnitudes using data from the Sumatra earthquake catalogue. Each data point represents an earthquake, plotted with its actual magnitude on the x-axis and its predicted magnitude on the y-axis. Ideally, perfect predictions would align along a 45-degree diagonal line, indicating exact correspondence between predicted and actual magnitudes. The values of predicted magnitudes have similar patterns with the actual magnitudes in ranges 5 to 6. Those predictions are unstable to be used to predict the large earthquakes, especially earthquakes with magnitude > 7 (total 19 events) due to the very few data compared to the smaller magnitudes.

The values of RMSE are similar between those algorithms, but the values of R^2 are slightly different, which the R^2 value of random forest algorithm is greater than the linear regression algorithm. Based on the RMSE and R^2 values, the random forest algorithm gives more robust results than the linear regression algorithm.

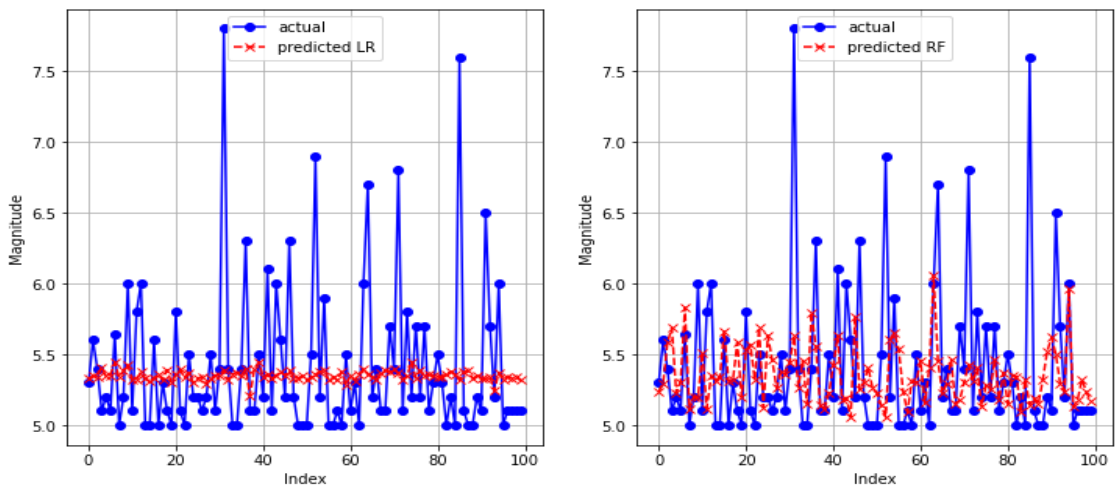

Figure 10. The comparison of actual and predicted magnitude (indices 0-100), between actual and predicted using linear regression (left) and between actual and predicted using random forest (right)

The visualization compares actual and predicted earthquake magnitudes for Sumatra, indexed from 0 to 100. Actual magnitudes (blue circles, solid line) and predicted magnitudes (red crosses, dashed line) show the random forest models are more effective in forecasting seismic events than the linear regression models. Minor discrepancies exist, but the model closely approximates actual magnitudes, reflecting its strong performance in modeling seismic patterns. Integrating detailed geological data could further enhance accuracy, underscoring the value of multidisciplinary approaches in seismic prediction.

Prediction using Actual Data

Linear regression and random forest algorithms are conducted using USGS data from July 1960 to April 2024. To verify the accuracy of the prediction, we test the algorithm using earthquakes data in Sumatra from May to June 2024. We use data derived from The Meteorological, Climatological and Geophysical Agency (BMKG) Indonesia.

Table 1. The actual vs predicted magnitude using linear regression and random forest (forecasting)

Earthquake Time and Location	Actual Magnitude	Predicted Magnitude (Linear Regression)	Predicted Magnitude (Random Forest)
25 May 2024 07:10:11 (Local Time) 94.73°E 3.66°N, 10 km (Calang-Acehjaya)	5.0	5.84	5.22
28 May 2024 18:52:38 (Local Time) 95.41°E 2.76°N, 12 km (Sinabang-Aceh)	6.2	5.84	5.64
5 June 2024 09:20:28 (Local Time) 100.54°E 3.76°S, 34 km (Mukomuko-Bengkulu)	5.7	5.82	5.33
5 June 2024 18:16:14 (Local Time) 98.5°E 0.42°N, 57 km (South Nias-North Sumatra)	5.7	5.81	5.66

The predicted magnitudes using linear regression have similar numbers due to limited capability. On the other hand, the predicted magnitudes using random forest are closer to the number of actual magnitudes.

Conclusion

This study establishes the effectiveness of machine learning models, specifically Random Forest and Linear Regression, in forecasting earthquake magnitudes in the Sumatran region. The Random Forest model demonstrates substantial proficiency in estimating earthquake magnitudes, achieving an R^2 value of 0.0367 and a Root Mean Squared Error (RMSE) of 0.205, thereby capturing the spatial distribution of seismic events more effectively than the Linear Regression model. Both models perform admirably to be used to predict earthquakes with magnitude 5 to 6, despite minor inconsistencies in predicting higher magnitude events.

Despite their strengths, the models exhibit limitations in accurately predicting the highest magnitude events, suggesting the necessity for further refinement and the incorporation of additional predictive variables. The integration

of detailed geological and geophysical data could enhance the models' accuracy, particularly for greater-magnitude earthquakes. This recommendation aligns with the multidisciplinary approach advocated by various studies to improve seismic hazard assessment and risk mitigation.

The successful application of these machine learning models for earthquake magnitude prediction in Sumatra underscores their potential utility for seismic hazard assessment and disaster preparedness. By providing more precise predictions of earthquake magnitudes, these models can inform risk mitigation strategies, ultimately contributing to improved disaster management and resilience in tectonically active regions. Random Forest and Linear Regression models prove to be valuable tools for predicting earthquake magnitudes in Sumatra, and their integration with comprehensive geological data holds promise for further advancements in predictive accuracy, enhancing our understanding and management of seismic risks in this vulnerable region.

References

[1] Chlieh, M., Avouac, J.P., Sieh, K., Natawidjaja, D.H., Galetzka, J., Heterogeneous coupling of the Sumatran megathrust constrained by geodetic and paleogeodetic measurements, *J Geophys Res Solid Earth*, Vol. 113, No. 5, May 2008, doi: 10.1029/2007JB004981.

[2] McCaffrey, R., The tectonic framework of the sumatran subduction zone, *Annual Review of Earth and Planetary Sciences*, Vol. 37. pp. 345–366, May 2009. doi: 10.1146/annurev.earth.031208.100212.

[3] Omang, A., Cummins, P., Robinson, D., Hidayati, S., Sensitivity analysis for probabilistic seismic hazard analysis (PSHA) in the Aceh Fault Segment, Indonesia, in *Geological Society Special Publication*, Vol. 441, No. 1, Geological Society of London, 2017, pp. 121–131. doi: 10.1144/SP441.5.

[4] Richter, C.F., An Instrumental Earthquake Magnitude Scale, *Bulletin of the Seismological Society of America*, Vol.25, No.1, 1935.

[5] Kanamori, H., The energy release in great earthquakes, *J Geophys Res*, Vol. 82, No. 20, pp. 2981–2987, Jul. 1977, doi: 10.1029/jb082i020p02981.

[6] Breiman, L., Random Forests, Machine Learning, 2001. 45(1), 5-32.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011, 12(85): 2825-2830.

[8] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J., Linear Regression, 2023, pp. 69–134. doi: 10.1007/978-3-031-38747-0_3.

[9] Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer Series in Statistics, 2009. doi.org/10.1007/978-0-387-84858-7

[10] Rong, S., Bao-Wen, Z., The research of regression model in machine learning field, in *MATEC Web of Conferences*, EDP Sciences, Jul. 2018. doi: 10.1051/matecconf/201817601033.

[11] Liaw, A., Wiener, M., Classification and Regression by RandomForest, R News, 2002, 2(3): 18-22.

[12] Rigatti, S.J., Random Forest, Journal of Insurance Medicine, 2017, 47(1):31-39. doi: 10.17849/insm-47-01-31-39.1

[13] Stein, S., Okal, E.A., Ultralong period seismic study of the December 2004 Indian Ocean earthquake and implications for regional tectonics and the subduction process.

[14] Sieh, K., Natawidjaja, D.H., Neotectonics of the Sumatran fault, Indonesia, Journal of Geophysical Research Solid Earth, Vol. 105, No. B12, pp. 28295–28326, Dec. 2000, doi: 10.1029/2000jb900120.

[15] Budiman, K., Ifriza, Y.N., Analysis of earthquake forecasting using random forest, *Journal of Soft Computing Exploration,* Vol. 2, No. 2, pp. 153-162, 2021, doi: 10.52465/joscex.v2i2.51

[16] Fauzan, A., Ahmad, D., Analisis Hasil Prediksi Magnitudo Gempa di Wilayah Kota Padang Menggunakan Teknik Random Forest, Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, Vol. 4, No. 3, pp. 1569-1576, 2023, doi: 10.46306/lb.v4i3

[17] Jufriansah, A., Khusnani, A., Saputra, S., Wahab, D.S., Forecasting the Magnitude Category Based on The Flores Sea Earthquake, Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), Vol. 7, No. 6, 2023, pp. 1439-1447. doi: 10.29207/resti.v7i6.1495

[18] Ommi, S., Hashemi, M., Machine learning technique in the north zagros earthquake prediction, Applied Computing and Geosciences, Vol. 22, 2024, doi: 10.1016/j.acags.2024.100163

[19] Madona, M., Rosid, M.S., Handoko, D., Riama, N.F., Saiful, D., Implementation of the Random Forest Algorithm for Magnitude Prediction in the Matano Fault Indonesia, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-4807, https://doi.org/10.5194/egusphere-egu24-4807, 2024.

[20] Ismail, I., Klasifikasi Area Gempa Bumi Menggunakan Algoritma Random Forest, Jurnal Ilmiah Informatika Komputer, Vol. 26, No. 1, 2021, pp. 56-64.