International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 2, March-April-2021, ISSN (Online): 2395-566X

Prediction of Earthquake Magnitude Based on the Clusters in Sulawesi Island, Indonesia

Fachrizal Fajrin Aksan[a*], Andi Azizah[a], Enggar Dwi Prihastomo[b]
[a]Politechnika Wrocławska, [b]Politechnika Warszawska
[*]fachrizal.aksan@gmail.com

***Abstract*- In this paper, we present an earthquake magnitude prediction model based on similar earthquake locations. To classify the earthquakes that occurred in Sulawesi Island, we use the K-Means clustering method to group the earthquakes based on the longitude and latitude of the earthquake. Support Vector Regression and Random Forest Regressor are proposed model to predict the magnitude in each cluster based on the longitude, latitude, and depth of the earthquakes. The data of past earthquakes are obtained from the USA Geological Survey and Meteorology, Climatology and Geophysical Agency of the Republic of Indonesia (BMKG). The optimal number of clusters is determined by the elbow method is 3. The prediction results show that the most accurate prediction model is Random Forest Regressor when the clustering approach is used.**

***Keywords*- K-Means, Support Vector Regression, Random Forest Regressor, Earthquake, Elbow method.**

I.INTRODUCTION

Earthquakes are natural disasters that are not commonplace, but can have a great negative impact on lives and property. An earthquake is caused by a sudden sliding on a fault. Tectonic plates always move slowly, but they get stuck at their edges due to friction. When the stress on the edge overcomes the friction, an earthquake occurs. This releases energy in waves that move through the earth's crust and cause the quake. Depending on the cause, earthquakes are classified as tectonic earthquakes, volcanic earthquakes, earthquake debris, and artificial earthquake. Tectonic earthquakes are the most common and damaging earthquakes among the other types of causes. The earthquake based on the depth below the earth's surface hypocentrum in kilometers can be categorized into 3 types: shallow which has depth below 70 km. Intermediate is in a range between 70 km and 300 km. And deep is a level which has a 300 km [1].

Sulawesi is one of islands which are categorized as a region that has specificity in earthquake prone. Sulawesi Island consists of six provinces. Where the northern part of Sulawesi is the most event of earthquake occurred. The earthquake data sources used in this study come from official website of Meteorlogy, Climatology, and Geophysical Agency of the Republic of Indonesia (BMKG). Earthquakes can cause severe damage and hazards, especially in areas near the epicenter. To minimize the damage, it is necessary to study the pattern of the earthquake event spatially, earthquakes contain huge amount of data. The discovery of interesting information known as data mining technique is the right way to make a better conclusion. In this study, the earthquake dataset was grouped based on the longitude and latitude of the earthquake using clustering analysis. The aim to separate the earthquake datasets in different clusters into different sub-datasets is to prepare the new dataset for magnitude prediction.

Cluster analysis is the task of grouping a set of objects so that the objects in the same group are more similar to each other than those in other groups [2]. K-Means clustering is a proposed algorithm to perform partitioning of earthquake focal zones. In the study of Novianti et al [3], the clustering of earthquake epicenters was analyzed using K-Means algorithm. This study was designed to find the optimal number of clusters of earthquake epicenters. The result showed that the optimal earthquake clusters in Bengkulu Province is 7.

Sulawesi is an earthquake prone area; the characteristics of each earthquake must be studied and analyzed. In this study, the K-Means method was used to separate earthquake clusters based on latitude and longitude in Sulawesi Island. Assuming that similar earthquakes are clustered correctly, the experiments should be able to develop a more accurate prediction model to predict the magnitude of the earthquake based on its other characteristics.

In the study of Hashemi et al [4], a different prediction model was required for each cluster to predict the magnitude using decision tree, SVM and KNN. The study was evaluated with 10-fold cross validation and showed that decision tree is more accurate than other algorithms.

However, in this paper the prediction model is proposed are Random forest regressor and Support Vector Regression. The root mean squared error (RMSE) is regression metric used to measure accuracy of model. The contribution of this paper are summarized in the following: K-Means algorithm with Euclidean distance

method is proposed to group the earthquake dataset based on its location and visualize the cluster using geopandas, The Support Vector Regression and Random forest regressor are models to predict the magnitude of earthquake in each cluster.

II. RELATED WORK

In this section, the related work for research paper is divided into two subsections, where the first subsection gives an overview of clustering analysis and the second subsection explains predictive analysis.

1. Clustering Analysis:
Cluster analysis is also known as unsupervised learning in related fields such as machine learning and data mining. It provides insight into data by classifying objects into groups (clusters) of objects, such that objects in one cluster are more similar to each other than objects in other clusters.

There are two purposes of cluster analysis to divide data into groups (clusters), namely understanding and utility. Clustering for understanding means finding the averaging of groups of objects that have a common characteristic and capture the natural structure of the data [5]. It plays an important role in assisting in the analysis of the data.

Clustering for usefulness works to explore the representative objects of individual data objects in the same clusters to discover useful starting points for other purposes, such as data summarization.

In this paper, K- Means is proposed as one of the cluster analysis algorithms to group the dataset based on the similarity of their distance.

1.1 K-Means Algorithm: K-Means clustering, due to its speed and scalability, is most commonly used to portion a given dataset into a set of k groups (k clusters), where k is the number of groups specified in the analysis process. It classifies objects into multiple groups, where objects in the same groups are as identical as possible, while objects from different clusters are as different as possible. Each cluster in K-means clustering is represented by its center (centroid), which is the average of the points assigned to the cluster [6].

The K-means algorithm is a clustering method with an iterative process in which the centers of the clusters, also called centroids, are shifted to the mean position of the points forming them and the instances are reassigned to the associated clusters. The parameters of K-means are the coordinates of the centroids of the clusters and the observation to which each cluster is assigned. The optimal parameter value of K-means is determined by minimizing an objective function [2].

The objective function for K- means is given by the following equation.

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \| x_i^{(j)} - C_j \|^2 \quad [1]$$

Where;
- J, is objective function
- k, is the number of clusters 2
- $\| x\,(j) - C \|$, is distance function $i\ j$
- n, is number of cases
- Cj, is centroid for cluster J
- $x\,(j)$, is case i

The choice of distance measure is a critical step in clustering because it affects the shape of the clusters. Normally, similarity and dissimilarity are measured between objects. A distance function returns a smaller value for pairs of objects that are more similar to each other. The distance measurement method used in this work is Euclidean distance, which is described below.

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad [2]$$

Where;
- n , is number of cases
- xi , is the number of features belongs to object x
- yi, is the number of features belongs to object y

In K-Means clustering, there are several steps that should be followed: Define the number k. The definition of k in the K-Means clustering algorithm is the number of centroids that determine the set of groups in the clustering. Choosing the wrong k can lead to a strange clustering result.

In some datasets, the fixed number of k may be known by prior knowledge about the application domain. To find a good k in K-means, the most popular method is the elbow method to determine the optimal number of clusters by computing the with-in-cluster sum-of-squares (WCSS) for different values of k while running the K- means clustering algorithm. The second step is Choosing Initial Centroids. The initialization of the centroid in clustering is sometimes chosen randomly by distributed data points in a graph.

However, if one sets the initial value in the wrong place, one approaches the wrong result. Another approach tries to make several runs with addressed initial centroids chosen randomly and select the best result.

2. Predictive Analysis:
Predictive analytics enables predictions based on past events. In most cases, historical data is used to create a

statistical model that captures key patterns. This predictive model is then applied to current data to anticipate what will happen next or recommend steps to take to achieve the best results [7]. This paper uses algorithm that are Random Forest Regressor and Support Vector Regression.

2.1 Support Vector Regression: Support Vector Machine (SVM) is quite multipurpose not only does it support nonlinear and linear classification, but also supports linear and nonlinear regression issues. The trick is to reverse the goal: Instead of trying to find the widest possible road between two classes while limiting edge violations, SVM regression tries to get as many instances as possible on the road while limiting edge violations. Support Vector Regression that is used to predict discrete values which has same principle as the SVMs. The main idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

There are a few important parameters of Support Vector Machine, which should be considered: Kernel, hyperplane and decision boundary. In Support Vector Regression, a kernel is used to find a hyperplane in the higher dimensional space without increasing the computational cost. The reason is that the computational cost increases when the dimension of the data increases. The increase in dimension is required when we cannot find a separating hyperplane in a particular dimension and need to move to a higher dimension. In the support vector machine, the hyperplane is generally the separating line between two classes of data. However, in this case, the line is used to predict the continuous output. The decision boundary is a hyperplane that divides the underlying vector space into two groups, one for each class.

2.2 Random Forest Regressor: Random Forest Regressor is a supervised learning algorithm that is an ensemble learning method for regression. It works by constructing a large number of decision trees at training time and outputting the mean prediction of each tree. The sample of random forest regressor can be seen at Fig 1 [8].

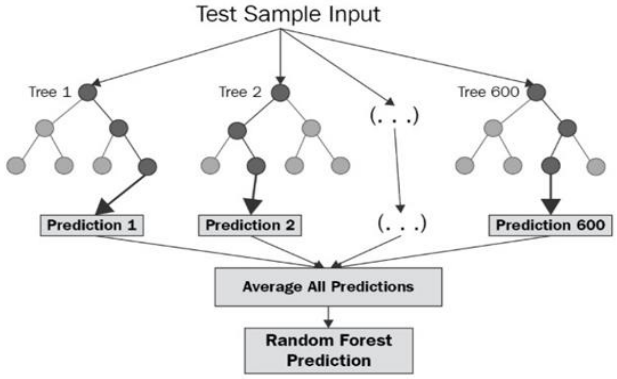

Fig 1. Random Forest Regressor Scheme.

The advantages of Random Forest are that it is one of the most accurate learning algorithms. It can handle multiple input variables without having to delete variables. However, it has the disadvantage of overfitting for some data sets with noisy.

III. THE STUDY AREA AND DATA

According to Meteorology, Climatology and Geophysics Agency (BMKG), the Sulawesi Island is classified as an earthquake-prone area. This earthquake hazard is not only in the northern part of Sulawesi, but in all areas of the island of Sulawesi [9].

Sulawesi is the eleventh largest island in the world and covers an area of 174,600 km2. The center of the island is mountainous with a rough surface, so the peninsulas in Sulawesi are basically far apart and easier to reach by sea than by road. There are three bays that separate the peninsulas in Sulawesi from north to south, namely the Gulf of Tomini, Tolo and Bone.

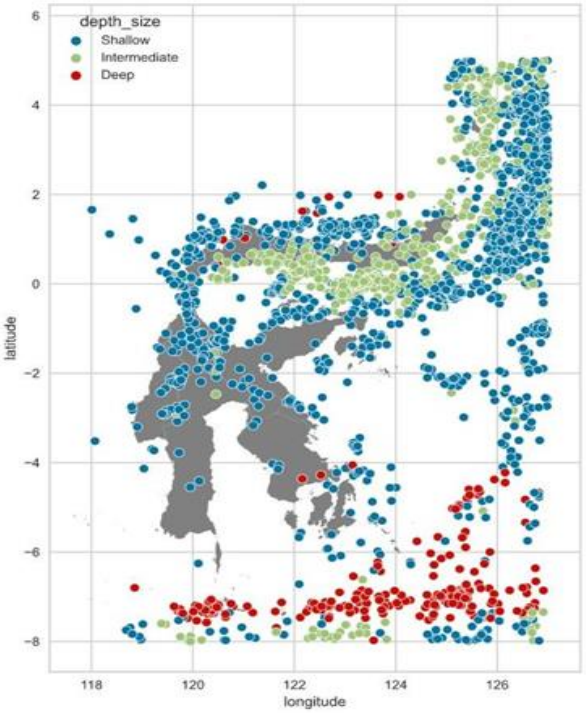

Fig 2. Earthquakes in Sulawesi Island Based on Depth Type.

The Sulawesi Island was formed by a deep shoreline embayment that surrounded it to form an inland area of high mountains, and is largely non-volcanic. According to plate reconstruction, the island is believed to have been formed by the process of terrane collision between the Asian Plate (forming the western and southwestern peninsula) and the Australian Plate (forming the southeastern peninsula and Banggai) with archipelagic arcs previously found in the Pacific Ocean (and forming the northern and southern peninsula. Timur). Due to the

instability of its tectonic history, various faults formed and the island is prone to earthquakes. The map of Sulawesi Island along with the earthquake events can be seen in Fig 2.

This paper uses earthquake record data from the U.S. Geological Survey and the Meteorology, Climatology and Geophysics Agency of the Republic of Indonesia (BMKG) [10]. The dataset consists of date of events, latitude, longitude, depth and magnitude that occurred in Indonesia from 1907 to 2020. In this paper, the dataset is filtered and only based on the events in Sulawesi Island. The statistical analysis of earthquake dataset in Sulawesi island is presented in Table 1.

Table 1. Statistical Description of Earthquake Dataset on Sulawesi Island.

	Latitude	Longitude	Depth	Magnitude
Count	3194	3194	3194	3194
Mean	0.166	124.687	94.685	5.379
Standard Deviation	3.092	2.211	135.267	0.447
Minimum	-7.999	118.009	0.9	5
Maximum	4.997	126.997	663.6	8.2

IV. METHOD

The proposed method is the K-Means algorithm for clustering the earthquake based on its latitude and longitude. After determining the optimal number of clusters by elbow method, the dataset is grouped to its cluster centroid.

The next step is to predict the earthquake magnitude in each cluster using the dependent variables of longitude, latitude and earthquake depth. The scheme of the method can be seen in the following Fig 3. The scheme consists of four major parts: Data acquisition, Data Preprocessing, Earthquake clusters, Magnitude prediction.

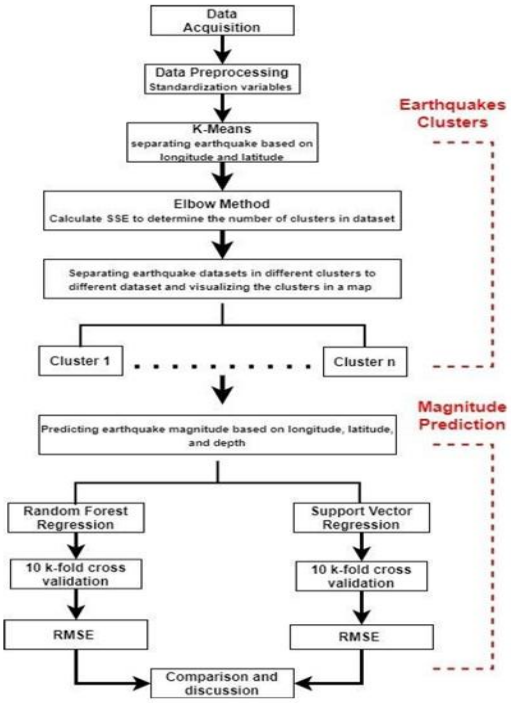

Fig 3. Scheme of the process to cluster the earthquakes and predict the magnitude.

1. Data Acquisition

Data are obtained from reliable single or multiple sources. This step is crucial for clustering and prediction, as the quality and quantity of the data set affects the accuracy of the output. The data collection used in this work is stored in a file in CSV format.

2. Data Preprocessing:

The past historical data is supposed to be pre-processed and reformatted. In this step several techniques are used to transform the raw data into an understandable format. Due to the features of dataset working with different unit, the standardization is necessary for scaling the features where the values are centered on the mean a unit standard deviation [11]. To obtain the standardized value of the feature, the calculation is determined by finding the mean and standard deviation of each feature. The formula to compute the standardization can be done by formula below.

$$X = \frac{X - \mu}{\sigma} \quad (3)$$

Where;

- μ is the mean value of total X
- σ is the standard deviation of the feature values

3. Earthquake Clusters:

This section is part of the clustering analysis, where the K-Means algorithm is used to cluster the earthquake dataset by separating it according to the location of the earthquake. The first task is to perform the K-Means algorithm on a given dataset for a range of values of the

number of clusters (k), in this work the range is between 1 and 9, and calculate the sum of squared errors (SSE) for each value of k. In order to determine the optimal number of clusters, it is necessary to draw a line plot of the SSE for each value of k. The optimal number of k is the one that looks like an arm in the plot line. Since the prediction models for earthquakes in each cluster are developed independently and separately, the reason to separate the earthquake datasets in different clusters into different sub-datasets is to prepare the new dataset for magnitude prediction.

4. Magnitude Prediction:

In predictive analytics, Support Vector Regression and Random Forest Regressor models are proposed for prediction models, which are built to predict the earthquake magnitude based on the predictors or dependent variables consisting of longitude, latitude, and depth of the earthquake in the sub-datasets from each cluster. The prediction model for the sub-dataset is evaluated with 10 k- fold cross validation. The root mean square error (RMSE) is the regression metrics used to calculate and evaluate the accuracy of the prediction model. Root mean square error is the square root of the mean of the square of all of the difference between the estimated and the actual value of the variable. To calculate the RMSE can be used by formula is defined as following [12]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Actual\ value - Predicted\ value)^2}{N}} \quad [4]$$

V. RESULT AND DISCUSSION

This section is divided into two subsections. The first subsection presents the clustering analysis with the result of the performance of the K-Means algorithm to cluster the earthquakes based on the latitude and longitude, displaying the earthquake cluster as shown at Fig 5 and reporting the statistical earthquake data from each cluster, while the second subsection shows the accuracy of the prediction model in each subdataset.

1. K-Means cluster analysis on earthquake data in Sulawesi Island:

The fundamental step in cluster analysis is to determine the optimum k number of cluster [3]. In this experiment, the elbow method is proposed method to determine the optimum k.

The Kelbow Visualizer library from yellowbrick is used to visualize the optimum k number of cluster which shows in Fig 4 .The plot of elbow method show that the optimum k number of cluster is 3 with distortion score is 1650.110.

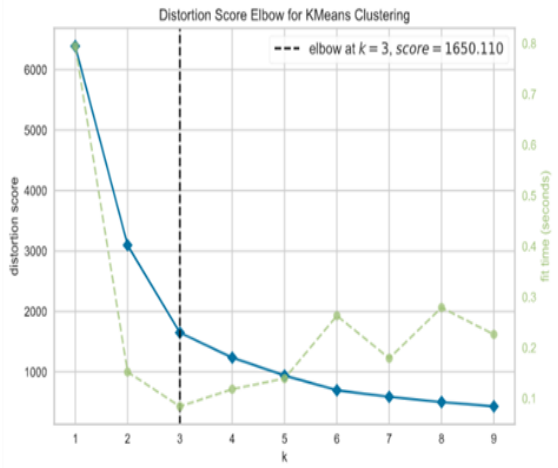

Fig 4. Elbow Method to Find Optimal Number of Cluster.

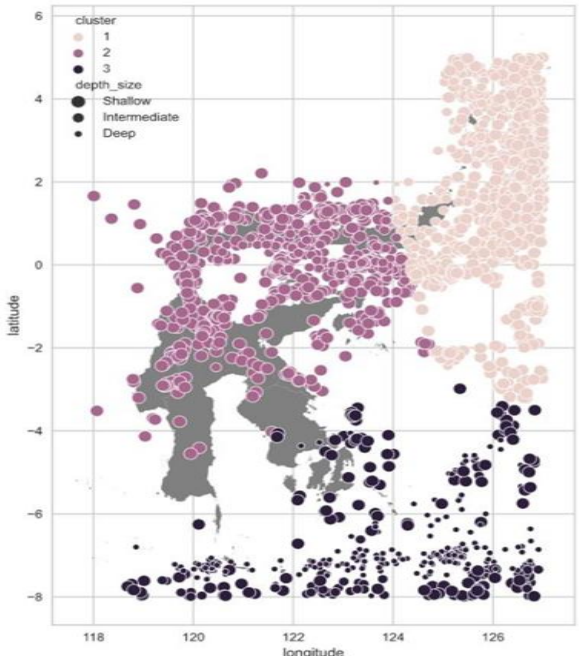

Fig 5. Earthquakes Clusters in Sulawesi Island.

The Table 2 shows that Cluster 1 has more frequent earthquakes than other clusters around 1860 and the average earthquake magnitude is in the range of 5.363.

The earthquake locations in cluster 1 occurred mostly in the northern part of Sulawesi Island, where the centroid of the cluster is located at latitude: 1.778 and longitude: 126.193, in this cluster the average depth of the earthquake is about 51.7 km. while in cluster 2 about 928 earthquakes are counted grouped in the center of Sulawesi Island to Gorontalo province as shown in Fig 5. Cluster 2 shows that the average magnitude of the earthquake is about 5.385 and the average depth is about 75.015 km.

The earthquakes in cluster 3 occurred mainly in Banda Sea and southeast Sulawesi with an average magnitude and depth of 5.442 and 336.453 km, respectively.

Table 2. Statistical Description of the Earthquakes data in Each Cluster of Sulawesi Island.

		Latitude	Longitude	Depth	Magnitude
Cluster 1	Count	1860	1860	1860	1860
	Mean	1.778	126.193	51.726	5.363
	Standard Deviation	1.615	0.648	36.266	0.423
	Minimum	-3.215	124.033	3.3	5
	Maximum	4.997	126.997	313	8.2
Cluster 2	Count	928	928	928	928
	Mean	-0.072	122.056	75.015	5.385
	Standard Deviation	1.150	1.456	70.687	0.467
	Minimum	-4.548	118.009	4.990	5.000
	Maximum	2.2076	124.708	615.6	8.1
Cluster 3	Count	406.000	406.000	406.000	406.000
	Mean	-6.670	123.801	336.453	5.442
	Standard Deviation	1.223	2.168	242.533	0.498
	Minimum	-7.999	118.687	0.900	5.000
	Maximum	-2.984	126.911	663.600	7.900

2. Performance Metrics of Magnitude Prediction:

There are many error metrics that can be used to evaluate and compare between prediction methods when assessing the fit of regression models. However, in this paper, the RMSE is proposed to evaluate the accuracy of model prediction. The RMSE is obtained by 10 k-fold cross validation. In this work, the metrics for regression are supported by scikit-learn python machine learning library [13]. The RMSE can range from 0 to ∞. It is negatively oriented values: smaller values are better [14].

Table 3. RMSE Result of Model Prediction.

Clusters	RMSE	
	Random Forest Regressor	Support Vector Regression
Cluster 1	0.405	0.417
Cluster 2	0.442	0.465
Cluster 3	0.483	0.488
Average (cluster1, cluster2, cluster3)	0.443	0.457

In this experiment, the Random Forest Regressor and Support Vector Regression are created by the scikit-learn Python machine learning library. The Random Forest Regressor [15] is set up with the number of trees in the forest: 100, the random state: 42 and the maximum depth of the three: 2.

On the other hand, Support Vector Regression [16] is built with a linear kernel. The experiment was conducted in each cluster as a subdataset and shows that the RMSE metric Random Forest Regressor has better accuracy in predicting earthquake magnitude than Support Vector Regression which shown in Table 3.

VI. CONCLUSION

The earthquake in Sulawesi Island and its surroundings occurs mainly in the northern part as cluster 1 and in the central part of the island as cluster 2. The optimal cluster obtained by using K-Means to calculate the sum of squared error and visualized by the elbow method is 3.

Using the clustering approach, the comparative model between Support Vector Regression and Random Forest Regressor shows that based on the RMSE evaluation metric Random Forest Regressor is more accurate in predicting the earthquake magnitude, with an average score of 0.443 compared to 0.457 for Support Vector Regression.

REFERENCES

[1] H. J. Wattimanela, U. S. Pasaribu, S. W. Indratno, and A. N. T. Puspito, “Earthquakes clustering based on the magnitude and the depths in Molluca Province,” Bandung, Indonesia, 2015, p. 020021, doi: 10.1063/1.4936449.

[2] B. Everitt, “Cluster Analysis,” p. 348.

[3] P. Novianti, D. Setyorini, and U. Rafflesia, “K-Means cluster analysis in earthquake epicenter clustering,” Int. J. Adv. Intell. Informatics, vol. 3, no. 2, p. 81, Jul. 2017, doi: 10.26555/ijain.v3i2.100.

[4] M. Hashemi and H. A. Karimi, “Seismic Source Modeling by Clustering Earthquakes and Predicting Earthquake Magnitudes,” in Smart City 360°, vol. 166,

[5] A. Leon-Garcia, R. Lenort, D. Holman, D. Staš, V. Krutilova, P. Wicher, D. Cagáňová, D. Špirková, J. Golej, and K. Nguyen, Eds. Cham: Springer International Publishing, 2016, pp. 468–478.

[6] “Springer Theses Recognizing Outstanding Ph.D. Research.”

[7] A. Kassambara, “Copyright ©2017 by Alboukadel Kassambara. All rights reserved.” p. 187.

[8] “2-Aurélien-Géron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_- Concepts-Tools-and-Techniques-to-Build- Intelligent-Systems-O’Reilly-Media-2019.pdf.”

[9] A.Chakure, “Random Forest and Its Implementation,” Medium, Nov. 06, 2020. https://medium.com/ swlh/ random-forest-and-its-implementation-71824ced 454f (accessed Mar. 15, 2021).

[10] “Infografis BMKG Sebut Pulau Sulawesi Rawan Gempa,” iNews. ID, Feb. 02, 2021. https://sulsel.

inews.id/infografis/infografis-bmkg-sebut-pulau-sulawesi-rawan-gempa (accessed Mar. 13, 2021).
[11] BMKG, "Gempabumi Terkini | BMKG," BMKG | Badan Meteorologi, Klimatologi, dan Geofisika. https://www.bmkg.go.id/gempabumi/gempabumi-terkini.bmkg (accessed Mar. 13, 2021).
[12] "Feature Scaling | Standardization Vs Normalization," Analytics Vidhya, Apr. 03, 2020. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization- standardization/ (accessed Mar. 13, 2021).
[13] "RMSE - Root Mean Square Error in Python - AskPython." https://www.askpython.com/python/examples/rmse-root-mean-square-error (accessed Mar. 13, 2021).
[14] "API Reference— scikit-learn 0.24.1 documentation." https://scikitlearn.org/stable/modules/classes.html#regression- metrics (accessed Mar. 14, 2021).
[15] "Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)." http://www.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/ uos3_ko1.htm (accessed Mar. 14, 2021).
[16] "sklearn.ensemble.RandomForestRegressor — scikit-learn 0.24.1 documentation." https://scikit- learn.org/stable/modules/generated/sklearn.ensembl e.RandomForestRegressor.html (accessed Mar. 14, 2021).
[17] "Sklearn.svm. SVR—scikit-learn0.24.1 documentation." https://scikit-learn.org/stable /modules/generated/sklearn.svm.SV R.html (accessed Mar. 14, 2021).