Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 6 (2024) 768 - 778 e-ISSN: 2580-0760

Comparative Study of Earthquake Clustering in Indonesia Using K-Medoids, K-Means, DBSCAN, Fuzzy C-Means and K-AP Algorithms

Nurfidah Dwitiyanti[1], Siti Ayu Kumala[2], Shinta Dwi Handayani[3]
[1,2,3]Informatics Engineering, Fakultas Teknik dan Ilmu Komputer, Universitas Indraprasta PGRI, Jakarta, Indonesia
[1]nurfidah.pulungan@gmail.com, [2]sitikumala7891@gmail.com, [3]shintadh.1109@gmail.com

Abstract

Indonesia's frequent earthquakes, caused by its position at the convergence of multiple tectonic plates, Indonesia's frequent earthquakes, caused by its position at the convergence of multiple tectonic plates, necessitate precise seismic zone identification to improve disaster preparedness. This research evaluates the effectiveness of five clustering algorithms—K-Medoids, K-Means, DBSCAN, Fuzzy C-Means, and K-Affinity Propagation (K-AP)—for analyzing earthquake data from January 2017 to January 2023. Using a dataset from BMKG encompassing 13,860 seismic events, each algorithm was assessed based on Silhouette Score and Cluster Purity metrics. Results indicated that K-Means provided the best balance, forming six clusters with a Silhouette Score of 0.3245 and Cluster Purity of 0.7366, making it the most suitable for seismic zone analysis. K-Medoids closely followed with a Silhouette Score of 0.3158 and Cluster Purity of 0.7190. Although DBSCAN effectively handled noise, its negative Silhouette values indicated poor clustering quality. Fuzzy C-Means and K-AP underperformed, with K-AP generating an impractically high number of clusters (196) and the lowest Silhouette Score (0.2550). This study offers a novel, comprehensive comparison of clustering algorithms for Indonesian earthquake data, emphasizing a dual-metric evaluation approach. By identifying K-Means as the most effective algorithm, provides valuable insights for disaster mitigation and seismic risk analysis.

Keywords: *cluster purity; comparative study; earthquake clustering; K-Means*

How to Cite: N. Dwitiyanti, Siti Ayu Kumala, and Shinta Dwi Handayani, "Comparative Study of Earthquake Clustering in Indonesia Using K-Medoids, K-Means, DBSCAN, Fuzzy C-Means and K-AP Algorithms", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 6, pp. 768 - 778, Dec. 2024.
DOI: https://doi.org/10.29207/resti.v8i6.5514

1. Introduction

Indonesia frequently experiences earthquakes, a common natural calamity in the region. The elevated seismicity and volcanic activity of the nation can be ascribed to its location at the intersection of numerous dynamic tectonic plates worldwide, such as the Eurasian, Pacific, and Indo-Australian plates, in addition to the Philippine microplate This geological setting results in the heightened susceptibility of Indonesia to seismic events [1]-[5]. The interaction of these tectonic plates situates Indonesia within a zone characterized by elevated levels of volcanic eruptions and seismic events [6]. Indonesia encounters 6,512 tectonic earthquake occurrences annually, with 543 incidents recorded monthly and 18 events on a daily basis, as indicated by statistical information [7]. Given the considerable seismic activity present, there are limited regions in Indonesia that remain unaffected by the possibility of experiencing earthquakes. Consequently, individuals residing in zones at high risk must maintain a constant state of vigilance, considering the unpredictability of earthquake occurrences.

Adopting novel approaches to disaster mitigation is imperative in diminishing the probability of damage and casualties resulting from these capricious seismic occurrences. Diverse stakeholders, encompassing governmental bodies, academia, the scientific sphere, and the wider populace, have endorsed a variety of strategies. Precise delineation of earthquake-prone regions stands as a pivotal element within these frameworks. The objective is to equip authorities with the requisite data to formulate tailored development blueprints for each high-risk zone through meticulous mapping exercises. This crucial realization further enables the implementation of focused mitigation

Received: 30-10-2024 | Accepted: 22-12-2024 | Published Online: 27-12-2024

actions aimed at mitigating seismic repercussions [8], [9].

Clustering methodologies represent a viable approach for delineating regions susceptible to seismic activity. These methodologies segment a collection of data into clusters of interconnected entities, referred to as clusters. Therefore, this division aids in the grouping together of similar entities or cases, guaranteeing that entities/cases within the identical cluster display a higher degree of similarity to one another when contrasted with entities/cases in separate clusters [10], [11]. Clusters enable the discovery of underlying patterns in datasets by employing the technique of clustering. Entities grouped together in a cluster display a notable degree of resemblance, whereas those found in separate clusters manifest remarkable differences. Essentially, entities within a single cluster are expected to display a notably greater degree of resemblance compared to those distributed across diverse clusters [8]. Clustering techniques can be employed to identify seismic or earthquake patterns based on unlabelled attributes [12].

Several clustering methods have been employed in past research for examining seismic event data. Methods like K-Medoids, K-Means, DBSCAN, Fuzzy C-Means, and K-AP exhibit diverse effectiveness in the clustering of earthquakes, a process assessed through metrics like Sum of Squared Errors (SSE), Davies-Bouldin Index (DBI), Dunn Index (DI), and silhouette score. K-Means, which is a type of partition-based clustering methodology, is renowned for its effectiveness and simplicity, often surpassing alternative techniques in terms of SSE, DBI, and DI. Consequently, it is frequently favored for the clustering of seismic data based on attributes such as longitude, latitude, magnitude, and depth [13], [14]. The K-Medoids algorithm, akin to K-Means but employing medoids in lieu of centroids, has demonstrated superior silhouette scores and computational efficacy in numerous research endeavors, notably in the clustering of earthquake data in Indonesia. This algorithm exhibits resilience towards outliers and excels in handling extensive datasets [15].

For example, in research conducted by [1] using earthquake data in Indonesia sourced from USGS data for the period 2014-2018 comparing the K-Medoids method with K-Means. The research variables are latitude, longitude, earthquake depth, and earthquake strength. According to the study, the K-Medoids approach produces a Silhouette Score = 0.4574067 with cluster 6 as the most optimal cluster compared to the K-Means method with k = 4 having a Silhouette Score = 0.3607622. Similar research conducted also by [8] shows the results of the Silhouette Score = 0.546 from the K-Medoids Algorithm are better than the results of the Silhouette Score = 0.516 from the K-Means algorithm for k = 5 in clustering earthquake data in Indonesia in 1973-2017 based on depth and magnitude. DBSCAN, a density-based algorithm, excels in handling noise and outliers, making it suitable for spatial clustering of seismic events, although it has difficulty in handling datasets with varying densities and large sizes [16]. For instance, a comparative analysis was conducted on the DBSCAN and PCA-DBSCAN algorithms used for clustering seismic regions [17]. The findings showed that the experiment using PCA resulted in the highest SI value, which was 0.4137. The study showed that the DBSCAN algorithm performed better than the K-Medoids algorithm in the classification of earthquake-prone areas [18]. Meanwhile, Fuzzy C-Means offers a robust clustering approach characterized by high precision and efficient computational speed [19]. The computational method identified as Fuzzy C-Means facilitates the allocation of data points to various clusters with diverse degrees of membership, hence providing flexibility in clustering ambiguous data. Nevertheless, the efficiency of this method might not reach its peak when encountering non-convex clusters [13], [20].

The K-AP algorithm, despite not being explicitly addressed in the provided context, is recognized for its capability to detect instances and create clusters without the need to specify the number of clusters, a feature that can be beneficial in scenarios involving dynamic earthquake datasets. Studies have been conducted simultaneously on the seismic event grouping in the region of Indonesia, employing the K-AP clustering method in contrast to the K-Means clustering approach. The results revealed that the assessments of K-Means and K-AP utilizing the C-Index, Davies Bouldin Index, and Connectivity Index determined the most suitable number of clusters for K-Means as 3 and 5 (C-Index=0.052; DBI=0.109 and CI=5.102). Conversely, for K-AP, the optimum cluster quantities were recognized to be 2 and 4 (C-Index=0.022; DBI=0.108 and CI=2.173). Through the utilization of cluster variance, it was ascertained that employing four clusters with K-AP was the optimal approach due to its lower Sw/Sb value in comparison to K-Means [4]. When considering the entirety of the situation, the selection of an algorithm is contingent upon the distinct attributes of the seismic data and the preferred equilibrium between computational efficacy and clustering precision. Every algorithm possesses its own set of advantages and disadvantages, and its effectiveness may exhibit notable discrepancies depending on the assessment criteria employed and the characteristics of the data [1], [13], [16], [20]-[22].

Therefore, by leveraging existing research, the current study seeks to perform a comprehensive comparative analysis of five different clustering methodologies - specifically K-Medoids, K-Means, DBSCAN, Fuzzy C-Means, and K-AP - within the framework of seismic data acquired from Indonesia spanning from 2017 to 2023. The investigation will concentrate on crucial variables such as latitude, longitude, depth, and magnitude to evaluate the efficacy of the aforementioned clustering techniques, this study will employ a range of evaluation metrics, such as the

Silhouette Score, and Cluster Purity, which will serve as a unique feature setting it apart from prior research. The use of cluster purity allows for an evaluation of whether the clustering results of a particular algorithm accurately reflect groupings of seismic events within the same geographic region or with similar characteristics. This indicator is considered more relevant and direct compared to the Silhouette Score, which primarily evaluates the internal cohesion of clusters without considering actual ground truth labels. While earlier research often compared a limited range of clustering algorithms based on the Silhouette Score metric, the main objective of this study is to pinpoint the most appropriate algorithm capable of accommodating the unique seismic attributes of Indonesia. A reliable clustering algorithm should also elucidate patterns of earthquakes, thus improving strategies for mitigation and preparedness.

2. Research Methods

Describe the research methods and research techniques The approach utilized in this research encompassed a systematic procedure consisting of five separate phases: gathering of data, preprocessing of data, implementation of clustering techniques, assessment of clustering methodologies, and ultimately, visualization and scrutiny of the cluster results. The procedural framework utilized in the investigation is depicted in Figure 1.

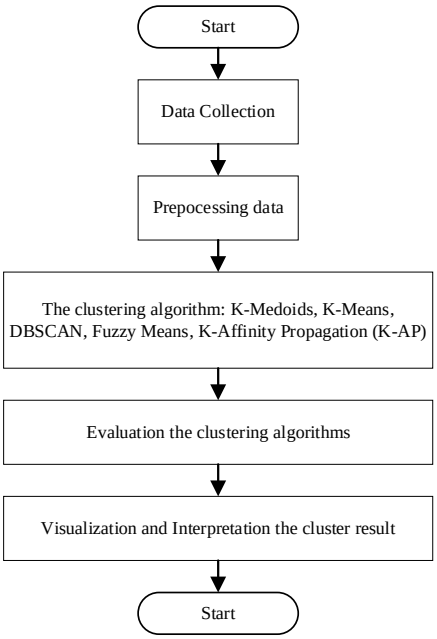

Figure 1. The Research Methodology

2.1 Data Collection

Acquiring seismic activity data in Indonesia spanning from January 2017 to January 2023 represents the foundational stage of the research. The seismic data employed in this investigation was acquired from the Indonesian Agency for Meteorological, Climatological, and Geophysics (BMKG) Contained within this dataset were a total of 13,860 seismic events registering magnitudes of three or higher, a significant observation [23]. The dataset comprises the exact coordinates (latitude and longitude), geographic position, depth, and magnitude of every seismic event. Table 1 provides a thorough analysis.

Table 1. Earthquake Data in Indonesia

Date	Lat	Lon	Depth	Mag	Location
01/01/2017	-6,02	103,78	10	4,1	Southwest of Sumatra, Indonesia
01/01/2017	-8,96	110	10	4	Java, Indonesia
...	...	...	...	...	...
31/01/2023	-3,43	127,43	10	4,5	Seram, Indonesia
31/01/2023	3,89	128,47	84	4,8	North of Halmahera, Indonesia

2.2 Data Preprocessing

Data preprocessing is the second step, which modifies the overall data format to make the data usable. The data cleaning, data selection, and data standardization steps make up this stage. Variable standardization is necessary because grouping and forecasting techniques may be disproportionately impacted by a variable having a much wider range than other variables. In order to guarantee balanced contributions in clustering and prediction models, all variables should be scaled uniformly [19].

Using ideas like mean and standard deviation, this method takes the unstructured data and scales, transforms, and organizes it before producing the normalized values or range of data. This process is known as Z-score Normalization [20]. Equation 1 is used to carry out this standardization process.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

The outcomes of the data standardization procedure for the chosen attributes are shown in Table 2.

Table 2. Earthquake Data in Indonesia

Date	Lat	Lon	Depth	Mag	Location
01/01/2017	-0,67	-1,63	-0,62	-0,99	Southwest of Sumatra, Indonesia
01/01/2017	-1,31	-1,07	-0,62	-1,25	Java, Indonesia
...	...	...	...	...	...
31/01/2023	-0,12	0,511	-0,62	-0,99	Seram, Indonesia
31/01/2023	1,45	0,60	0,06	0,25	North of Halmahera, Indonesia

2.3. Algorithms for Clustering

Using K-Medoids, K-Means, DBSCAN, Fuzzy C-Means, and K-AP clustering, the third stage involves cluster formation.

The K-Medoids algorithm is used to locate medoids in a cluster. The research steps for applying the K-

Medoids clustering method are as follows [18]: Choose k out of n data points at random to be the medoids; Connect every data point with the nearest medoid; calculate the total cost of the configuration, which is the average dissimilarity of o to all the data points associated with m, swap m and o for each medoid m and each data point o associated with m. Choose the configuration medoid that has the lowest cost; Continue steps 2 and 3 until the medoids remain unchanged.

One of the easiest algorithms to comprehend is K-means [24]. The research steps for utilizing K-means in optimization are as follows[25]: Choose k out of n data points at random to be the starting centroids; Assign each piece of data to the cluster whose centroid is closest; Determine the centroid and average of the data in each cluster anew; Continue steps 2-3 until either a stopping criterion is met or the centroids do not change.

DBSCAN is an algorithm that belongs to the class of density-based clustering, which is the process of creating clusters according to the degree of density between objects in the dataset in terms of proximity or distance. The DBSCAN algorithm has several advantages over other clustering algorithms, such as its ability to detect outliers/noise [26]. Additionally, DBSCAN can identify clusters of arbitrary shapes and does not require a number of clusters for initialization.[27] Here is a thorough step-by-step breakdown of the DBSCAN clustering method [26], [28], [29]: Determine every point that is within the current point's eps distance; Form a cluster if min_samples or more points are found within the eps distance. If not, label the point as irrelevant; Repeat steps 2 and 3 for each neighboring point. Include it in the cluster if it is dense; Repeat the procedure for each of the dataset's unexplored points. Once every point has been visited and classified as core, border, or noise, the algorithm comes to an end.

A clustering technique called fuzzy C-Means enables a single data point to be a part of several clusters with different levels of membership. When compared to hard clustering techniques, fuzzy C-Means yield a more nuanced clustering result. When the boundaries between clusters are unclear, it is especially helpful [30], [31]. The following steps can be used to summarize the FCM algorithm: Set up k cluster centers at random; Assign every piece of data, with varying membership levels, to every cluster; Each cluster center should be updated using the weighted average of all earthquakes, with the membership degrees serving as the weights; Until the cluster centers stabilize or a stopping criterion is satisfied, repeat steps 2-3.

The Affinity Propagation (AP) algorithm is modified by KAP to find the ideal number of sample sets. [32] The following steps can be used to describe the K-Affinity Propagation algorithm [4]: Start the similarity matrix with negative squared differences in the attributes (depth, magnitude, etc.) of the earthquakes; Refresh availability and responsibility matrices with an iterative set of "message-passing" rules; Determine the exemplars, or cluster centers, by adding the matrices for availability and responsibility; Assign the nearest exemplar to earthquakes; Continue steps 2-4 until a stopping criterion is satisfied or the exemplars stabilize.

2.4 Clustering Algorithm Evaluation

The assessment of clustering algorithms has significance in assessing their effectiveness in partitioning data into coherent clusters. In this research, various evaluation metrics will be presented to measure the performance of clustering algorithms including:

The Silhouette Score serves as an intrinsic evaluation metric designed to measure the degree of similarity between data points and their respective clusters in relation to alternative clusters. The Silhouette Score has a value range of -1 to 1. The quality of data clustering increases the closer the silhouette coefficient value is to 1. Conversely, the worse the clustering of data in a cluster, the closer the silhouette score value is to -1. [33] Equation 2 is used to determine the Silhouette score:

$$s = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (2)$$

The evaluation of Cluster Purity serves as a metric external to the clustering algorithm, employed for assessing the level of prevalence demonstrated by the predominant class within a specific cluster. The range of Cluster Purity values spans from 0 to 1, with a score of 1 denoting flawless clustering. The calculation of the Cluster Purity value is achievable by utilizing the subsequent formula in Equation 3.

$$Purity = \frac{1}{N}\sum_k max_j \left|C_k \cap L_j\right| \quad (3)$$

The Silhouette Score is utilized for the purpose of optimizing clustering hyperparameters and assessing the effectiveness of clustering algorithms. Through the incorporation of various clustering validity metrics like the Silhouette Coefficient and Cluster Purity can effectively evaluate and improve clustering results, thereby achieving more accurate and meaningful insights from the data.

The ultimate phase involves the visual representation and analysis of the top cluster outcomes.

3. Results and Discussions

We used the k-medoids technique to classify the data in our research of 13,860 earthquake cases from Indonesia into groups of two to ten clusters. We identified the clusters and their centers, and then calculated the Euclidean Distance between each center and the closest non-centroid earthquake data point. The silhouette method was employed in our study to determine the optimal number of clusters. Better clustering quality is indicated by a higher silhouette value. Figure 2 shows how our findings are represented visually:

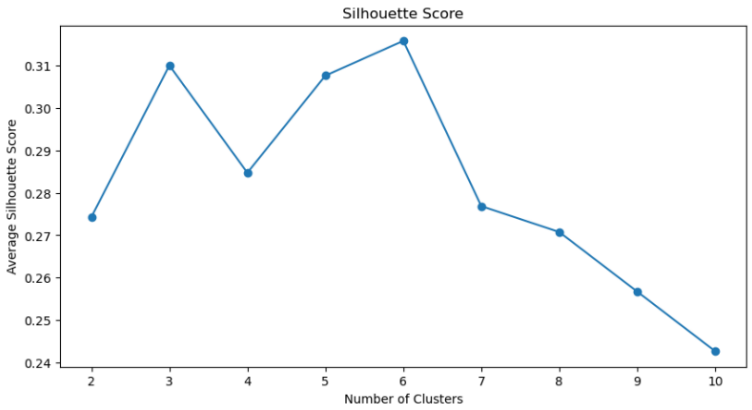

Figure 2. K-Medoids silhouette average graph

The silhouette value noticeably decreases to somewhat less than 0.25 when we use the k-medoids approach, as shown in Figure 2, with k assigned at 10. Conversely, when k is set to 6, we observe a significant rise, with the value rising over 0.31. Table 3 provides a detailed summary of the silhouette results obtained using the k-medoids approach on the full dataset.

Table 3. K-Medoids for Silhouette Score

No	Number of Clusters	Silhouette Score
1	2	0.274307
2	3	0.310028
3	4	0.284735
4	5	0.307621
5	6	0.315820
6	7	0.276869
7	8	0.270761
8	9	0.256713
9	10	0.242656

The maximum silhouette value, with k=6, is 0.315820, according to Table 3.

Table 4 displays the K-Means cluster results. Table 4 demonstrates that, with a silhouette value of 0.324548, six clusters is the ideal amount for K-Means.

Table 4. Shilhoutte Score Using K-Means

No	Number of Clusters	Silhouette Score
1	2	0.273463
2	3	0.321254
3	4	0.308797
4	5	0.312178
5	6	0.324548
6	7	0.297986
7	8	0.313608
8	9	0.315864
9	10	0.319170

The number of clusters from the DBSCAN clustering using various Eps and MinPts settings is shown in Table 5. The present study employs a MinPts interval of 3 to 6 and an Eps value of 0.7 to 0.9.

Table 5 shows that, with two clusters produced and 29 noise, the maximum silhouette score is 0.563362, or at MinPts=4 and Eps=0.9. The items in the dataset appear to be quite strongly grouped based on these results; that is, they have low intra-cluster distances and high inter-cluster distances (the distance between an object and other objects in the same cluster). This shows that although each cluster member has a very high density with other items in the same cluster, each cluster is relatively isolated from the others [26].

Table 5. Silhouette Score Calculated with DBSCAN

No	Eps	MinPts	Shilhoutte Score	Number of Clusters	Noise
1	0.7	3	0.345012	5	48
2	0.7	4	0.469739	4	63
3	0.7	5	0.256496	2	81
4	0.7	6	0.533721	1	93
5	0.8	3	0.540739	2	33
6	0.8	4	0.540311	3	36
7	0.8	5	0.551297	2	44
8	0.8	6	0.550882	2	53
9	0.9	3	0.541971	2	21
10	0.9	4	0.563362	2	29
11	0.9	5	0.560521	2	31
12	0.9	6	0.553540	3	34

3.4 Fuzzy C-Means Cluster Analysis

Table 6 illustrates that the ideal number of clusters when utilizing Fuzzy C-Means is 5, with a Shilhoutte Score of 0.297185.

Table 6. Fuzzy C-Means-Based Shilhoutte Score

No	Number of Clusters	Silhouette Score
1	2	0.253232
2	3	0.296397
3	4	0.276139
4	5	0.297185
5	6	0.269957
6	7	0.268747
7	8	0.288997
8	9	0.290241
9	10	0.276212

The Shilhoutte score and number of clusters using the K-Affinity Propagation technique are displayed in Figure 3. With 196 clusters, the optimal Shilhoutte value was 0.2550.

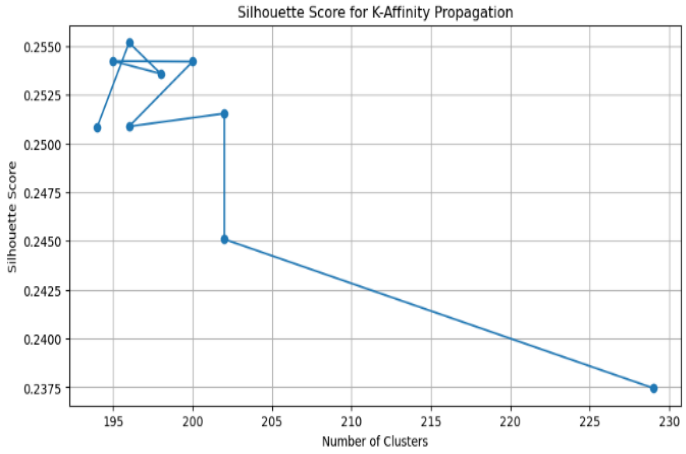

Figure 3. Silhouette average graph from K-AP

To find a better technique for clustering earthquake data in Indonesia from January 2017 to January 2023, the K-Medoids, K-Means, DBSCAN, Fuzzy C Means, and K-Affinity Propagation algorithms were used to analyze the data. The results were then compared. The optimal silhouette score is calculated by comparing the cluster results to decide which cluster is the best.[34] Figure 4 displays a visualization plot of the average silhouette score for every cluster created using the five techniques.

The plot of the average silhouette score for each cluster in each of the five algorithms is shown to be negative in Figure 4. This suggests the data has been placed in the incorrect cluster.[35] DBSCAN and K Means, two of the five algorithms, have the best Shilhoutte scores when compared to the other techniques, at 0.563362 and 0.324548, respectively. This suggests that there are strong connections between the objects in the cluster. Nevertheless, if the silhouette score in any of the first three clusters is negative, it means that the data belongs to the incorrect cluster. Nevertheless, negative silhouette scores are found in DBSCAN cluster one, while negative silhouette coefficients are found in K-Means algorithm clusters one, four, and six. This signifies that data has entered the wrong cluster.

Based on the results, the silhouette coefficient values obtained by the two algorithms show quite considerable disparities. If viewed from the average silhouette coefficient value of the two algorithms, the DBSCAN algorithm is a better strategy for clustering earthquake data. However, looking at the silhouette coefficient plot of the two algorithms, both techniques exhibit negative silhouette widths. This indicates a mistake in the clustering procedure. The error value in the silhouette width of the K-Means algorithm is less than the DBSCAN algorithm. Therefore, the K-Means approach was determined as a better technique for grouping earthquake data in Indonesia from January 2017 to January 2023.

Moreover, in order to assess the efficacy of the data clustering generated by the five clustering algorithms, cluster purity is employed as a metric to gauge the degree to which items grouped within a particular cluster share identical labels or classes, particularly in the context of clustering seismic regions in Indonesia. This evaluation is conducted using the BMKG catalogue to quantify the cluster purity[1], The number of clustering of earthquake areas in Indonesia based on Depth and Magnitude consists of 9 clusters. Table 7 shows the validation test with silhouette score and cluster purity.

Table 7. Validation Test

Cluster Methods	Number of Clusters	Silhouette Score	Cluster Purity
K-Medoids	6	0.315820	0.718975
K-Means	6	0.324548	0.736580
DBSCAN	2	0.563362	0.567965
Fuzzy C-Means	5	0.297185	0.691341
K-AP	196	0.2550	0.931313

From Table 7 shows that K-Medoids method produced 6 clusters with a moderate Silhouette Score of 0.315820, indicating that the clusters are reasonably compact and well-separated. The Cluster Purity of 0.718975 suggests that approximately 72% of the points within each cluster belong to the dominant class. Similar to K-Medoids, K-Means also produced 6 clusters but with a slightly higher Silhouette Score of 0.324548, indicating better-defined clusters. The Cluster Purity of 0.736580 is also higher, indicating that around 74% of the points within each cluster belong to the dominant class.

Fuzzy C-Means method produced 5 clusters with the lowest Silhouette Score of 0.297185 among the non-Affinity Propagation methods, indicating less defined clusters. The Cluster Purity of 0.691341 indicates that approximately 69% of the points within each cluster belong to the dominant class. K-Affinity Propagation produced a very large number of clusters (196) with the lowest Silhouette Score of 0.2550, indicating poorly defined clusters. However, it has the highest Cluster Purity of 0.931313, suggesting that the clusters formed are very pure, with about 93% of the points within each cluster belonging to the dominant class. This result indicates that while clusters are highly pure, the large number of clusters might indicate overfitting.

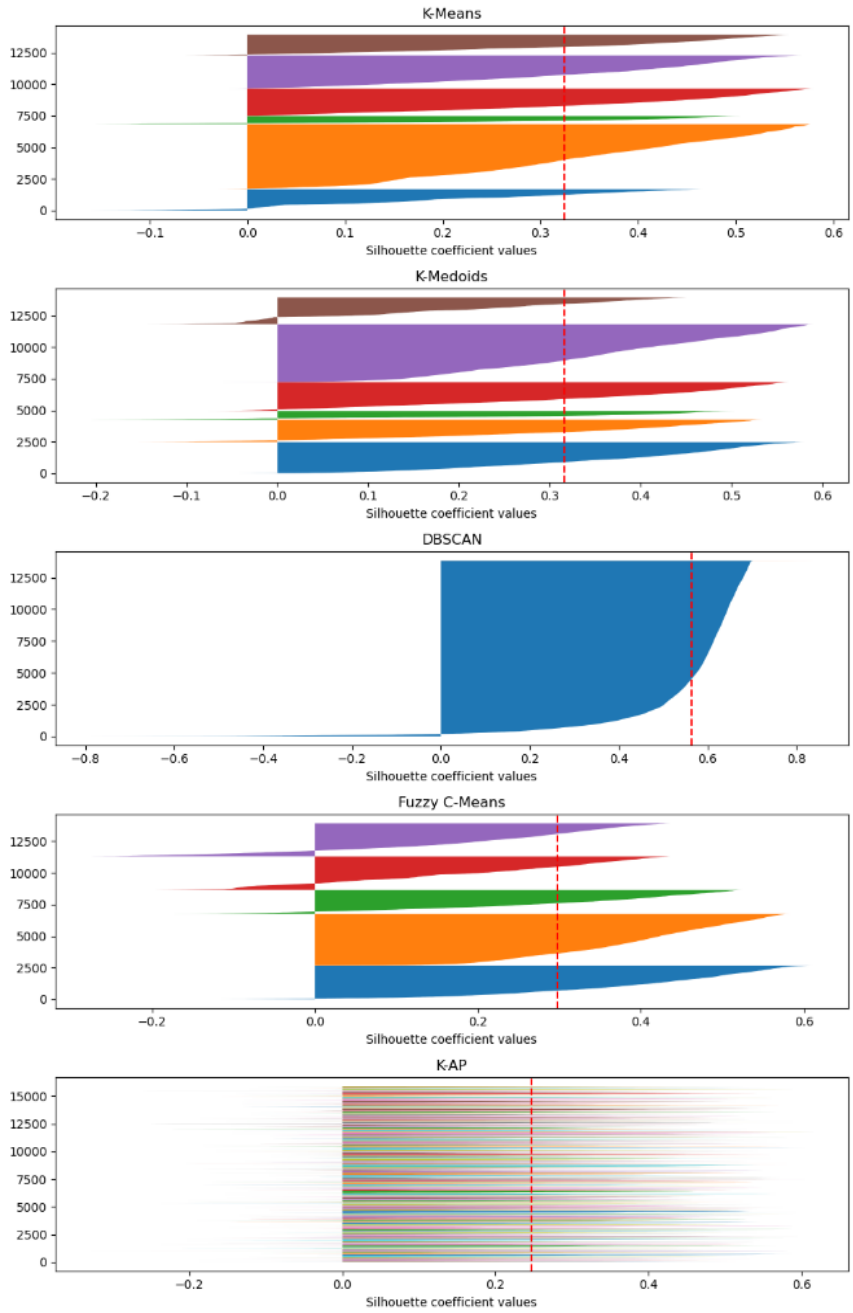

Figure 4. Average silhouette scores for K-Means, K-Medoids, DBSCAN, Fuzzy C Means, and K-AP are plotted in a visualization

While DBSCAN exhibits a lower Cluster Purity, its elevated Silhouette Score and capacity to manage noise render it a robust contender for clustering seismic data in Indonesia. Nevertheless, the negative Silhouette values depicted in Figure 4 signify substandard clustering efficacy and the presence of overlapping clusters. Considering these factors, K-Means emerges as the most appropriate clustering technique for seismic data in Indonesia due to its ability to strike a favorable equilibrium between Silhouette Score (0.324548) and Cluster Purity (0.736580). This results in well-defined and easily interpretable clusters that demonstrate a certain level of homogeneity, thus proving to be well-suited for practical applications in the analysis of seismic events.

After finding that the K-Means technique is a better approach for grouping earthquake data in Indonesia, the following step is to describe the earthquake data from the clustering results.

From Table 8, it is known that earthquakes included in clusters 2 to 5 are shallow earthquakes with less than 70 kilometers, with typical damage being light and moderate in class. In this cluster, earthquakes are regularly felt but only cause small damage including damage the building and other structures.[35] Meanwhile, the earthquakes contained in clusters 1 is intermediate where its depth between 70 km up to 300 km, and cluster 6 is deep earthquakes with a depth of more than 300 kilometers and minor up to strong damages. In this cluster, the earthquake caused modest

damage to buildings and other structures, up to serious damage.

An image of mapping earthquake zones in Indonesia based on K-Means clustering algorithm illustrated in Figure 5.

Table 8. Descriptive Statistic for The Cluster Result of K-Means

Cluster	Variabel	Statictic				Cluster Members
		Mean	Median	Min	Max	
1	Latitude	-6.7273	-6.89	-10.49	1.28	2164
	Longitude	128.7728	130	107.14	140.76	
	Depth	152.17	157.00	10	313	
	Magnitude	4.51	4.50	4	5.4	
2	Latitude	-0.2066	-0.04	-10.86	6	1682
	Longitude	126.8654	127	96.05	140.97	
	Depth	49.59	34.00	5	352	
	Magnitude	5.16	5.00	4.8	7.5	
3	Latitude	-1.0336	-0.77	-9.27	5.99	1633
	Longitude	99.4685	100	94.7	113.53	
	Depth	42.00	24.00	5	266	
	Magnitude	4.51	4.50	4	5.7	
4	Latitude	-8.6599	-8.68	-11	-3.5	2624
	Longitude	114.5574	116	101.25	130.42	
	Depth	28.91	12.00	5	233	
	Magnitude	4.40	4.30	4	5.3	
5	Latitude	0.3792	0,11	-6,77	6	5125
	Longitude	127,7926	127	115,61	140,92	
	Depth	42,93	14,00	5	280	
	Magnitude	4,32	4,30	4	4,7	
6	Latitude	-3,6876	-6,18	-8,1	5,93	632
	Longitude	123,6913	125	96,93	129,99	
	Depth	451,62	442,00	265	700	
	Magnitude	4,53	4,50	4	7,2	

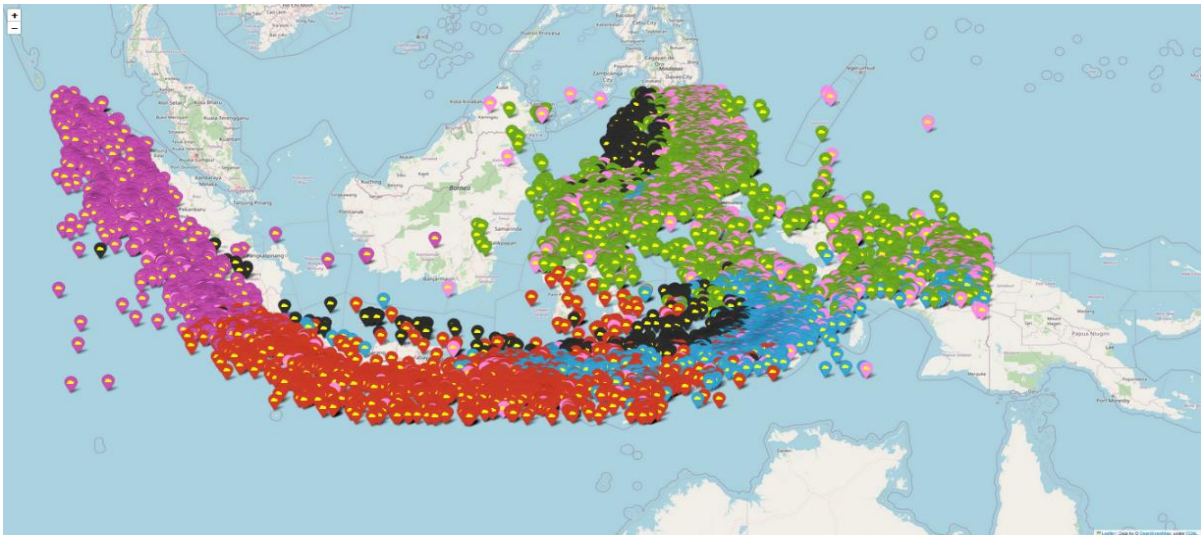

Figure 5. Results of Mapping K-Means Clusters with k=6

From the cluster results using the K-Means algorithm, it was revealed that cluster 1 had an average magnitude of 4.51 M and an average depth of 152,17 km with 2164 earthquake occurrences occurring in 22 regions of Indonesia, including the Arafura Sea, Aru Islands Region, Bali Region, and its environs. This cluster, on average, has an intermediate earthquake depth (70-300 km). Cluster 2 has an average magnitude of 5.16 and an average depth of 49.59, with the most earthquakes occurring at 1682 in 26 regions of Indonesia, including the Aru Islands, Banda Sea, Kalimantan, Buru, and adjacent areas. This cluster is shallow earthquakes with mild seismic consequences.

Cluster 3 has average magnitude of 4.51, with an average depth of 42.00 and an earthquake occurrence rate of 1633 across 46 regions of Indonesia. Earthquakes in this cluster caused modest damage. Cluster 4 has multiple earthquake events, 2624 with an average magnitude of 4.40 and an average depth of 28.91. This cluster is classified a shallow earthquake because it has an average depth (<60 kilometers). There are 21 Indonesian regions in cluster 4, including the

Bali Region, Bali Sea, Banda Sea, Flores Region, Flores Sea, and neighboring territories.

Meanwhile, cluster 5 has an average magnitude of 4.32 and an average depth of 42.93, with the number of earthquakes happening being 5125 across 9 regions of Indonesia, including Kalimantan, Java Sea, Java, North Sumatra, and adjacent areas. The last cluster is cluster 6, which had a minor earthquake incidence compared to other clusters, namely 632 earthquakes with an average magnitude of 4.53 and an average depth of 451.62 in 9 regions of Indonesia, namely the Bali Sea, Banda Sea, Sulawesi Sea, Flores Sea, and adjacent areas. From the results of the clustering mapping, it can be concluded that there are several areas that need to carry mitigation steps because of the considerable potential for disasters that will occur, such as areas included in clusters 2 where these areas have an average of earthquake magnitude 5.16, namely moderate in class and also has average of shallow earthquake.

The implementation of the K-Means clustering algorithm on seismic activity data in Indonesia spanning from January 2017 to January 2023 resulted in the establishment of six discrete clusters. Each cluster was analyzed based on key seismic attributes such as latitude, longitude, depth, and magnitude. The visualization of these clusters, as presented in Table 8, offers insights into the geographical and seismological patterns across the Indonesian archipelago. Previous research results k=3 for earthquake clustering in Indonesia using K-Means [2], [36], but the data used was only up to 2019. This research has included earthquake data until 2023 so the data is the latest ones.

The K-Means algorithm identified that Cluster 2, which includes 26 regions of Indonesia such as the Aru Islands and Banda Sea, has the highest average earthquake magnitude (5.16) and shallow depth (49.59 km). This suggests a higher likelihood of seismic events causing moderate damage in these regions, necessitating targeted disaster preparedness and mitigation strategies. Clusters 1 and 6, with deeper earthquakes (average depths of 152.17 km and 451.62 km, respectively), predominantly encompass regions like the Arafura Sea and Flores Sea. These clusters signify seismic phenomena that are less probable to inflict surface harm; nonetheless, they present considerable hazards owing to the possibility of aftershocks and ancillary consequences such as tsunamis.

The K-Means algorithm, which generated six clusters, achieved the highest Silhouette Score of 0.324548 and a Cluster Purity of 0.736580. This indicates that the clusters formed are both well-defined and exhibit a high degree of homogeneity. In comparison to earlier studies, which often relied solely on Silhouette Score or other single metrics, our dual-metric approach offers a more nuanced evaluation. For example, previous research by [1] reported a Silhouette Score of 0.4574067 using the K-Medoids algorithm, but it did not incorporate Cluster Purity, which is crucial for understanding the actual grouping accuracy in a seismic context. In contrast, the inclusion of Cluster Purity in our analysis reveals the extent to which earthquakes within the same cluster share similar characteristics, providing deeper insights into the reliability of the clustering.

Moreover, the findings show that the K-Means algorithm outperforms not only K-Medoids but also other clustering methods such as DBSCAN, Fuzzy C-Means, and K-Affinity Propagation (K-AP). DBSCAN, while effective in handling noise, exhibited a negative Silhouette Score, indicating poor clustering quality. Fuzzy C-Means and K-AP were less effective, with K-AP producing an impractically high number of clusters (196), leading to a lower overall effectiveness despite its high Cluster Purity. This demonstrates that while DBSCAN and K-AP have their strengths in specific contexts, K-Means provides a more balanced and practical approach for seismic data clustering.

The novelty of this study lies in its comprehensive comparative analysis, which is the first of its kind to apply both Silhouette Score and Cluster Purity metrics to earthquake data clustering in Indonesia. In this research we compares five methods that have never been carried out by previous research, where at most they only compare three methods [15], [18]. This dual approach not only confirms the internal coherence of the clusters but also verifies their external validity by ensuring that the clusters reflect actual seismic groupings. This is a significant advancement over previous studies, which often did not consider how well the clusters correspond to real-world seismic phenomena.

4. Conclusions

Earthquake data in Indonesia from January 2017 to January 2023 has been effectively grouped using the K-Medoids, K-Means, DBSCAN, Fuzzy C Means and K-Affinity Propagation algorithms. From the analysis results, the K-Medoids and K-Means algorithms formed 6 clusters using the silhouette approach. In comparison, the DBSCAN algorithm succeeded in generating 2 clusters, Fuzzy C-Means succeeded in generating 5 clusters, and K-Affinity succeeded in generating 196 clusters. Based on the silhouette scores and cluster purity acquired by the five algorithms, K-Means is the most appropriate clustering algorithm for earthquake data in Indonesia. It provides a balanced performance with well-defined, homogeneous clusters that are easy to interpret and practical for real-world applications in seismic activity analysis. Future research should investigate adding another evaluation model to show a more ideal comparison of clustering findings.

Acknowledgements

The author would like to thanks to the Directorate of Research, Technology, and Community Service, and the Directorate General of Higher Education, Ministry of Education, for supporting this research under Contract No. 179/E5/PG.02.00.PL/2023, dated 19 June 2023.

References

[1] F. R. Senduk, Indwiarti, and F. Nhita, "Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm," Indones. J. Comput., vol. 4, no. 3, pp. 65–76, 2019, doi: 10.21108/indojc.2019.4.3.359.

[2] A. Jufriansah, Y. Pramudya, A. Khusnani, and S. Saputra, "Analysis of Earthquake Activity in Indonesia by Clustering Method," J. Phys. Theor. Appl., vol. 5, no. 2, p. 92, 2021, doi: 10.20961/jphystheor-appl.v5i2.59133.

[3] M. N. Shodiq, D. H. Kusuma, M. G. Rifqi, A. R. Barakbah, and T. Harsono, "Neural Network for Earthquake Prediction Based on Automatic Clustering in Indonesia," Int. J. Informatics Vis., vol. 2, no. 1, pp. 37–43, 2018, doi: 10.30630/joiv.2.1.106.

[4] M. Muhajir and N. N. Sari, "K-Affinity Propagation (K-AP) and K-Means Clustering for Classification of Earthquakes in Indonesia," in Proceeding - 2018 International Symposium on Advanced Intelligent Informatics: Revolutionize Intelligent Informatics Spectrum for Humanity, SAIN 2018, IEEE, 2018, pp. 6–10. doi: 10.1109/SAIN.2018.8673344.

[5] B. A. P. Martadiputra, D. Rachmatin, and A. S. Hidayat, "Analysis of Characteristics of Earthquake Area in Indonesia in 2020 with Cluster Analysis as Natural Disaster," Int. J. Sci. Res., vol. 9, no. 11, pp. 1243–1250, doi: 10.21275/SR201122121148.

[6] M. N. Shodiq, A. Ridho Barakbah, and T. Harsono, "Spatial Analysis of Earthquake Distribution with Automatic Clustering for Prediction of Earthquake Seismicity in Indonesia," in The Forurth Indonesian-Japanese Conference on Knowledge Creation dan Intelegent Computing (KCIC) 2015, pp. 47–55.

[7] A. Sabtaji, "Statistics of Tectonic Earthquake Events Each Proovince in Indonesia Territory for 11 Years of Observation (2009-2019)."

[8] Prihandoko, Bertalya, and M. I. Ramadhan, "An Analysis of Natural Disaster Data by Using K-Means and K-Medoids Algorithm of Data Mining Techniques," in 2017 15th International Conference on Quality in Research (QiR) : International Symposium on Electrical and Computer Engineering, Nusa Dua, Bali, Indonesia: IEEE, Jul. 2017, pp. 221–225. doi: 10.1109/QIR.2017.8168485.

[9] P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means Cluster Analysis in Earthquake Epicenter Clustering," Int. J. Adv. Intell. Informatics, vol. 3, no. 2, pp. 81–89, doi: 10.26555/ijain.v3i2.100.

[10] A. Ahmad and N. Irsalinda, "Cluster Analysis of Earthquake's Data Clustering in Indonesia using Fuzzy K-means Clustering," in Proceeding International Conference on Science and Engineering, pp. 3–7. doi: 10.14421/icse.v3.457.

[11] M. Nishom, S. F. Handayani, and D. Dairoh, "Pillar Algorithm in K-Means Method for Identification Health Human Resources Availability Profile in Central Java," JUITA J. Inform., vol. 9, no. 2, p. 145, 2021, doi: 10.30595/juita.v9i2.9860.

[12] A. Kusmiran, "Clustering and Risk Analysis of The Earthquake in Sulawesi Using Mini Batch K-Means, K-Medoids, and Maximum Likelihood Method," Elkawnie J. Islam. Sci. Technol., vol. 9, no. 1, pp. 1–23, doi: 10.22373/ekw.v8i2.13027.

[13] R. Adolph, "Analysis of Seismic Data Using Partition-Based Clustering Techniques," in 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT), New Delhi, India, 2022, pp. 1–6. doi: 10.1109/GlobConPT57482.2022.9938362.

[14] A. Ashabi, S. Bin Sahibuddin, and M. S. Haghighi, "The Systematic Review of K-Means Clustering Algorithm," in ICNCC '20: Proceedings of the 2020 9th International Conference on Networks, Communication and Computing, pp. 13–18. doi: https://doi.org/10.1145/3447654.3447657.

[15] I. H. Rifa, H. Pratiwi, and Respatiwulan, "Implementasi Algoritma Clara untuk Data Gempa Bumi di Indonesia," in Seminar Nasional Penelitian Pendidikan Matematika (SNP2M), 2019, pp. 161–166. doi: http://dx.doi.org/10.31000/cpu.v0i0.1694.

[16] H. V. Singh, A. Girdhar, and S. Dahiya, "A Literature survey based on DBSCAN algorithms," in 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India: IEEE, May 2022, pp. 751–758. doi: 10.1109/ICICCS53718.2022.9788440.

[17] Mustakim, E. Rahmi, M. R. Mundzir, S. T. Rizaldi, Okfalisa, and I. Maita, "Comparison of DBSCAN and PCA-DBSCAN Algorithm for Grouping Earthquake Area," in 2021 International Congress of Advanced Technology and Engineering (ICOTEN), IEEE, Jul. 2021, pp. 1–5. doi: 10.1109/ICOTEN52080.2021.9493497.

[18] Mustakim, M. Z. Fauzi, Mustafa, A. Abdullah, and Rohayati, "Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms," J. Phys. Conf. Ser., vol. 1783, no. 1, p. 012016, Feb. 2021, doi: 10.1088/1742-6596/1783/1/012016.

[19] R. R. Arista, R. A. Asmara, and D. Puspitasari, "Pengelompokan Kejadian Gempa Bumi Menggunakan Fuzzy C-Means Clustering," J. Teknol. Inf. dan Terap., vol. 4, no. 2, pp. 103–110, Apr. 2019, doi: 10.25047/jtit.v4i2.67.

[20] B. Tavakkol and Y. Son, "Fuzzy Kernel K-Medoids Clustering Algorithm for Uncertain Data Objects," Pattern Anal. Appl., vol. 24, no. 3, pp. 1287–1302, Aug. 2021, doi: 10.1007/s10044-021-00983-z.

[21] M. N. Bangun, O. Darnius, and S. Sutarman, "Optimization Model in Clustering The Hazard Zone After an Earthquake Disaster," Sink. J. dan Penelit. Tek. Inform., vol. 7, no. 3, pp. 2089–2095, Aug. 2022, doi: 10.33395/sinkron.v7i3.11598.

[22] I. H. Rifa, H. Pratiwi, and R. Respatiwulan, "Clustering of Earthquake Risk in Indonesia Using K-Medoids and K-Means Algorithms," Media Stat., vol. 13, no. 2, pp. 194–205, Dec. 2020, doi: 10.14710/medstat.13.2.194-205.

[23] N. Dwitiyanti, S. Ayu Kumala, and S. Dwi Handayani, "Implementation of K-Means Method in Classterization of Earthquake Prone Areas in Indonesia," Pros. Semin. Nas. UNIMUS, vol. 6, pp. 1029–1037, 2023.

[24] C. Shah and A. Jivani, "Comparison of data mining clustering algorithms," in 2013 Nirma University International Conference on Engineering (NUiCONE), IEEE, Nov. 2013, pp. 1–4. doi: 10.1109/NUiCONE.2013.6780101.

[25] R. Refianti, A. B. Mutiara, A. Juarna, and S. N. Ikhsan, "Analysis and Implementation of Algorithm Clustering Affinity Propagation and K-Means at Data Student Based On GPA and Duration of Bachelor-Thesis Completion," J. Theor. Appl. Inf. Technol., vol. 35, no. 1, pp. 69–76.

[26] M. Tanzil 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, Jul. , doi: 10.21776/ub.jeest.2016.003.01.1.

[27] S. Augustine and S. Augustine, "Comparative Performance Analysis of Clustering Techniques in Educational," IADIS Int. J. Comput. Sci. Inf. Syst., vol. 10, no. 2, pp. 65–78, 2015.

[28] J. Han, M. Kamber, and J. Pei, Data Mining. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.

[29] A. Kristianto, E. Sediyono, and K. D. Hartomo, "Implementation DBSCAN Algorithm to Clustering Satellite Surface Temperature Data in Indonesia," Regist. J. Ilm. Teknol. Sist. Inf., vol. 6, no. 2, pp. 109–118, 2020, doi: 10.26594/register.v6i2.1913.

[30] C. S. Putra, "Usulan penelitian dosen pemula," no. Anggota 1, pp. 1–27, 2015.

[31] Y. Lu, T. Ma, C. Yin, X. Xie, W. Tian, and S. M. Zhong, "Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data," Int. J. Database Theory Appl., vol. 6, no. 6, pp. 1–18, 2013, doi: 10.14257/ijdta.2013.6.6.01.

[32] X. Zhang, W. Wang, and K. Nørv, "K-AP : Generating Specified K Clusters by Efficient Affinity Propagation K -AP : Generating Specified K Clusters by Efficient Affinity Propagation," no. December, 2010, doi: 10.1109/ICDM.2010.107.
[33] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 9, pp. 723–732, doi: 10.1109/EUMC.2008.4751704.
[34] M. Bariklana and A. Fauzan, "Implementation of The DBSCAN Method for Cluster Mapping of Earthquake Spread Location," BAREKENG J. Math. Its Appl., vol. 17, no. 2, pp. 867–878, 2023.
[35] A. Kusmiran, "Clustering and Risk Analysis Of The Earthquake in Sulawesi Using Mini Batch K-Means , K-Medoids, and Maximum Likelihood Method," J. Islam. Sci. Technol., vol. 9, no. 1, pp. 1–23, 2023, doi: 10.22373/ekw.v9i1.13027.
[36] A. Prasetio, M. M. Effendi, and M. N. Dwi M, "Analisis Gempa Bumi Di Indonesia Dengan Metode Clustering," Bull. Inf. Technol., vol. 4, no. 3, pp. 338–343, Sep. 2023, doi: 10.47065/bit.v4i3.820.

