Indonesian Physical Review
Volume 5 Issue 3, September 2022
P-ISSN: 2615-1278, E-ISSN: 2614-7904

Applying K-Means Algorithm for Clustering Analysis Earthquakes Data in West Nusa Tenggara Province

Kertanah[1*], Irwan Rahadi[2], Baiq Aryani Novianti[3], Khaerus Syahidi[3], Sapiruddin[3], Hadian Mandala Putra[4], Muhammad Gazali[1], Ristu Haiban Hirzi[1], Sabar[5]

[1] *Department of Statistics, FMIPA, Universitas Hamzanwadi, Indonesia. E-mail: kertha90@gmail.com*
[2] *Department of Tourism, FBSH, Universitas Hamzanwadi, Indonesia. E-mail: irwanrahadi1@gmail.com*
[3] *Department of Physics Education, FMIPA , Universitas Hamzanwadi, Indonesia. E-mail: khaerussyahidi@hamzanwadi.ac.id, aryani.novidnd@gmail.com, Zafir.Addin@yahoo.com,*
[4]*Department of Computer Engineering, FT, Universitas Hamzanwadi, Indonesia. E-mail: hadian_mandala@hamzanwadi.ac.id*
[5]*Department of Instrumentation and Automation Engineering, Institut Teknologi Sumatera, Indonesia, E-mail: sabar@staff.ac.id*

ARTICLE INFO

Article info:
Received: 08-04-2022
Revised: 01-08-2022
Accepted: 23-08-2022

Keywords:
K-means; Clustering; Earthquakes; West Nusa Tenggara

How To Cite:
Kertanah, I. Rahadi, B. A. Novianti, K. Syahidi, H. M. Putra, M. Gazali, R. H. Hirzi, and Sabar. "Applying K-Means Algorithm for Clustering Analysi Earthquakes Data in West Nusa Tenggara", Indonesian Physical Review, vol. 5, no. 3, p, 197-207, 2022

DOI:
https://doi.org/10.29303/ipr.v5i3.148

ABSTRACT

This study aims to cluster and visualize the earthquake data on a geographical map to determine earthquakes' characteristics using the k-means algorithm. Cluster analysis using the k-means algorithm was carried out on the earthquake data. K-means algorithm is familiar and is one of the well-known techniques to have been applied in cluster analysis. One of Its advantages in cluster analysis is scaling large datasets, for example, earthquake data. The data used in this study is earthquake data in the West Nusa Tenggara from 1991 to 2021. Applying the proposed k-means algorithm, the optimal number of clusters (k) used in this clustering is 2, based on the highest silhouette score of 0.749. The cluster analysis on the geographical map showed that the epicenters of the earthquakes were pretty spread out before 2018, and the number of earthquakes in the eastern region of West Nusa Tenggara is more than in the western area. However, in 2018, the clusters were all bunched in the northern Lombok region. There were a few earthquakes in the west region in 2018, but they happened before August 5. Even after 2019, most earthquakes continue to occur, with the epicenters clustered close to the northern Lombok region

Introduction

Almost every year, earthquakes frequently occur all over the world. Earthquakes are one of the most destructive forces [1] and one of the major natural disasters which cause severe casualties and economic losses to human societies [2]. Indonesia is one of the countries where earthquakes have frequently occurred globally. Some earthquakes are volcanic, tectonic, artificial, and debris [3]. Tectonic and volcanic earthquakes are two kinds of earthquakes that often occur in Indonesia. However, tectonic earthquakes have predominantly occurred since Indonesia is around major tectonics [4]. West Nusa Tenggara Province has become one of the 34 provinces in Indonesia, which is frequently hit by earthquakes. In the last three years, significant earthquakes occurred in Lombok on August 5, 2018 [5]. Clustering earthquakes is a crucial matter of seismicity that gives the primary information on earthquakes. The clustering method is an unsupervised machine learning algorithm to find similarity and relationship patterns among data samples. Additionally, they cluster those samples into groups having similarities based on features. There are many clustering algorithms, but the K-means algorithm is a familiar and well-known clustering approach that is more efficient in a large dataset, for instance, earthquake data.

Many researchers carried out the study of cluster analysis. Weatherill and Burton [6] researched the delineation of shallow seismic source zones in the Aegean region using k-means cluster analysis. Indonesia's earthquake risk clustering has been studied from 1973 to 2017 sing the k-means [7]. Jufriansah et al. [8] analyzed earthquake activity in Indonesia using the k-means clustering method. The k-means method was applied to categorize natural disaster-prone areas in Indonesia [9]. Analyze earthquakes using the k-means algorithm in Ecuador [10]. Kuyuk et al. [11] utilized k-means to classify seismic activity in Istanbul. Applying the k-means algorithm to classify nonlinear seismic responses in a case study of the Hokkaido Iburi-Tobu earthquake in Japan [12].

As the backdrop of research explained above, this paper proposes a model of the k-means algorithm with Elbow and Silhouette methods used to search for the optimal number of clusters and apply it to cluster earthquake data in West Nusa Tenggara province. In addition, it also visualizes the distribution of the earthquake magnitude data on a geographical map for the last thirty years.

Theory

K-means Algorithm

Unsupervised machine learning is a machine learning algorithm used for cluster analysis. K-means algorithm is a type of unsupervised machine learning algorithm, which means that it is utilized when having unlabeled data. The primary purpose of the K-means algorithm is to manage data into clusters such that there are high and low intra-cluster similarities [13]. K-Means clustering works by constantly attempting to find a centroid with closely held data points. It means that each cluster will have a centroid, and the data points in each group will be closer to its centroid than to the other centroids. The K-means objective function is one of the most popular clustering objectives. In K-means, the data is partitioned into disjoint sets C_1; C_2; C_3;……; C_k, in which a centroid represents each Ci μ_i. It is assumed that the input set X is embedded in some larger metric space (X', d) (so that $X \subseteq X'$) and centroids are members

$X^{'}$. The K-means objective function measures the squared distance between each point in X and its clusters' centroid. The centroid of C_i is defined to be [14]

$$`\mu_i(C_i) = \underset{\mu \in X^{'}}{argmin} \sum_{\mu \in C_i} d(x,\mu)^2 \quad (1)$$

Elbow Method

The elbow method is appropriate for relatively small *k* values. It calculates the squared difference of different *k* values. When the *k* value upsurges, the average distortion degree becomes smaller. As the *k* goes up, the position where the improvement effect of the distortion degree levels out the most is the *k* value corresponding to the elbow. An ideal way to find out the optimal number of clusters would be by calculating the within Cluster-Sum of Squares (WCSS), which is the sum of squares of the distances of each data point in all clusters to their respective centroids. The better clustering, the lower the overall WCSS. The following is the elbow equation for WCSS [15].

$$WCSS = \sum_{p_i in\ Cluster\ 1} distance(P_i, C_1)^2 + \sum_{p_i in\ Cluster\ 2} distance(P_i, C_2)^2 + \sum_{p_i in\ Cluster\ 3} distance(P_i, C_3)^2 \quad (2)$$

The equation 2 above can be written in general form as below.

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i in}^{d_m} \right) \quad (3)$$

where:

C: cluster of centroids.

d: the data point in each cluster.

The steps of working of K-means clustering algorithm are as follows.

Stage I: Choosing a suitable value for K, the number of clusters or centroids.

Stage II: Choosing centroids at random for each cluster.

Stage III: Each data point is assigned to its nearest centroid.

Stage IV: Step 4 involves adjusting the centroid for the newly formed cluster.

Stage V: Repeating steps 4 and 5 till all the data points are perfectly organized within a cluster space

Average Silhouette Method

In 1986, Peter J. Rousseuw described the silhouette method, which aims to explain the consistency within-cluster data. The silhouette value will range between -1 and 1 and a a high value indicates that items are well matched within clusters and weakly matched to neighboring clusters [14].

The range of Silhouette's score is between -1 and 1. Its analysis is as follows:

- **+1 score:** near +1 Silhouette score indicates that the sample is far away from its neighboring cluster.
- **0 score:** 0 Silhouette score suggests that the sample is close to the decision boundary separating two neighboring clusters.
- **-1 score:** -1 Silhouette score indicates that the sample has been assigned to the wrong cluster.

The Silhouette score can be calculated by using the following equation:

$$Silhouette\ score = \frac{(p-q)}{\max(p,q)}$$

Where: p: mean distance to the points in the nearest cluster.
q: mean intra-cluster distance to all the points.

Method

Data used

Earthquakes data is a study case located in the West Nusa Tenggara Province. The recorded data captured earthquakes with a reported magnitude of 3.0 or higher through the last thirty years from 1991 to 2021. This case study's data is obtained from the United States Geology Survey (USGS) global public seismic catalogs, which provide real-time earthquake data on past earthquakes [16]. With 22 available variables in earthquake data, this study only focuses on using six variables: time, longitude, latitude, depth, and mag, place.

Proposed model of K-means Algorithm

In this case, the building model of the k-means algorithm uses the Elbow method and Silhouette analysis method to find the optimal number of clusters and the best number of clusters, respectively. The proposed algorithm is utilized in this study, written in pseudocode algorithm.

Optimal number of clusters
(K-Means)

Declaration:
Variables: data frame (df), number of cluster (k), random_state

Algorithm:

Step1: Start
Step 2: *Input df, number of clusters (k), random_state* ##df=earthquakes data
Step 3: *Build model K-means*
Model=KMeans(random_state=n)## n =1,2,3,……..n
Step 4: *Find optimal K value using Elbow Method*
Elbow_Method = KElbowVisualizer(model,k=(2,10))## k values taken is from 2 to 10.
Step 5: *Find the best K value using Silhouette Method*
Silhouette_Method= KElbowVisualizer(model,k=(2,10),metric='Silhouette')
Step 6: Stop

This study uses python programming to assist in data analysis and visualization. The following is a flowchart of the data analysis and visualization process using python programming, as depicted in Figure 2.

Figure 1. Flowchart of data processing using python

The data processing using python was demonstrated. First, the libraries of pandas, numpy, matplotlib, and geopandas are imported. Pandas is the main library for data processing; Numpy is a multidimensional matrix operation library, Matplotlib is a python plotting library, and Geopandas is one of the python libraries to make choropleth maps. Then, the earthquake data is read into the memory. The exploratory data analysis is to check any data that has missing values or is empty; it will be filled in by 0 or means of data and get a statistical summary. The data preparation step is clean data and gathering data used in the subsequent data processing and analysis. Therefore, the model of the K-means algorithm is built by using the Elbow method and the Silhouette method to find the optimal number of clusters and choose the optimal number of clusters to be applied in the clustering process, respectively. Data visualization is the last stage in this process to visualize the clustering data on the geographical map.

Result and Discussion

Referring to the proposed model of the k-means algorithm in the previous section above, the randomly generated dataset for k-means clustering uses *KMeans* class and fit function available in the python *sklearn* package. The Elbow and Silhouette methods are used to find the optimal number of clusters in the earthquake dataset. The random state is set in an integer to obtain the same output when running it multiple times so that the chosen value of the random state is 1 in this case, while the number of clusters (k) is from 2 to 10. The following graph finds the optimal number of clusters (k) using the elbow method.

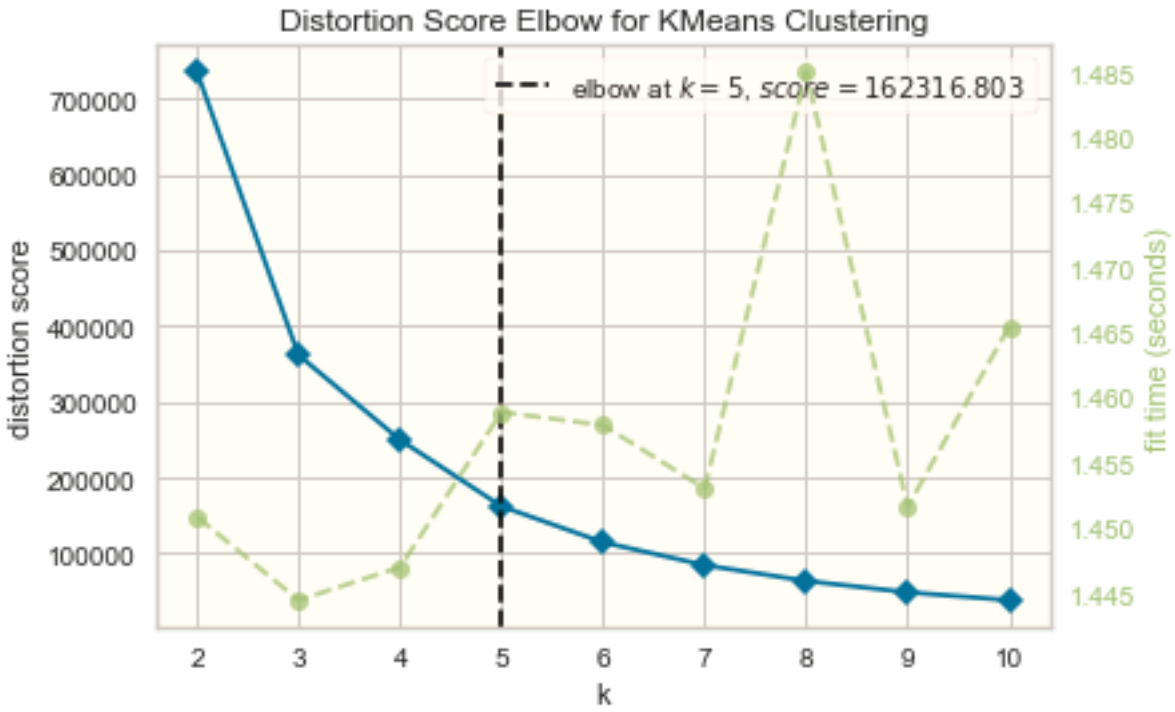
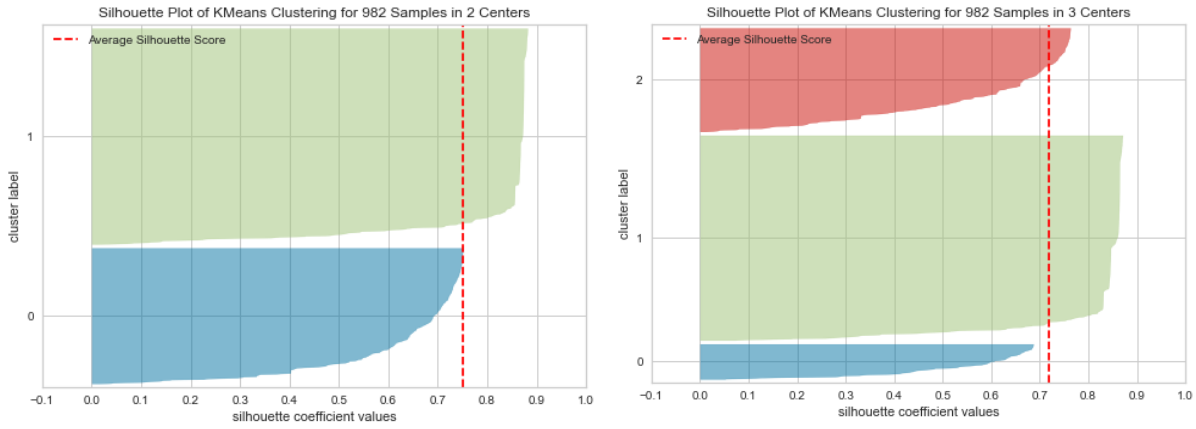
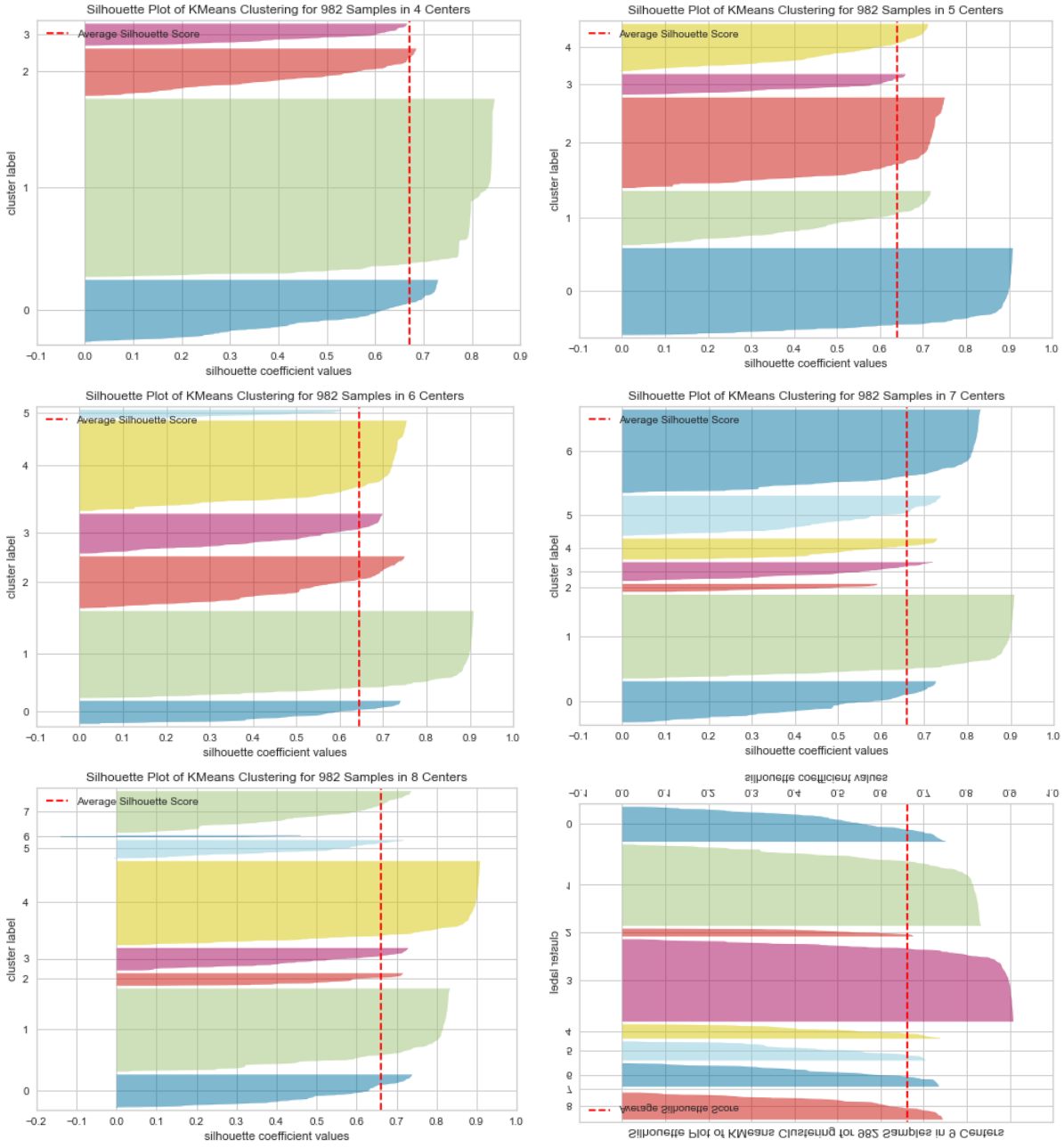


Figure 2. Number of clusters (k) against distortions

As depicted in Figure 2, the optimal number of clusters is plotted against the distortion (total of within-cluster sum of squares for a given number of *k*). The optimal number of clusters is a point at which there is a bend in the curve (elbow or knee). The graph above shows the reduction of a distortion score as the number of clusters increases. However, there is no clear "elbow" visible. The underlying algorithm suggests 5 clusters. A choice of 5 or 6 seems to be fair. Another way to choose the best number of clusters is using the silhouette method. Here is the silhouette score plotted in the number of clusters against silhouette coefficient values.




Silhouette Plot of KMeans Clustering for 982 Samples in 4 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
Silhouette Plot of KMeans Clustering for 982 Samples in 5 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
Silhouette Plot of KMeans Clustering for 982 Samples in 6 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
Silhouette Plot of KMeans Clustering for 982 Samples in 7 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
Silhouette Plot of KMeans Clustering for 982 Samples in 8 Centers
Average Silhouette Score
cluster label
silhouette coefficient values

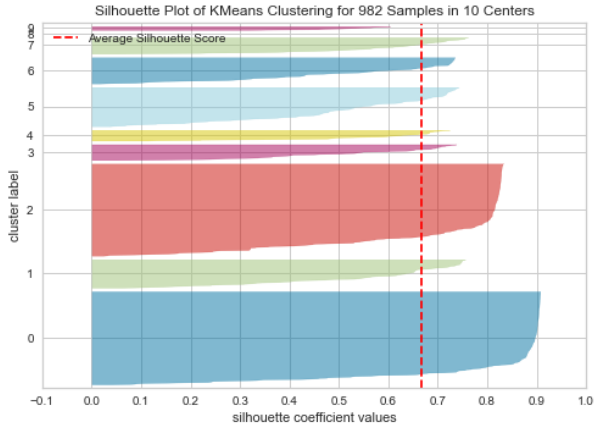


Figure 3. Number of clusters (k) silhouette scores

The silhouette scores of each number of clusters are indicated in Table 1, as below.

Table 1. Number of clusters against silhouette scores

No	Number of clusters (k)	Silhouette scores
1	2	0.749
2	3	0.720
3	4	0.670
4	5	0.639
5	6	0.643
6	7	0.659
7	8	0.660
8	9	0.662
9	10	0.666

Table 1 above shows the silhouette score for the different number of clusters. The underlying algorithm indicates that the highest score is depicted by the number of clusters (k) equal to 2. The highest silhouette score presents the optimal number of clusters, so the number of clusters equals two used in this clustering process.

An earthquake can be detected anywhere on the earth's surface between 0 and 700 km and is classified as shallow, intermediate, or deep based on its depth. These earthquakes can be noticed both on land and on the seafloor. However, earthquakes of similar magnitude can have varying depths if the detected location is below the ocean surface instead of on the ground. The clustered earthquake magnitude with depth was plotted and divided into 2 clusters based on the best number of clusters using the silhouette method above, as depicted in Figure 4 below.

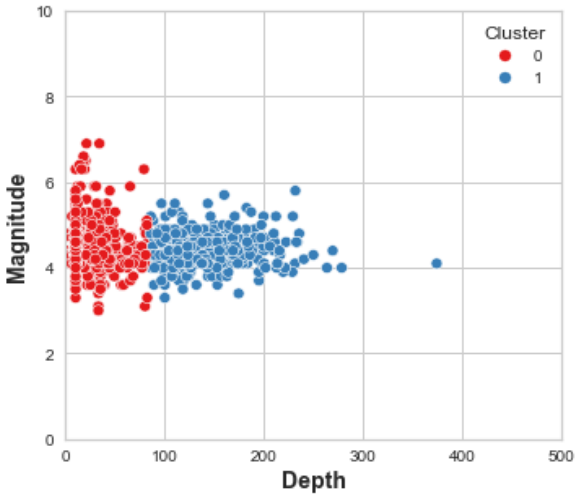


Figure 4. Plot clustered earthquakes magnitude against depth

The number of earthquake magnitude clusters is shown in red and blue in the first cluster (cluster 0) and the second cluster (cluster 1). Most earthquakes' maximum depth occurred between 0 and 374.5 kilometers, while the highest magnitude of 6.9 occurred at 21.6 kilometers on the earth's surface. Out of the 982 earthquakes in the data set, 603 earthquakes are in cluster 1 (red color), and the remaining 379 earthquakes are in cluster 2 (blue color). Plotting the latitude and longitude of various earthquakes in the West Nusa Tenggara province gives the data more context. It helps understand how different-clustered earthquakes are distributed across West Nusa Tenggara province based on their source.

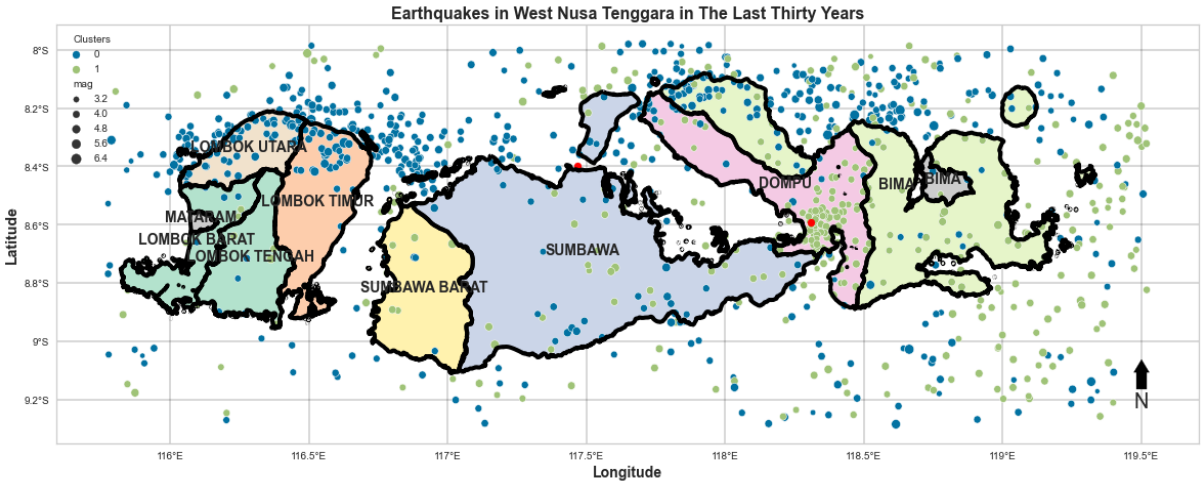


Figure 5. Distributions of clustered magnitude earthquakes data for the last thirty years

The centroid epicenters for the two clusters are also red on the map in Figure 6. The Euclidian distance between the centroids and the data points was used to generate the clusters.

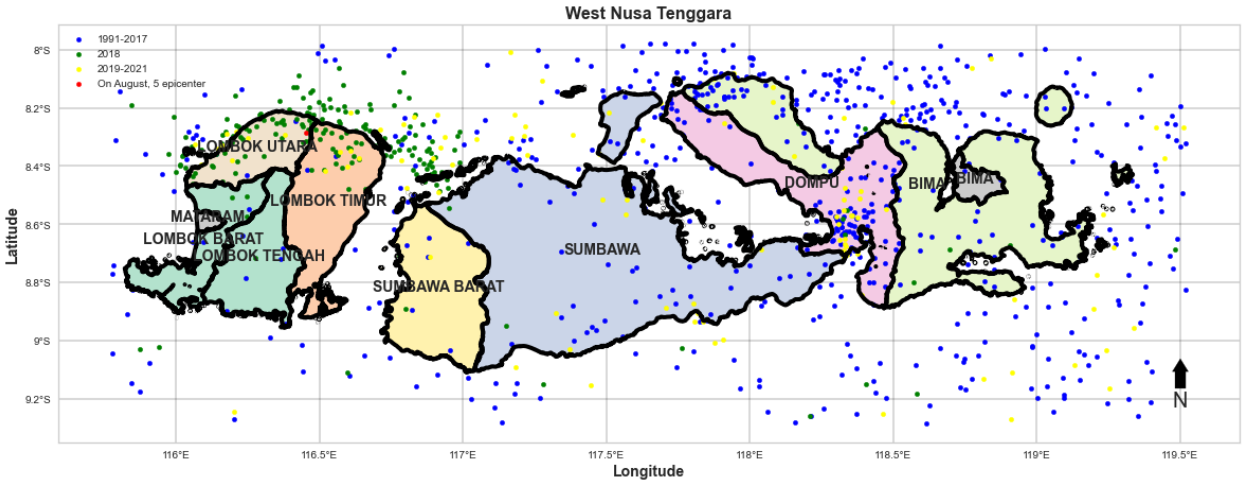


Figure 6. Spread of clustered magnitude earthquakes data from 1991 to 2021

The map depicted in Figure 7 shows that the epicenters of the earthquakes were pretty spread out before 2018. The number of earthquakes in the eastern region (Sumbawa Island) appears more than in the western area (Lombok Island). However, in 2018, the clusters were all bunched in the northern Lombok region. There were a few earthquakes in the western region in 2018, but they likely happened before August 5 (shown in blue color). Even after 2019, most earthquakes continue to occur, with the epicenters clustered close to the northern Lombok region (shown in red).

Conclusion

The K-means algorithm is a popular machine learning algorithm used in cluster analysis. It was used to analyze earthquake data in West Nusa Tenggara province. The cluster analysis on the geographical map shows that the eastern region of West Nusa Tenggara has more earthquakes than the western region. In 2018, the clusters gathered in the northern part of Lombok Island. There were a few earthquakes in the west part of Lombok Island, but they most likely happened before August 5, with the epicenters clustered near the northern Lombok region.

References

[1] Hsieh, T., & Chen, C. (2010). *Visualizing Field-Measured Seismic Data*. 65–72.

[2] Bao, Z.; Zhao, J.; Huang, P. ., & Yong, S.; Wang, X. (2021). *A Deep Learning-Based Electromagnetic Signal for Earthquake Magnitude Prediction*. 21, 4434.

[3] Novianti, P., Setyorini, D., & Rafflesia, U. (2017). *K-Means cluster analysis in earthquake epicenter clustering*. *3*(2), 81–89. https://doi.org/http://dx.doi.org/10.26555/ijain.v3i2.100

[4] Wattimanela, H. J. (2019). *Grouping of Tectonic Earthquakes in the Province of Nusa Tenggara Barat Indonesia with K-Means Cluster Method Approach and Determination of Distribution Type*. *2*(3), 177–191. https://doi.org/https://doi.org/10.30598/SNVol2Iss3pp177-191year2019

[5] Harini, S., Fahmi, H., D. Mulyanto, A., & Khudzaifah, M. (2020). *The earthquake events and*

impacts mapping in Bali and Nusa Tenggara using a clustering method The earthquake events and impacts mapping in Bali and Nusa Tenggara using a clustering method. https://doi.org/10.1088/1755-1315/456/1/012087

[6] Weatherill, G., Burton, P. W., & Burton, W. (2009). *Delineation of shallow seismic source zones using K -means cluster analysis , with application to the Aegean region*. 565–588. https://doi.org/10.1111/j.1365-246X.2008.03997.x

[7] Rifa, I. H., Pratiwi, H., Sciences, N., & Maret, U. S. (2020). *Clustering of earthquake risk in indonesia using k-medoids and k-means algorithms*. *13*(2), 194–205. https://doi.org/10.14710/medstat.13.1.194-205

[8] Jufriansah, A., Pramudya, Y., Khusnani, A., & Saputra, S. (2021). *Analysis of Earthquake Activity in Indonesia by Clustering Method*. *5*(2), 92–103. https://doi.org/10.20961/jphystheor-appl.v5i2.59133

[9] Supriyadi, Bambang., Windarto, Agus Perdana., Soemartono, Triyuni., U. (2018). Classification of Natural Disaster Prone Areas in Indonesia using K-. *International Journal of Grid and Distributed Computing*, 87–98. https://doi.org/http://dx.doi.org/10.14257/ijgdc.2018.11.8.08

[10] Ricardo, J. E., Juan, J., Menéndez, D., Manuel, J., Bermúdez, M., Lemus, N. M., … Barcos, I. F. (2021). *Neutrosophic Sets and Systems Neutrosophic K-means for the analysis of earthquake data in Ecuador Neutrosophic K-means for the analysis of earthquake data in Ecuador*. 44.

[11] Kuyuk, H. S., Yildirim, E., Dogan, E., & Horasan, G. (2012). *Nonlinear Processes in Geophysics Application of k -means and Gaussian mixture model for classification of seismic activities in Istanbul*. 411–419. https://doi.org/10.5194/npg-19-411-2012

[12] Ji, K., Wen, R., Ren, Y., & Dhakal, Y. P. (2018). Nonlinear seismic site response classification using K-means clustering algorithm: Case study of the September 6, 2018 Mw6.6 Hokkaido Iburi-Tobu earthquake, Japan. *Soil Dynamics and Earthquake Engineering*. https://doi.org/https://doi.org/10.1016/j.soildyn.2019.105907

[13] Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps*. Bangalore, Karnataka, India: Apress Media.

[14] Shalev-Shwartz, Shai.Ben-David, S. (2014). *Understanding Machine Learning : From Theory to Algorithms*. New York, The United States of America: Cambridge University Press.

[15] Cui, M. (2020). *Introduction to the K-Means Clustering Algorithm Based on the Elbow Method Mengyao*. 5–8. https://doi.org/10.23977/accaf.2020.010102

[16]USGS. (2022). *Search Earthquake Catalog*. Available online from: https://earthquake.usgs.gov/earthquakes/search/. [Accessed March 22, 2022].