P-ISSN : 2715-2448 | E-ISSN : 2715-7199
Vol.4 No.2 July 2023
Buana Information Technology and Computer Sciences (BIT and CS)

Agglomerative Clustering of 2022 Earthquakes in North Sulawesi, Indonesia

Berton Maruli Siahaan[1], Afrioni Roma Rio[2]
[1,2]Department of Physics, Faculty of Mathematic and Natural Science, Sam Ratulangi University
Kampus UNSRAT, Bahu, Manado, North Sulawesi, INDONESIA
[1]bertonsiahaan@unsrat.ac.id, [2]afrioni@unsrat.ac.id

Abstract
This paper presents a cluster analysis of earthquake data in the surrounding region of North Sulawesi, Indonesia. The dataset comprises seismic data recorded throughout the year 2022, obtained from the BMKG earthquake repository. A total of 211 earthquakes were included in the analysis, with a minimum magnitude threshold of 2.5 and a maximum depth of 300 km. The agglomerative clustering technique, combined with the elbow method, was employed to determine the optimal and distinct number of clusters. As a result, four unique clusters were identified. Cluster 1 exhibited high magnitudes, with an average magnitude of 4.4, and shallow depths, averaging at 20 km. Cluster 2 also had high magnitudes, averaging at 4.4, but deeper depths, with an average of 199 km. Cluster 3 consisted of earthquakes with low magnitudes, averaging at 3.4, and shallow depths, averaging at 21 km. Lastly, Cluster 4 comprised earthquakes with low magnitudes, averaging at 3.4, but deeper depths, with an average of 136 km. Among the 211 earthquakes, 29 were assigned to Cluster 1, 39 to Cluster 2, 100 to Cluster 3, which had the highest population, and 43 to Cluster 4. This study provides valuable insights into the clustering patterns and characteristics of earthquakes in the region, contributing to a better understanding of seismic activity in North Sulawesi, Indonesia.

Keywords: Earthquake clustering; earthquakes data, machine learning; agglomerative clustering

Abstrak
Artikel ini membahas analisis klasterisasi pada data gempa bumi di sekitar Sulawesi Utara, Indonesia. Data yang digunakan adalah data gempa bumi yang terjadi selama tahun 2022 yang diperoleh dari repositori gempa BMKG. Terdapat 211 gempa bumi yang dianalisis dengan batas magnitudo terendah 2,5 dan kedalaman maksimum 300 km. Dalam penelitian ini, digunakan teknik klasterisasi aglomeratif dan metode elbow untuk menentukan jumlah klaster yang optimal dan unik. Hasilnya, ditemukan empat klaster yang unik. Klaster 1 memiliki gempa bumi dengan magnitudo tinggi rata-rata sebesar 4,4 dan kedalaman dangkal rata-rata sebesar 20 km. Klaster 2 juga memiliki gempa bumi dengan magnitudo tinggi rata-rata 4,4, namun kedalaman lebih dalam rata-rata sebesar 199 km. Klaster 3 terdiri dari gempa bumi dengan magnitudo rendah rata-rata 3,4 dan kedalaman dangkal rata-rata sebesar 21 km. Klaster 4 terdiri dari gempa bumi dengan magnitudo rendah rata-rata 3,4, namun kedalaman lebih dalam rata-rata sebesar 136 km. Dari total 211 gempa bumi yang dianalisis, terdapat 29 gempa bumi dalam Klaster 1, 39 gempa bumi dalam Klaster 2, 100 gempa bumi dalam Klaster 3 yang memiliki populasi terbanyak, dan 43 gempa bumi dalam Klaster 4. Penelitian ini memberikan pemahaman yang lebih baik mengenai pola dan karakteristik gempa bumi di wilayah Sulawesi Utara, Indonesia.

Kata kunci: *Klasterisasi gempa bumi; data gempa; machine learning: agglomerative clustering*

I. Introduction

Earthquakes are natural phenomena that frequently occur in various regions around the world, including Indonesia, which is known for its high seismic activity. This region is located around the Pacific Ocean, known as the Pacific Ring of Fire (ROF). The Pacific Ring of Fire is a long chain of active volcanoes and tectonic structures encircling the Pacific Ocean. It stretches along the western coast of South and North America, passes through the Aleutian Islands in Alaska, extends along the eastern coast of Asia through New Zealand, and reaches the northern coast of Antarctica. The Pacific Ring of Fire is one of the most active geological areas on Earth and often experiences powerful earthquakes and volcanic eruptions. It is home to over 450 active and inactive volcanoes. Most of these volcanoes are formed through the process of subduction, where dense oceanic plates collide and slide beneath lighter continental plates. Material from the ocean floor melts as it enters the Earth's mantle and then rises to the surface as magma. The deepest trench in the ocean, the Mariana Trench, is located along the western part of the Pacific Ring of Fire. The majority of the world's largest earthquakes also occur within this ring. These earthquakes are caused by sudden movements of rocks laterally or vertically along plate boundaries. Approximately 81% of the world's largest earthquakes occur within the Pacific Ring of Fire [1]. In Indonesia, the frequency of earthquakes is exceptionally high, with an average of 6,512 tectonic earthquake events per year, equivalent to 543 events per month and 18 earthquake events per day [2].

This study focuses on the clustering of earthquakes in the vicinity of North Sulawesi. The data used is sourced from the earthquake repository of the Meteorology, Climatology, and Geophysics Agency (BMKG) for a one-year period in 2022 [3]. A total of 211 earthquakes have been analyzed, meeting the minimum magnitude criteria of 2.5 and a maximum depth of 300 km. The aim of this research is to identify clustering patterns that can provide deeper insights into the seismic activity in the area.

Clustering is an effective approach for analyzing and grouping earthquake data based on their characteristics. Several studies have applied this clustering method to categorize disaster data, disaster impacts, and tsunami potentials [4]–[7]. In this study, the agglomerative clustering method is used to partition the data into closely related groups based on attribute similarities. The validation of the unique and optimal number of clusters is performed using the elbow method.

The cluster analysis results reveal the presence of four unique clusters. Each cluster exhibits distinct characteristics, including magnitude scale and depth. Understanding the earthquake clustering patterns in the North Sulawesi region can significantly contribute to disaster mitigation efforts. By identifying the specific clusters and their characteristics, authorities and disaster management agencies can gain valuable insights into the distribution and behavior of earthquakes in the area. This knowledge can aid in the development of more targeted and effective disaster preparedness plans, early warning systems, and evacuation strategies. Additionally, it can inform infrastructure development and building codes to ensure resilience against seismic events. Ultimately, this research serves as a valuable resource for stakeholders involved in disaster management, enabling them to make informed decisions and implement proactive measures to reduce the potential impact of earthquakes and enhance the overall resilience of the region.

II. Methods

This study encompassed various stages, which involved conducting a review of relevant literature, collecting earthquake data, processing the data, applying the elbow method, performing agglomerative clustering, and conducting cluster analysis (refer to Fig. 1).

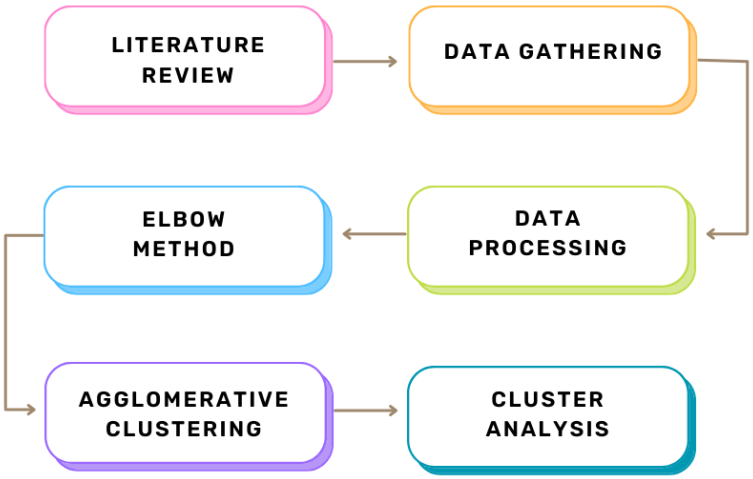

Figure 1. Flowchart of research methodology.

A. Literature Review

In a high-quality research study, it is crucial to have a relevant collection of literature and references as a strong foundation. Therefore, the initial stage of this research involved gathering literature related to the research topic. The collected literature encompassed various important aspects pertaining to the research subject.

During the initial stage, the researcher gathered information related to earthquakes. This aimed to understand the characteristics, causes, and impacts of earthquakes that are relevant to the context of this research. Additionally, the researcher studied machine learning, which is the technique or algorithm used for data processing in this study. A profound understanding of machine learning is crucial as it forms the basis for the data clustering conducted in this research.

Furthermore, the researcher obtained an understanding of the clustering technique using the agglomerative clustering method. Clustering is a method used to group data with similar characteristics into specific clusters. In the context of this research, the application of the agglomerative clustering method enables the identification of unique clustering patterns and structures within the earthquake data in the North Sulawesi region

B. Data Gathering

To obtain earthquake data in North Sulawesi for the year 2022, the data source used is the official website of the Meteorology, Climatology, and Geophysics Agency (BMKG), accessible at https://repogempa.bmkg.go.id/ [3]. The geographic region parameter is set to 0°N to 3°N latitude and 123°E to 126°E longitude. Although the data available on this website is limited to a single month, by leveraging knowledge of the API (application programming interface), we can efficiently access and extract data using for loops and the requests module in the Python programming language [8].

The process of retrieving data through the BMKG API will generate an HTML file that needs further processing. To convert it into a more structured format, the data needs to be parsed into a tabular form using the BeautifulSoup library [9]. Once the data has been successfully parsed and organized, the next step is to save it in CSV format for easier further data processing.

C. Data Processing

After successfully obtaining the earthquake data in North Sulawesi for the year 2022, the next step is to process and explore the data. The data processing is performed using the Python programming language [8], utilizing several packages and libraries as described below:

To perform numerical calculations on arrays or matrices, the NumPy library is used [10]. Additionally, for data analysis and processing in the form of dataframes, the Pandas library is employed

[11]. For visualizing graphs and plotting data distribution on maps, the Matplotlib, Seaborn, and Plotly libraries (including the API provided by Mapbox) are utilized [12]–[15].

Before proceeding with the agglomerative clustering method for clustering, the selected parameters for analysis, namely the earthquake magnitude scale (M) and depth (km), undergo a data preprocessing step using the standard scaler method.

The standard scaler is a widely used technique in data preprocessing that transforms the data by subtracting the mean and dividing by the standard deviation. This process ensures that the data is centered around zero with a standard deviation of 1, making it more suitable for clustering algorithms. By applying the standard scaler, the magnitude and depth values are normalized to a comparable scale, eliminating any potential biases caused by differences in their measurement units. This normalization allows for a fair comparison and accurate clustering based on the similarity of the scaled features.

D. Elbow Method

The elbow method is a visual technique utilized to determine the optimal number of clusters for clustering algorithms. It involves plotting the explained variance by each cluster against the number of clusters and observing the point of inflection, commonly referred to as the "elbow," where adding more clusters no longer significantly improves the explained variance [16].

In simpler terms, the elbow method assists in selecting the appropriate number of clusters for the given data by identifying the point at which adding more clusters does not yield substantial enhancements in the clustering outcomes.

To apply the elbow method, we initially perform clustering with various numbers of clusters and plot the explained variance against the corresponding number of clusters. The elbow point on the plot represents the stage at which the explained variance begins to level off, indicating that the inclusion of additional clusters does not contribute significantly to the improvement.

Once the elbow point is determined, we can choose the number of clusters that strikes a balance between the explained variance and the simplicity of the model.

E. Agglomerative Clustering

Once the optimal number of clusters has been determined using the elbow method, the subsequent step involves applying agglomerative clustering to the preprocessed data. Agglomerative clustering is a hierarchical clustering technique that begins with each data point assigned as a separate cluster and gradually merges the closest clusters until all data points are grouped into a single cluster [17].

In this research, we utilized the agglomerative clustering algorithm provided by the Python Scikit-learn library [18]. The algorithm requires specifying the number of clusters, which is set to the optimal number obtained from the elbow method. For this study, Ward's linkage criterion was employed, aiming to minimize the sum of squared differences within all clusters.

F. Cluster Analysis

Afterwards, the resulting clusters are analyzed to identify patterns or trends within the data. This analysis involves examining the characteristics of each cluster, such as the average values of relevant variables, as well as utilizing visualizations like box plots to depict the clusters.

By conducting cluster analysis, we can gain insights into the data structure and identify meaningful clusters among the population of 211 earthquakes in North Sulawesi. These clusters can be further investigated to obtain a deeper understanding of the characteristics specific to each cluster.

III. Results And Discussion

In this section, we will delve into the findings derived from the analysis of earthquake clusters in the Sulawesi Utara region of Indonesia for the year 2022. The application of the elbow method resulted in the identification of four distinct clusters, as illustrated in Fig. 2. These clusters represent groups of earthquakes that share similar characteristics in terms of their magnitudes and depths.

Following the determination of the optimal number of clusters, the agglomerative clustering technique was employed to further examine the data. This hierarchical clustering method starts by considering each earthquake event as an individual cluster and then iteratively merges the two closest clusters until all data points are grouped into a single cluster. By applying this method to the identified four clusters, we were able to discern notable dissimilarities among them, particularly in terms of the magnitudes and depths of the earthquakes they encompass.

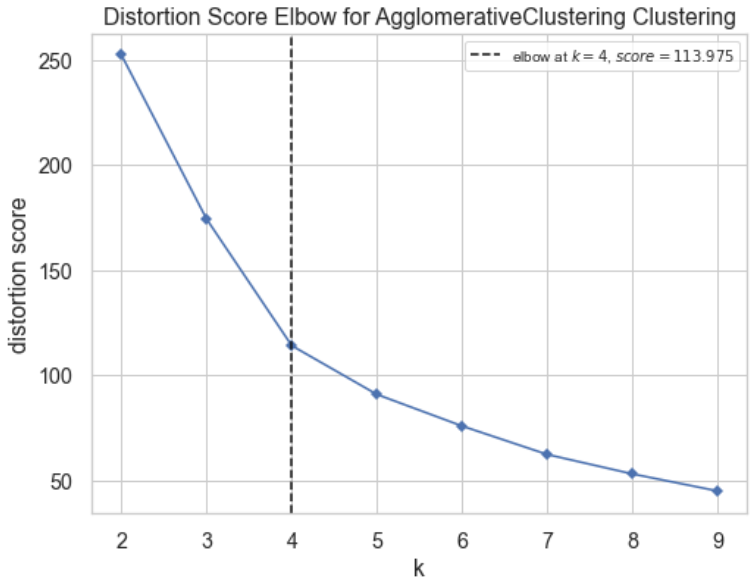

Figure 2. Elbow method with agglomerative clustering

Out of the total of 211 earthquakes analyzed, we observed that the first cluster consisted of 29 earthquakes, the second cluster contained 39 earthquakes, the third cluster emerged as the largest group with 100 earthquakes, and the fourth cluster comprised 43 earthquakes. These cluster-specific earthquake populations can be visualized in Fig. 3.

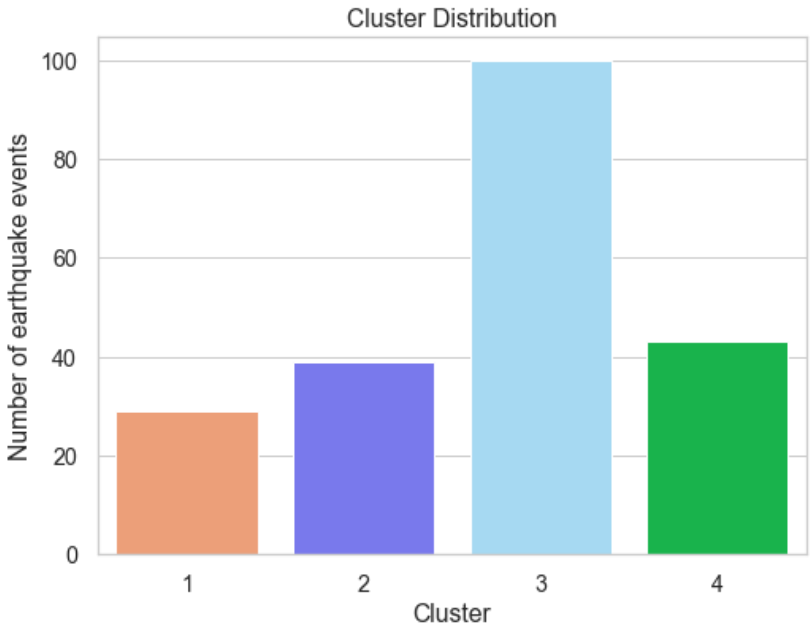

Figure 3. Distribution of earthquake events based on cluster.

The distribution of earthquake data and the formed clusters can be found on the map displayed in Fig. 4. In the cluster analysis, Cluster 3 emerged as the most dominant cluster, primarily distributed in the offshore area. This cluster exhibits characteristics of low magnitude scale, with an average of 3.4 M, and shallow depths, with an average of 21 km. Therefore, this area tends to be relatively safe from the impacts of earthquakes.

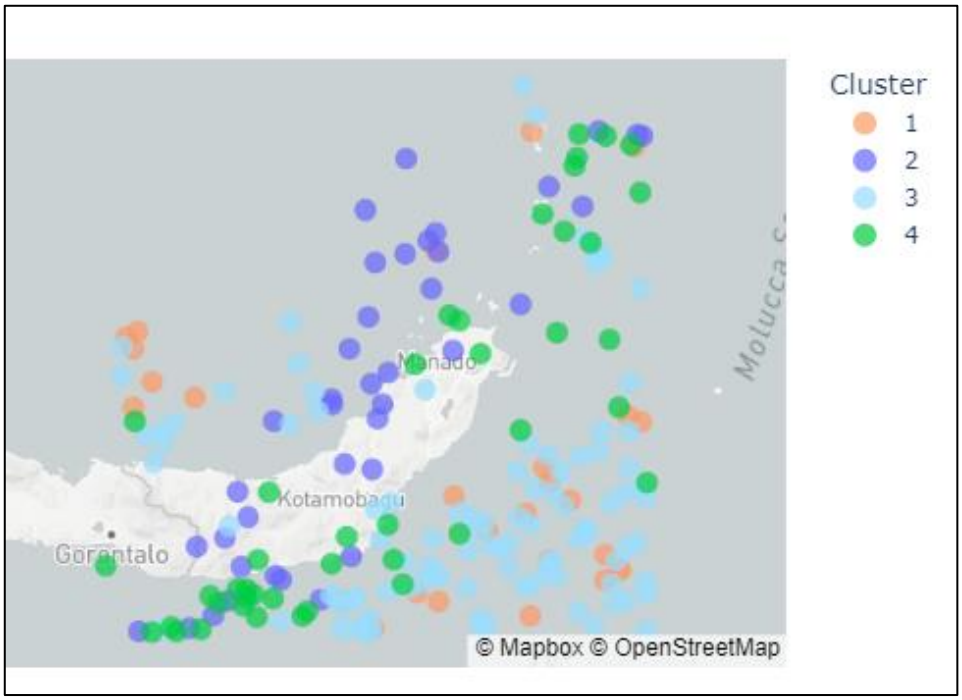

Figure 4. Clustering results: grouping of North Sulawesi earthquake events in 2022

No cluster 1 was found in the main island area. This cluster needs to be closely monitored as it exhibits a relatively high magnitude scale, with an average of 4.4 M, and shallow depths, with an average of 20 km. Shallow depths can increase the potential for damage compared to deeper depths. However, in the northern island areas of mainland Sulawesi, there are several earthquake sources that fall within cluster 1. Therefore, this area requires careful mitigation planning to reduce the risk for the local population.

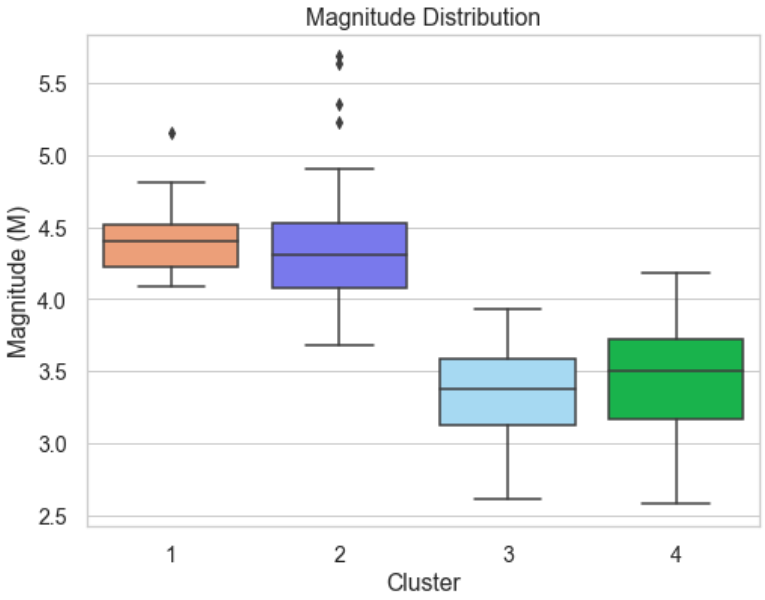

Figure 5. Distribution of magnitude based on cluster.

Cluster 2 also displays a high magnitude scale, with an average of approximately 4.4 magnitude, and deep depths with an average of 199 km. This type of earthquake often occurs in the main island area, which may be related to the activities of active volcanoes in the North Sulawesi region.

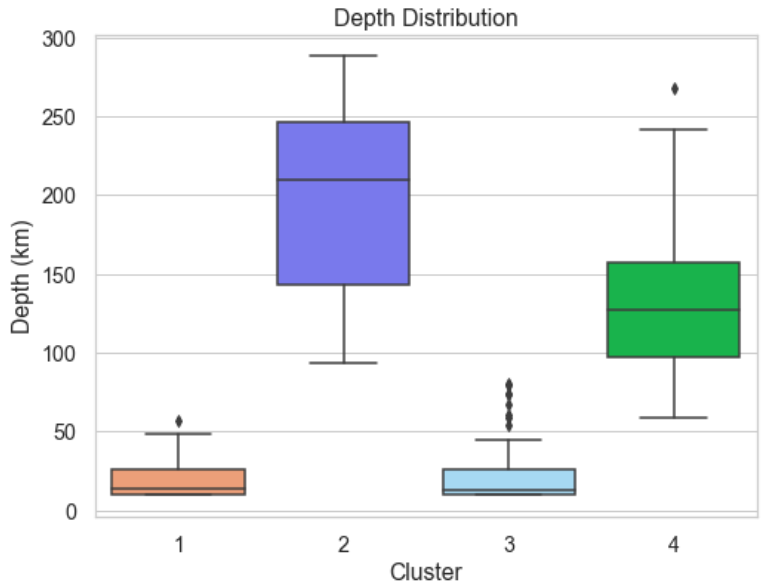

Figure 6. Distribution of depth based on cluster.

Cluster 4 exhibits a low magnitude scale and deep depths, making it relatively safe. However, it is still important to remain vigilant regarding the potential earthquakes within this cluster. The distribution and descriptive data of the clustering results can be observed in Fig. 5 and Fig. 6, along with Table 1.

Table 1. Data description based on cluster

Cluster	Number of earthquakes	Average magnitude (M)	Average depth (km)
1	29	4.4	20
2	39	4.4	199
3	100	3.4	21
4	43	3.4	136

Having an understanding of the characteristics of the formed clusters, this information can be utilized in disaster mitigation efforts and more effective planning to protect the community and reduce the risks associated with earthquake impacts in North Sulawesi. By incorporating this knowledge, appropriate measures can be taken to enhance preparedness, response, and resilience in the face of seismic events. It is crucial to prioritize the safety and well-being of the population and ensure that strategies are in place to mitigate the potential consequences of earthquakes. You can access the results of the clustering analysis through the following link: https://sl.unsrat.ac.id/EQ-AC.

IV. Conclusions

In conclusion, the application of agglomerative clustering to analyze earthquake data in the Sulawesi Utara region resulted in the identification of four distinct clusters with varying characteristics in terms of magnitude and depth. These clusters offer valuable insights into the seismic activity patterns in the area. The findings of this study have significant implications for disaster mitigation and risk reduction efforts in Sulawesi Utara. By understanding the clustering patterns of earthquakes, stakeholders can develop more effective strategies to protect the local population and minimize the impact of seismic events.

Future research in this field could focus on expanding the dataset by incorporating data from additional sources and over a longer time period. This would provide a more comprehensive understanding of the seismic activity and clustering patterns in Sulawesi Utara. Additionally, investigating the correlation between these earthquake clusters and geological features, such as fault lines or volcanic activity, could provide valuable insights for predicting and mitigating future seismic events.

V. References

[1] M. Masum and M. A. Akbar, “The Pacific ring of fire is working as a home country of geothermal resources in the world,” in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, 2019, p. 012020.

[2] A. Sabtaji, “Statistik kejadian gempa bumi tektonik tiap provinsi di wilayah Indonesia selama 11 tahun pengamatan (2009-2019),” *Bul. Meteorol. Klimatol. Dan Geofis.*, vol. 1, no. 7, pp. 31–46, 2020.

[3] B. M. K. dan Geofisika, “EQ Repository,” 2023. https://repogempa.bmkg.go.id/

[4] P. Novianti, D. Setyorini, and U. Rafflesia, “K-Means cluster analysis in earthquake epicenter clustering,” *Int. J. Adv. Intell. Inform.*, vol. 3, no. 2, pp. 81–89, 2017.

[5] M. Murdiaty, A. Angela, and C. Sylvia, “Pengelompokkan Data Bencana Alam Berdasarkan Wilayah, Waktu, Jumlah Korban dan Kerusakan Fasilitas Dengan Algoritma K-Means,” *J. Media Inform. Budidarma*, vol. 4, no. 3, pp. 744–752, 2020.

[6] M. T. Furqon and L. Muflikhah, “Clustering the potential risk of tsunami using Density-Based Spatial clustering of application with noise (DBSCAN),” *J. Environ. Eng. Sustain. Technol.*, vol. 3, no. 1, pp. 1–8, 2016.

[7] A. Wahyu and R. Rushendra, “Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa,” *JEPIN J. Edukasi Dan Penelit. Inform.*, vol. 8, no. 1, pp. 174–179, 2022.
[8] M. F. Sanner and others, “Python: a programming language for software integration and development,” *J Mol Graph Model*, vol. 17, no. 1, pp. 57–61, 1999.
[9] L. Richardson, “Beautiful soup documentation.” April, 2007.
[10] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.
[11] W. McKinney and others, “pandas: a foundational Python library for data analysis and statistics,” *Python High Perform. Sci. Comput.*, vol. 14, no. 9, pp. 1–9, 2011.
[12] P. Barrett, J. Hunter, J. T. Miller, J.-C. Hsu, and P. Greenfield, “matplotlib–A Portable Python Plotting Package,” in *Astronomical data analysis software and systems XIV*, 2005, p. 91.
[13] M. L. Waskom, “Seaborn: statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021.
[14] P. T. Inc, “Collaborative data science,” 2015. https://plot.ly
[15] Mapbox, “Map, Geocoding and Navigations APIs at: https://docs.mapbox.com/api/maps/.” Accessed, 2023.
[16] P. Bholowalia and A. Kumar, “EBK-means: A clustering technique based on elbow method and k-means in WSN,” *Int. J. Comput. Appl.*, vol. 105, no. 9, 2014.
[17] D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” *ArXiv Prepr. ArXiv11092378*, 2011.
[18] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.