**TELKOMNIKA Telecommunication, Computing, Electronics and Control**
Vol. 18, No. 3, June 2020, pp. 1331∼1342
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i3.14756 

# Comparison of machine learning performance for earthquake prediction in Indonesia using 30 years historical data

**I Made Murwantara[1], Pujianto Yugopuspito[2], Rickhen Hermawan[3]**
[1,2]Informatics Department, Faculty of Computer Science, Universitas Pelita Harapan, Indonesia
[3]Undergraduate Program, Informatics Department, Faculty of Computer Science, Universitas Pelita Harapan, Indonesia

**Article Info**

***Article history:***

Received Aug 15, 2019
Revised Jan 13, 2020
Accepted Feb 24, 2020

***Keywords:***

Big data
Earthquake
Machine learning
Multinomial logistic regression
Naïve bayes
Prediction
SVM

**ABSTRACT**

Indonesia resides on most earthquake region with more than 100 active volcanoes, and high number of seismic activities per year. In order to reduce the casualty, some method to predict earthquake have been developed to estimate the seismic movement. However, most prediction use only short term of historical data to predict the incoming earthquake, which has limitation on model performance. This work uses medium to long term earthquake historical data that were collected from 2 local government bodies and 8 legitimate international sources. We make an estimation of a medium-to-long term prediction via machine learning algorithms, which are multinomial logistic regression, support vector machine and Naïve Bayes, and compares their performance. This work shows that the support vector machine outperforms other method. We compare the root mean square error computation results that lead us into how concentrated data is around the line of best fit, where the multinomial logistic regression is 0.777, Naïve Bayes is 0.922 and support vector machine is 0.751. In predicting future earthquake, support vector machine outperforms other two methods that produce significant distance and magnitude to current earthquake report.



***Corresponding Author:***

I Made Murwantara,
Informatics Department,
Faculty of Computer Science, Universitas Pelita Harapan,
Tangerang, Banten, 15811, Indonesia,
Email: made.murwantara@uph.edu

## 1. INTRODUCTION

An earthquake is a natural disaster that occurs as a result of rocks layer movement or displacement of the earth tectonic plate. This precipitous movement releases a huge amount of energy that creates a kind of seismic waves. The vibration results that passed through the earth surface caused damage for the population that lives on the earthquake impact areas. Indonesia with more than 300 million inhabitants is a country located in the most frequent earthquake region as it has about 127 active volcanoes [1], which usually called the Ring of Fire area that become the most active tectonic movement. Moreover, Indonesia also has the Great Sumatran Fault that span 1900 km length and the Banda Sea convergent flat margin that creates even more seismic activities [2, 3].

Nowadays, the earthquake warning system already installed in many remote and volcanic areas that might increases the number survivor expectation. Moreover, many research outcomes also gain more information about earthquake characteristics and impacts to the surrounding area. machine learning has also been used to make advancement on the information and prediction results. However, some machine learning work result still has not provided accurate prediction, and sometimes rise up a false alarm because of lack of the volume of data or the prediction method [4]. In our knowledge, the application of the earthquake prediction still has a space for us to augment into a certain point that gives us more confidence and better results. Furthermore, a good and reasonable prediction will provide opportunities to manage the emergency route path for evacuation which may reduce the casualties.

In order to provide data for prediction, we utilize the data collection from several earthquake and seismological repositories. The list of data resources for our research as follows, the United States Geological Survey (USGS) [5], Incorporated Research Institution for Seismology(IRIS) [6], National Oceanic and Atmospheric Administration (NOAA) [7], European-Mediteranian Seismological Centre (EMSC) [8], International Seismological Centre (ISC) [9], Istituto Nazionale di Geofisica e Vulcanologia (INGV) [10], GeoForschungZentrum (GFZ) [11, 12], Indonesia Tsunami Early Warning System (InaTEWS) [13], Global Historical Earthquake Archive(GHEA) [14, 15], and Badan Meteorologi, Klimatology dan Geofisika (BMKG) Indonesia [16]. The volume of the data collection produces more than 1TB. After cleansing to have only data within Indonesia region, we have around 375 GB data which is used as training and testing data. Considering the volume of data, this work is a Big Data research.

In this work, we compare the performance of three machine learning approaches, which are multinomial logistic regression [17, 18], Naïve Bayes [2, 19–21] and support vector machine (SVM) [4, 22–25] to the earthquake data. Where, Logistic Regression provides information of relationship between variant and to find out how close is one or more variable to another one. Naïve Bayes approach allows us to compute the probability that is taken from new information. SVM is used for classification and regression analysis of separation hyperplane. The contribution of this paper is twofold:

(a) In predicting a disaster such as earthquake, a comparison between different machine learning algorithms may give light for a new approach. We propose a technique that is comparable to other approach for earthquake prediction in Indonesia region. Our method facilitates of prediction and visualization that range within 50 years of seismic historical data which is particularly helpful to classify of how different machine learning performance could put light on our method of prediction. To this, our approach can also adjust the size of data for better prediction. This is useful since the size of data, sometimes, influence the training and testing process for ultimate prediction. Other than that, we have flexibility on testing our results.

(b) The data collection and cleansing includes massive volume of data which creates rich resources for prediction. We collect the data from legitimate organization all over the world that compares with the local monitoring by the government bodies in Indonesia. The data cleansing also takes most of our time which is not only retrieve raw data, it is also through web scrapping and data transformation. Some information need to be inspected carefully, as the monitoring data may be irrelevant for our work. To this, we analyze the data based on whether the location of monitoring and its data relevant. For example, the earthquake data that released by a resource that taken from third party or not primarily generated by a specific seismic monitoring station.

## 2. RESEARCH METHOD

### 2.1. Relevant works

The improvement of earthquake prediction has been utilized via historical seismic data. The most promising technique is to use the Artificial Intelligence (AI) and machine learning (ML) has gained further knowledge [26]. In [27], Bertrand et al. identify the possibility of upcoming earthquake by forecasting the laboratory quake cycle, which reveals the timing of the event will probably occurs.In general, earthquake prediction is categorized into three different terms that is based on the length of the historical data source. Short term earthquake prediction needs a precursor to strengthen its accuracy [28], while intermediate and long term prediction makes estimation on statistical probability approach. Syifa et al. [29] uses SVM to analyze post earthquake situation to assess the distribution of seismic destruction, which can be useful for evacuation and mitigation plan. Another technique to address the prediction of earthquake uses the meteorological data [30]

based on the particle filter-based and support vector regression. This technique obtained natural information, such as air temperature, gas concentration and wind speed to estimate the precursor of earthquake.

### 2.2. Background

This section will discuss the background theory of the work that covers the earthquake theory and machine learning approaches. The earthquake background theory is categorized into earthquake types, seismic wave and earthquake phenomena in Indonesia. The machine learning covers the multinomial logistic regression, Naïve Bayes and support vector machine.

#### 2.2.1. Earthquake

An earthquake is a natural disaster that creates tremor or vibration in the impacted area as a result of earth rocks layers movement or displacement because of the tectonic dislocation. This vibration will reach the earth surface that causes massive destruction. There are four types of earthquake, which are tectonic, volcanic, collapse and explosion. As shown in Figure 1, three types of of surface movement that caused an earthquake that appears not on every place in the earth. In general, the movement of earth surface as the cause of an earthquake when (a) two plates moves away to different direction, (b) two plates move in to the same point of line and (c) these plates move side-by-side on opposite direction.

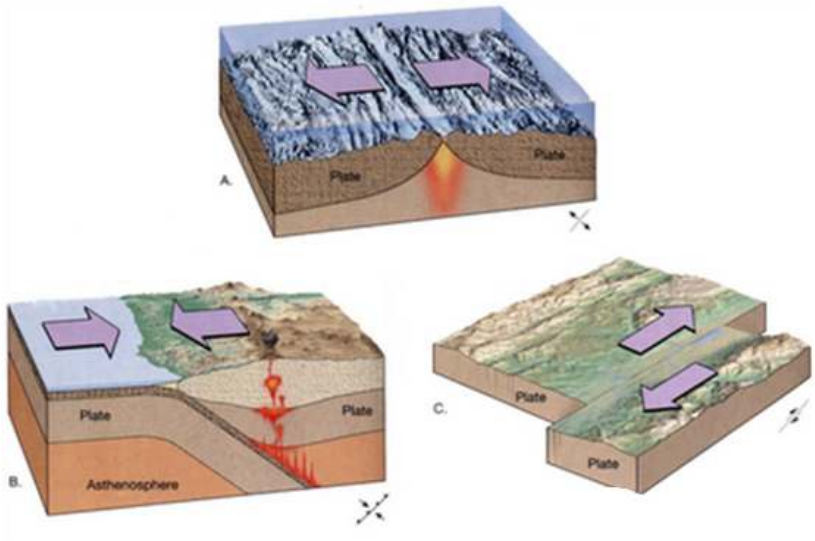



Figure 1. Earthquake types (a) divergent, (b) convergent and (c) transform

The layer of earth skin has high temperature that distributes its heath into surrounding area. In general, this volcanicactivityknown as the heath flow convection. This kind of activity pushes the magma into the surface which creates volcanoe. Indonesia is an archipelago that located in the Circum-Pacific and Mediteranian which has a lot of numbers of active volcanoes. To this, Indonesia becomes one of the high risk countries on earthquake disaster. In term of earthquake prediction, it is categorized based on how the earthquake occurs. There are three category of prediction. The first is long term prediction, where this prediction rarely implemented as it gets the range of more than 10 years of historical data and some additional information from sequential earthquake as a result of fault location. The second is the intermediate prediction that obtained information from the earthquake location, time and destruction power within several years. The last one is the short-term prediction that makes an earthquake estimation using several days of data set.

#### 2.2.2. Machine learning

machine learning builds an insight from one or more dataset via some specific algorithms. In this work, we compare the performance of three machine learning algorithms, namely Naïve Bayes, support vector machine (SVM) and multinomial regression.

a. SVM

In general, SVM is used to solve classification and regression problem. However, SVM has gained its popularity as it has good performance on empirical data. SVM conceptually simple, it has fast learning algorithm and very often produce accurate results. This is because SVM is a machine learning that is developed based risk minimization principle. In SVM, a training data set D is given as, D=$\{(x_i, y_i) \mid x_i \in R^p, y_i \in \{-1, 1\}\}_{i=1}^n$, $y_i$ is -1 or 1 indicating the class input which is a threshold wavelet coefficients $x_i$ to describe low or high magnitude. For each $x_i$ is the $p$ dimensional vector. A Hyperplane is used to separate between class input which is good when its position between classes. So that, if $wx_1 + b = +1$ is a supporting hyperplane of class +1, then $wx_2 + b = -1$ is the hyperplane to support class -1. In order to count the gap margin between two classes, we can find the distance between two supporting hyperplanes. This margin can be identified via $(wx_1 + b = +1) - (wx_2 + b = -1) = w(x_1 - x_2)$, so that, $\frac{w(x_1-x_2)}{||w||} = \frac{2}{||w||}$. For Linear classification, it will be $min_{(w,b)} \frac{1}{2} w^2$, and for non-linear $\hat{a} = arg\ min_a \frac{1}{2} \Sigma_{i,j=1}^m\ a_i a_j y_i y_j K(x_i, x_j) - \Sigma_{i=1}^m a_i$ where $K(x_i, x_j)$ is a kernel function.

b. Multinomial logistic regression

This method analyzes the relation between bounded and unbounded variable that have more than two variables which generalize logistic regression into multiclass regression. Multinomial logistic regression model with three categories will have formula as follow,

$$P(Y = i \mid x) = \pi_y(x) = \frac{exp(g_i(x))}{1 + \sum_{h=1}^{2} exp(g_h(x))} \tag{1}$$

c. Naïve bayes

Naïve Bayes is a simple classification for counting the probability of combinations of a certain data set. This method assumes there is no dependency between classes to a value in class variable. Bayes theorem, as shown below, derives the posterior probability of two antecedents, which are prior probability and a likelihood function.

$$P(X \mid H) = \frac{P(X \mid H).P(H)}{P(H)} \tag{2}$$

Where, $X$ is the data with unknown class, $H$ is the hypothesis data for class specification, aa is the probability of hypothesis $H$ based on the posterior probability $(X)$, $P(H)$ is the prior probability, $P(X \mid H)$ is the probability observing $X$ given $H$, and $P(X)$ is the marginal evidence of probability of $X$.

d. Evaluation method

In order to evaluate the machine learning performance, we make use of confusion matrix, mean absolute error (MAE), mean Absolute percentage error (MAPE), mean square error (MSE) and root mean square error (RMSE). Confusion matrix describes the performance of classification model from different classes. The classifier has done its work when it gained the information of true positive (TP) and true negative (TN). And, when it classifies the negative value it will produce the false positive (FP) and false negative (FN). In measuring machine learning performance, we evaluates for their accuracy (percent of correctness over all test instances) and precision.In this paper, we measure the performance using mean absoule error (MAE), mean absolute percentage error (MAPE), mean square error (MSE) and root mean square (RMSE),

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (\hat{y}_i - y_i)^2} \tag{3}$$

As shown in the evaluation formula above, $\hat{y}_i$ is the predicted earthquakes, $y_i$ is the data of earthquake from the resources and $T$ is the number of examples used for testing. MAE measures whether our computation towards under and over estimations [28]. MSE is the most common way to evaluate the prediction results, where the error is the differences between the estimation result and its data. MAPE is the evaluation to indicate error when predicting between the original data and its result. MAPE useful when the size of variable is important to evaluate the prediction. Meanwhile, RMSE measurement emphasizes large errors more. RMSE

evaluates how close the observed data points are to the models' predicted values and MAE describes uniformly distributed errors. It is worth to note that the RMSE value is similar to the unit of the outcome. For example, when it measure the depth of an earthquake then the unit is km.

### 2.3. Data collection

This stage begins all of our work by collecting data from different location and various formats. The challenge in this activity is that some data can be retrieved directly from repository as ready to use data. In this work, the data collection activity is categorized into 3 methods, as follow:

(a) Retrieve directly from the repository as it is provided in a ready to use format, such as comma separated value (CSV).
(b) Retrieve a web site, manually, in a hypertext markup language (HTML) format. Then web-scraping to get the information we need from within the HTML text file.

Several techniques applied to different data source. We retrieve the EMSC data by accessing or download of each web page within 14 years (2004 – 2018). The webscraping technique is applied to resources from NOAA, EMSC, ISC, INGV, GFZ and BMKG. For InaTEWS, we downloaded manually. Other data set also downloaded directly, such as GHEA where the data format is not in CSV.

USGS data is in CSV format that we can downloaded almost all the data that range from 1st January 1900 until 31st August 2018. For IRIS data set we obtained data range 1968 to 2018. INGV data set ranges from 1985 to 2018, and for BMKG data set range 2008 to 2018.

### 2.4. Data pre-processing

This stage prepares the data before we make any prediction. Most of the work in this stage is filtering the information such as to identify whether the date, time, latitude, longitude, magnitude and depth exist within the data set. We also remove the data that has magnitude values 0 to avoid any misclassification during processing stage. Data merges also done in this stage. For example, we make classification of data within the same range of dates into 10 years and 30 years. In doing so, we obtained the intersection of data from different resources.

### 2.5. Prediction stage

This stage predicts the data set for specific group of 10 and 30 years. We split the work into two parts. In the first part, we train the data using set of group based on time, date, latitude, longitude, magnitude and depth to find the location and the possibility energy of earthquake. In the next part, we split the dataset into train and test that already categorized into 4 groups which are latitude, longitude, magnitude and depth, where the split ratio is 0.8 over 1.0. We make use R [31] as a tool to make prediction and its library implement some machine learning methods that we implement to. For Naïve Bayes we use the function Naive Bayes and SVM for support vector machine from library e0171 [32]. multinomial logistic regression uses multinom function from library NNET [33].

To predict the earthquake, the object is splitted to have specific result. For example, we predict the location of earthquake as the first step. Then, the magnitude and depth of earthquake is predicted based on the new location that already estimated in the previous step. The result of prediction is the combination of, both, the first step and the second step. In predicting the location of earthquake, we have implemented two techniques. First, we make use of Geohash library to merge the latitude and longitude. Second, we also predict the location of earthquake using only latitude and longitude. We split our prediction based on location as shown in Table 1. It is worth noting that the latitude and longitude is in degrees using decimal fraction.

Table 1. Prediction Factor Based on Location

| Method | | | Machine Learning | | | |
|---|---|---|---|---|---|---|
| Location | | GeoHash | | | Latitude Longitude | |
| Data | Depth | Depth+Magnitude | Magnitude | Depth | Depth+Magnitude | Magnitude |

In predicting the magnitude values of an earthquake, we factorize the prediction into two factors. First, in order to get into magnitude prediction the latitude and longitude are used to get the power of earthquake. Second, we predict via the combination of location and depth, as depicted in Table 2. For the depth of

earthquake, we factorized into the opposite of the magnitude prediction, as shown in Table 3. To visualize our results, we make use of R tool with Shiny [34] library that overlay on top of map that retrieved from google map using ggmap [35] library. The final application of this work is a web-based system.

Table 2. Prediction Factor Based on Depth

| | Machine Learning | | | | | |
|---|---|---|---|---|---|---|
| | Prediction Location Based on Depth | | Prediction Location Based on Depth and Magnitude | | Prediction Location Based on Magnitude | |
| Data | Longitude +Latitude | Longitude + Latitude + Depth | Longitude +Latitude | Longitude + Latitude + Depth | Longitude +Latitude | Longitude + Latitude + Depth |

Table 3. Prediction Factor Based on Magnitude

| | Machine Learning | | | | | |
|---|---|---|---|---|---|---|
| | Prediction Location Based on Depth | | Prediction Location Based on Depth and Magnitude | | Prediction Location Based on Magnitude | |
| Data | Longitude +Latitude | Longitude + Latitude + Magnitude | Longitude +Latitude | Longitude + Latitude + Magnitude | Longitude +Latitude | Longitude + Latitude + Magnitude |

## 3. RESULTS AND ANALYSIS

### 3.1. Analysis

In this work, we make prediction, solely, based on the earthquake data set. Data processes in two condition, first, we grouped into 10 Years and 30 Year, second, without grouping or individual data. Other than that, Naïve Bayes cannot create prediction for 10 and 30 Year individual data set because of imbalance data set. We split the training and testing data into 60% and 40%. We take into account the smaller error will guide us into more accurate prediction. To reduce the complexity of our work, we manage the prediction using a catalog that describe the method and data set, as shown in Table 4.

As shown in Table 5, the actual data that is grouped into 10 years using different evaluation techniques. SVM shows good result for Magnitude prediction and multinomial logistic regression has better results for data with Depth. Naïve Bayes is not included into 10 years analysis. On the other hand, SVM outperforms other method for 30 years dataset with grouping on Magnitude and Depth, as shown in table 5. It shows that the prediction accuracy as shown by MAE has 0.598473 which explicate that the prediction results of earthquake is quite precision than other method.

In making prediction using 10 years of data without grouping, SVM outperforms other algorithm which predict the earthquake location based on Magnitude and Depth. In this prediction, SVM solely predict the factor of latitude and longitude. The result, as depicted in table 6, shows that the prediction has achieved good result when the information of Magnitude and Depth estimates the coordinate location.

In predicting earthquake for 30 years dataset without grouping, multinomial logistic regression (MLR) exceeds other algorithm. It shows that using Magnitude and Depth data, as shown in Table 6, MLR has smaller error than SVM, where in this prediction Naïve Bayes is not included because of imbalance data.

In the next step, we would like to find out which method of machine learning suitable to predict earthquake. To this, we calculate the average of data set to give us an insight of which data set can provide small error rate. As shown in figure 7, the most applicable data set is for 30 year grouping data and 10 years not grouping data, as both shows low level of error rate. And we analyze that those data set has a chance to have good prediction. In more detail, both, the 30 years grouping and 10 years not grouping data set, SVM outperfoms other data with small error rate on using Magnitude information, which also shows smaller error compares to the Depth information. So that, we analyze that SVM will predict earthquake much better when using solely, on Magnitude information.

From the information in Table 7, we analyze that the earthquake prediction should be more accurate when we use Magnitude data as reference. In contrast, when the Depth data are used as reference, we might encounter the accuracy and, probably, has problem to predict the earthquake location prediction. These data give us vision that the depth data might have its use to predict the destruction that might appear to the location prediction.

In measuring the performance of which machine learning method that suitable for earthquake prediction in Indonesia, we compare the average error rate for not grouping and grouping data set. Our result shows that the 30 Years grouping and 10 years not grouping data set give us a reasonable values. As shown in Table 8, SVM outperforms multinomial logistic regression and Naive Bayes. And also, 10 years not grouping data set, SVM shows better performance than Multinomial Logistic Regresion, as depicted in Table 9. Where in 10 Years not grouping data set, because of imbalance data, we cannot obtain result from Naïve Bayes method. Overall, our evaluation on machine learning performance shows that the grouping and not grouping data set which uses Magnitude as grouping reference performs better than using Depth values. Moreover, SVM method show better performance than other algorithm. Due to that we believe the prediction of earthquake that make use of SVM would provide better accuracy than multinomial logistic regression and Naive Bayes using similar data set.

Table 4. An excerpt of 10 years group for prediction method and dataset

| No | Method | Location | Data |
|---|---|---|---|
| 1 | MultiLogReg | Depth | Predict(NonDepth) |
| 2 | MultiLogReg | Depth | Predict(NonDepthNonMag) |
| 3 | MultiLogReg | Depth | Predict(NonMag) |
| 4 | MultiLogReg | Depth | PredictGeoHash(NonDepth) |
| 5 | MultiLogReg | Depth | PredictGeoHash(NonDepthNonMag) |
| 6 | MultiLogReg | Depth | PredictGeoHash(NonMag) |
| 7 | MultiLogReg | Depth+MAG | Predict(NonDepth) |
| 8 | MultiLogReg | Depth+MAG | Predict(NonDepthNonMag) |
| 9 | MultiLogReg | Depth+MAG | Predict(NonMag) |
| 10 | MultiLogReg | Depth+MAG | PredictGeoHash(NonDepth) |
| 11 | MultiLogReg | Depth+MAG | PredictGeoHash(NonDepthNonMag) |
| 12 | MultiLogReg | Depth+MAG | PredictGeoHash(NonMag) |
| 13 | MultiLogReg | MAG | Predict(NonDepth) |
| 14 | MultiLogReg | MAG | Predict(NonDepthNonMag) |
| 15 | MultiLogReg | MAG | Predict(NonMag) |
| 16 | MultiLogReg | MAG | PredictGeoHash(NonDepth) |
| 17 | MultiLogReg | MAG | PredictGeoHash(NonDepthNonMag) |
| 18 | MultiLogReg | MAG | PredictGeoHash(NonMag) |
| 19 | SVM | Depth | Predict(NonDepth) |
| 20 | SVM | Depth | Predict(NonDepthNonMag) |
| 21 | SVM | Depth | Predict(NonMag) |
| 22 | SVM | Depth | PredictGeoHash(NonDepth) |
| 23 | SVM | Depth | PredictGeoHash(NonDepthNonMag) |
| 24 | SVM | Depth | PredictGeoHash(NonMag) |
| 25 | SVM | Depth+MAG | Predict(NonDepth) |
| 26 | SVM | Depth+MAG | Predict(NonDepthNonMag) |
| 27 | SVM | Depth+MAG | Predict(NonMag) |
| 28 | SVM | Depth+MAG | PredictGeoHash(NonDepth) |
| 29 | SVM | Depth+MAG | PredictGeoHash(NonDepthNonMag) |
| 30 | SVM | Depth+MAG | PredictGeoHash(NonMag) |
| 31 | SVM | MAG | Predict(NonDepth) |
| 32 | SVM | MAG | Predict(NonDepthNonMag) |
| 33 | SVM | MAG | Predict(NonMag) |
| 34 | SVM | MAG | PredictGeoHash(NonDepth) |
| 35 | SVM | MAG | PredictGeoHash(NonDepthNonMag) |
| 36 | SVM | MAG | PredictGeoHash(NonMag) |
| 37 | NaiveBayes | Depth | Predict(NonDepth) |
| 38 | NaiveBayes | Depth | Predict(NonDepthNonMag) |
| 39 | NaiveBayes | Depth | Predict(NonMag) |
| 40 | NaiveBayes | Depth | PredictGeoHash(NonDepth) |
| 41 | NaiveBayes | Depth | PredictGeoHash(NonDepthNonMag) |
| 42 | NaiveBayes | Depth | PredictGeoHash(NonMag) |
| 43 | NaiveBayes | Depth+MAG | Predict(NonDepth) |
| 44 | NaiveBayes | Depth+MAG | Predict(NonDepthNonMag) |
| 45 | NaiveBayes | Depth+MAG | Predict(NonMag) |
| 46 | NaiveBayes | Depth+MAG | PredictGeoHash(NonDepth) |
| 47 | NaiveBayes | Depth+MAG | PredictGeoHash(NonDepthNonMag) |

Table 5. Grouping dataset

| Method | Magnitude | Depth |
|---|---|---|
| | 10 Years Evaluation | |
| RMSE | Method(25, 26)0.839928006 | Method(34)123.7999 |
| MAPE | Method (30) 0.186486 | Method (14, 15) 0.712816 |
| MSE | Method (25, 27) 0.705479 | Method (34) 15326.42 |
| MAE | Method (30) 0.681305 | Method (31) 64.91890744 |
| | 30 Years Evaluation | |
| RMSE | Method (25, 26) 0.751008212 | Method (28) 120.3226 |
| MAPE | Method (34, 35) 0.156257 | Method (32, 33) 0.809354 |
| MSE | Method (25, 26) 0.564013 | Method (28)14477.52 |
| MAE | Method (34, 35) 0.598473 | Method(28) 64.5761601 |

Table 6. Ungrouping dataset

| Method | Magnitude | Depth |
|---|---|---|
| | 10 Years Evaluation | |
| RMSE | Method (19, 20) 0.805136856 | Method (23,24) 101.4409 |
| MAPE | Method (19, 20) 0.135727 | Method (23, 24) 1.835921 |
| MSE | Method (19, 20) 0.648245 | Method (23, 24)10290.26 |
| MAE | Method (19, 20) 0.618199 | Method(23, 24) 76.15196673 |
| | 30 Years Evaluation | |
| RMSE | Method (15) 3.663452813 | Method (2) 107.2547 |
| MAPE | Method (15) 0.539494 | Method (1) 0.701563 |
| MSE | Method (15) 13.42089 | Method (1)11503.57 |
| MAE | Method (15) 2.310839 | Method(1) 70.64115023 |

Table 7. Average evaluation result

| Data Set | RMSE_MAG | MAPE_MAG | MSE_MAG | MAE_MAG |
|---|---|---|---|---|
| | Magnitude | | | |
| Data 10 Years (Grouping) | 0.963318 | 0.21023 | 0.94712 | 0.777716 |
| Data 30 Years (Grouping) | 0.854072 | 0.173682 | 0.746437 | 0.676576 |
| Data 10 Years (No Grouping) | 0.868458 | 0.147251 | 0.757441 | 0.672579 |
| Data 30 Years (No Grouping) | 5.051307 | 0.866291 | 25.78514 | 3.706884 |
| | Depth | | | |
| Data 10 Years (Grouping) | 127.0155 | 1.070409 | 16153.99 | 68.82178 |
| Data 30 Years (Grouping) | 125.8881 | 1.162366 | 15885.88 | 70.96083 |
| Data 10 Years (No Grouping) | 109.1246 | 2.463045 | 11940.31 | 80.3022 |
| Data 30 Years (No Grouping) | 109.8351 | 0.765595 | 12066.61 | 72.89245 |

Table 8. Machine learning performance for 30 years

| Method | RMSE_MAG | MAPE_MAG | MSE_MAG | MAE_MAG |
|---|---|---|---|---|
| | Grouping Data Based on Magnitude | | | |
| Multinomial Logistic Regression | 0.777235 | 0.160233 | 0.604094 | 0.61487 |
| SVM | 0.751008 | 0.156257 | 0.564013 | 0.598473 |
| Naïve Bayes | 0.922814 | 0.183305 | 0.851585 | 0.716253 |
| | Grouping Data Based on Depth | | | |
| Multinomial Logistic Regression | 121.9435 | 0.817061 | 14870.22 | 67.01762 |
| SVM | 120.3226 | 0.809354 | 14477.52 | 64.57616 |
| Naïve Bayes | 123.5369 | 1.308522 | 15261.35 | 70.61942 |

Table 9. Machine learning performance for 10 years

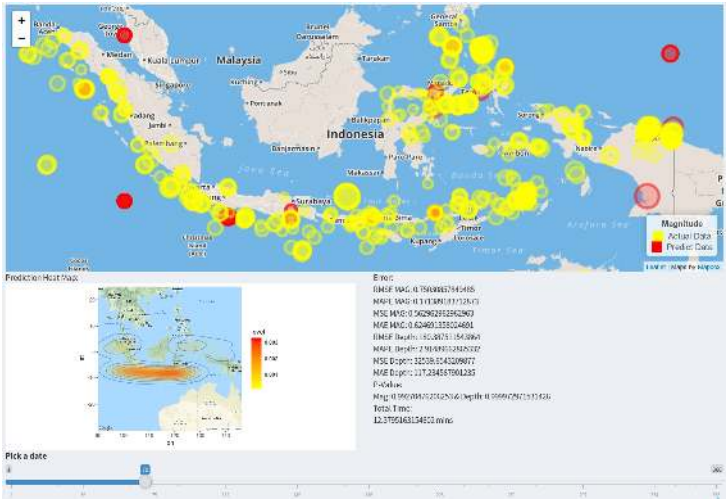
| Method | RMSE_MAG | MAPE_MAG | MSE_MAG | MAE_MAG |
|---|---|---|---|---|
| | Not Grouping Data Based on Magnitude | | | |
| Multinomial Logistic Regression | 0.884768 | 0.150343 | 0.782815 | 0.687099 |
| SVM | 0.805137 | 0.135727 | 0.648245 | 0.618199 |
| | Not Grouping Data Based on Depth | | | |
| Multinomial Logistic Regression | 109.8913 | 2.797098 | 12076.09 | 80.97818 |
| SVM | 101.4409 | 1.835921 | 10290.26 | 76.15197 |

### 3.2. Results

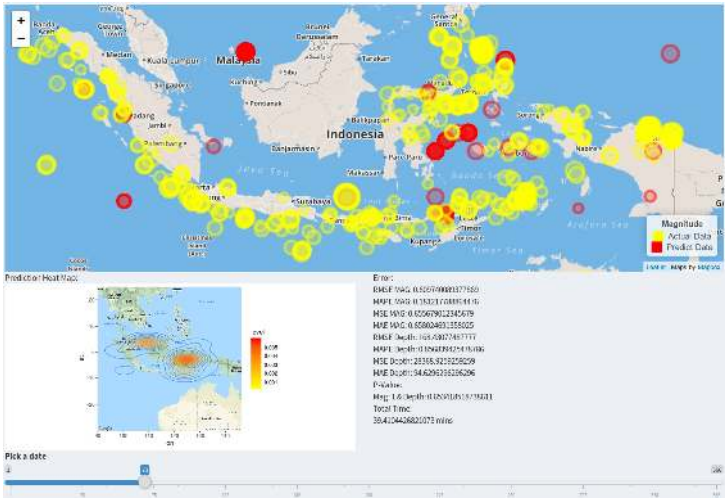
To show the implementation of our prediction into a more visualize information, a web service presentation is shown using R Shiny system. An original information of earthquake is retrieved from Indonesian Geological center. shown in Figure 2(a). We compare the earthquake report from the BMKG Indonesia, as shown in Figure 2(a), and compare it to the prediction results we made before the date of event that is depicted in Figure 2(b), 2(c) and 2(d). Our prediction is based on the number of day within a year. For example if we want to predict earthquake in March 11, 2019, then we count number of days from the beginning of the year up until the D day, where from the calculation we have 70 days. Then, we select the value of day, which is 70 days, into the web-system. In our map, the red colour shows the prediction result and the yellow colour shows the original data.

In comparing the earthquake report from BMKG Indonesia and our prediction result shows that prediction using Naïve Bayes, as shown in 2(b), based on the original learning data is not good enough. multinomial logistic regression performs better than Naïve Bayes, as shown in 2(c), the earthquake location slightly close to the report from BMKG. support vector machine (SVM) achieve better results for eastern Indonesia region, which is out performs other methods. It is worth to note that the training data influence the prediction results. Overall, the prediction results have updated our knowledge that different machine learning may perform differently, although similar data sets were used for training. In our analysis, SVM may have a chance for better earthquake prediction.

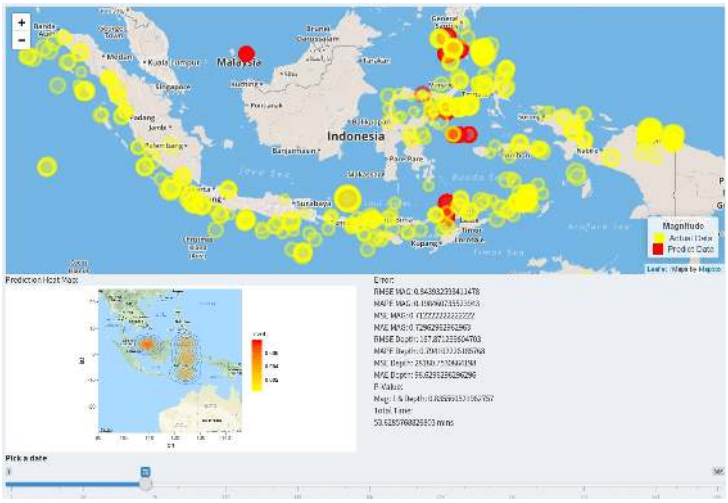
(a)




(b)




(c)




(d)

Figure 1. Earthquake occurs on March 11, 2019, (a) original information from BMKG Indonesia [16], (b) prediction using Naïve Bayes, (c) prediction using multinomial logistic regression, (d) prediction Using SVM.

## 4. CONCLUSION

We have compared machine learning method to predict earthquake location, depth and magnitude for Indonesia region. In order to visualize the prediction results, a web-based application has also been demonstrated. The conclusion we obtained from this work as follow, Naïve Bayes method is not good enough to predict for a grouping data set for only one year, and it is applicable for multi year grouping data. Considering the average error rate, SVM method outperforms other algorithm where using Magnitude data as reference provides better results than using the Depth data. This information leads us into an insight that the Depth can be used as the addition factor for better prediction. We deal with day, month and year as date property for prediction, and our observation shows that prediction based on day performs better. For overall data set, as we already expected, SVM outperforms other method that is followed by multinomial logistic regression in predicting. Naïve Bayes performed worst from all prediction results.

**ACKNOWLEDGEMENT**

This work has been funded by the Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Pelita Harapan, Project No. P-094-FIK-/III/2019. Pujianto Yugopuspito was supported by the Indonesian Directorate General of Strengthening Research and Development of the Ministry of Research, Technology and Higher Education, Contract No. 26/AKM/PNT/2019, focusing on research's infrastructure, and decree No. 7/E/KPT/2019.

**REFERENCES**

[1] P. R. Cummins, "Geohazards in indonesia: Earth science for disaster risk reduction – introduction," *Geological Society, London, Special Publications*, vol. 441, no. 1, pp. 1–7, 2017. [Online]. Available: https://doi.org/10.1144/sp441.11

[2] S. Stein and E. A. Okal, "Speed and size of the sumatra earthquake," *Nature*, vol. 434, no. 7033, pp. 581–582, Mar. 2005. [Online]. Available: https://doi.org/10.1038/434581a

[3] M. Osada and K. Abe, "Mechanism and tectonic implications of the great banda sea earthquake of november 4, 1963," *Physics of the Earth and Planetary Interiors*, vol. 25, no. 2, pp. 129–139, Apr. 1981. [Online]. Available: https://doi.org/10.1016/0031-9201(81)90146-1

[4] A. Ruano, G. Madureira, O. Barros, H. Khosravani, M. Ruano, and P. Ferreira, "Seismic detection using support vector machines," *Neurocomputing*, vol. 135, pp. 273–283, Jul. 2014. [Online]. Available: https://doi.org/10.1016/j.neucom.2013.12.020

[5] U.S. Geological Survey, "Advanced national seismic system (anss) comprehensive catalog," 2017. [Online]. Available: http://earthquake.usgs.gov/earthquakes/search/

[6] Incorporated Research Institutions for Seismology (IRIS), "Data service iris," 2019. [Online]. Available: http://service.iris.edu/

[7] National Geophysical Data Center, "Global significant earthquake database," 1972. [Online]. Available: https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153

[8] Re3data.Org, "European-mediterranean seismological centre," 2016. [Online]. Available: http://service.re3data.org/repository/r3d100011729

[9] I. Bondár and D. Storchak, "Improved location procedures at the international seismological centre," *Geophysical Journal International*, vol. 186, no. 3, pp. 1220–1244, Jul. 2011. [Online]. Available: https://doi.org/10.1111/j.1365-246x.2011.05107.x

[10] Istituto Nazionale Di Geofisica E Vulcanologia (INGV), Istituto Di Geologia Ambientale E Geoingegneria-Consiglio Nazionale Delle Ricerche (IGAG-CNR), Istituto Per La Dinamica Dei Processi Ambientali-Consiglio Nazionale Delle Ricerche (IDPA-CNR), Istituto Di Metodologie Per L'Analisi Ambientale-Consiglio Nazionale Delle Ricerche (IMAA-CNR), and Agenzia Nazionale Per Le Nuove Tecnologie, L'energia E Lo Sviluppo Economico Sostenibile (ENEA CRE Casaccia), "Rete del centro di microzonazione sismica (centromz), sequenza sismica del 2016 in italia centrale," 2018. [Online]. Available: http://cnt.rm.ingv.it/instruments/network/3A

[11] W. Hanka and R. Kind, "The geofon program," *Annals of Geophysics*, vol. 37, no. 5, Sep. 1994. [Online]. Available: http://doi.org/10.4401/ag-4196

[12] R. Steed and A. Fuenzalida, "Dataset for article "crowdsourcing triggers rapid, reliable earthquake locations" by steed et al. (2019)," 2019. [Online]. Available: http://dataservices.gfz-potsdam.de/panmetaworks/showshort.php?id=escidoc:3686893

[13] Indonesia Tsunami Early Warning System (InaTEWS), "Data online bmkg," 2018. [Online]. Available: https://inatews.bmkg.go.id/

[14] P. Albini, R. M. W. Musson, A. Rovida, M. Locati, A. A. G. Capera, and D. Viganò, "The global earthquake history," *Earthquake Spectra*, vol. 30, no. 2, pp. 607–624, May 2014. [Online]. Available: https://doi.org/10.1193/122013eqs297

[15] P. Albini, R. M. Musson, A. A. Gomez Capera, M. Locati, A. Rovida, M. Stucchi, and D. Viganò, "Gem global historical earthquake archive," 2013. [Online]. Available: http://www.globalquakemodel.org/what/seismic-hazard/GHEA

[16] Badan Meteorologi, Klimatology dan Geofisika(BMKG), "Data online bmkg," 2019. [Online]. Available: http://dataonline.bmkg.go.id

[17] S. Shapira, L. Novack, Y. Bar-Dayan, and L. Aharonson-Daniel, "An integrated and interdisciplinary

model for predicting the risk of injury and death in future earthquakes," *PLOS ONE*, vol. 11, no. 3, p. e0151111, Mar. 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0151111

[18] A. J. Yazdi, T. Haukaas, T. Yang, and P. Gardoni, "Multivariate fragility models for earthquake engineering," *Earthquake Spectra*, vol. 32, no. 1, pp. 441–461, Feb. 2016. [Online]. Available: https://doi.org/10.1193/061314eqs085m

[19] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[20] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*. Springer Berlin Heidelberg, 1998, pp. 4–15. [Online]. Available: https://doi.org/10.1007/bfb0026666

[21] G. Zazzaro, F.M. Pisano, and G. Romano, "Bayesian networks for earthquake magnitude classification in a early warning system," 2012.

[22] C. Jiang, X. Wei, X. Cui, and D. You, "Application of support vector machine to synthetic earthquake prediction," *Earthquake Science*, vol. 22, no. 3, pp. 315–320, Jun. 2009. [Online]. Available: https://doi.org/10.1007/s11589-009-0315-8

[23] R. Niu, X. Wu, D. Yao, L. Peng, L. Ai, and J. Peng, "Susceptibility assessment of landslides triggered by the lushan earthquake, april 20, 2013, china," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 9, pp. 3979–3992, Sep. 2014. [Online]. Available: https://doi.org/10.1109/jstars.2014.2308553

[24] G. T. Kaya, O. K. Ersoy, and M. E. Kamasak, "Support vector selection and adaptation for classification of earthquake images," in *2009 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul. 2009. [Online]. Available: https://doi.org/10.1109/igarss.2009.5418229

[25] W. Astuti, R. Akmeliawati, W. Sediono, and M. J. E. Salami, "Hybrid technique using singular value decomposition (SVD) and support vector machine (SVM) approach for earthquake prediction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 5, pp. 1719–1728, May 2014. [Online]. Available: https://doi.org/10.1109/jstars.2014.2321972

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[27] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, "Machine learning predicts laboratory earthquakes," *Geophysical Research Letters*, vol. 44, no. 18, pp. 9276–9282, Sep. 2017. [Online]. Available: https://doi.org/10.1002/2017gl074677

[28] S. Uyeda, T. Nagao, and M. Kamogawa, "Short-term earthquake prediction: Current status of seismo-electromagnetics," *Tectonophysics*, vol. 470, no. 3-4, pp. 205–213, May 2009. [Online]. Available: https://doi.org/10.1016/j.tecto.2008.07.019

[29] M. Syifa, P. Kadavi, and C.-W. Lee, "An artificial intelligence application for post-earthquake damage mapping in palu, central sulawesi, indonesia," *Sensors*, vol. 19, no. 3, p. 542, Jan. 2019. [Online]. Available: https://doi.org/10.3390/s19030542

[30] P. Hajikhodaverdikhan, M. Nazari, M. Mohsenizadeh, S. Shamshirband, and K. wing Chau, "Earthquake prediction with meteorological data by particle filter-based support vector regression," *Engineering Applications of Computational Fluid Mechanics*, vol. 12, no. 1, pp. 679–688, Jan. 2018. [Online]. Available: https://doi.org/10.1080/19942060.2018.1512010

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[32] D. Meyer and T. U. Wien, "Support vector machines. the interface to libsvm in package e1071. online-documentation of the package e1071 for quot;r," 2001.

[33] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. Springer New York, 2002. [Online]. Available: https://doi.org/10.1007/978-0-387-21706-2

[34] K.-W. Moon, "Make a plot with a click," in *Use R!* Springer International Publishing, 2016, pp. 1–14. [Online]. Available: https://doi.org/10.1007/978-3-319-53019-2-1

[35] D. Kahle and H. Wickham, "ggmap: Spatial visualization with ggplot2," *The R Journal*, vol. 5, no. 1, pp. 144–161, 2013. [Online]. Available: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf