2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI) | 979-8-3315-0861-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICoCSETI63724.2025.11020099

Mapping Indonesia's Regions Based on Carbon Emissions Using the K-Means Algorithm

William
Fakultas Teknologi Informasi
Universitas Tarumanagara
Jakarta, Indonesia
william.535210013@stu.untar.ac.id

Luhur Bayuaji
Fac. of Data Science and Information Tech.
INTI International University
Kuala Lumpur, Malaysia
luhur.bayuaji@newinti.edu.my

Novario J. Perdana
Fakultas Teknologi Informasi
Universitas Tarumanagara
Jakarta, Indonesia
novariojp@fti.untar.ac.id

Teny Handhayani
Fakultas Teknologi Informasi
Universitas Tarumanagara
Jakarta, Indonesia
tenyh@fti.untar.ac.id

***Abstract*—Carbon emissions, also known as greenhouse gas (GHG) emissions, refer to the release of gases that trap heat in the Earth's atmosphere, contributing to the greenhouse effect and climate change. These gases, primarily carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) are produced through various human activities. GHGs are generated from sectors such as energy, industry, agriculture, forestry, and waste, each contributing emissions with distinct characteristics in every region. GHG data from 34 provinces in Indonesia from 2000 to 2023 were processed using the K-Means algorithm for clustering to facilitate the analysis of emission patterns, Clustering was based on the similarity of emission characteristics across these five sectors. Clustering results were evaluated using the Silhouette Coefficient to assess the quality of the grouping. Visualization in an interactive map allows users to understand the distribution patterns of emissions between provinces. The analysis process includes several stages from steps of data collection, data preprocessing, clustering, evaluation, and visualization. The K-Means algorithm has proven effective in grouping provinces based on the similarity of GHG emission profiles in each sector as well as combined sectors. Evaluation using the Silhouette Coefficient showed that clustering data into three clusters obtained an average score of 0.62. This result indicates a medium level of similarity among provinces within a cluster. Riau Province was identified as the highest emitter, while Papua and West Papua were recognized as the provinces acting as the highest absorbers. The interactive map successfully demonstrated the spatial distribution of emissions.**

***Index Terms*—clustering, greenhouse gas emissions, k-means, silhouette coefficient, spatial visualization.**

I. Introduction

Carbon emissions, also known as greenhouse gas (GHG) emissions, refer to the release of gases that trap heat in the Earth's atmosphere, contributing to the greenhouse effect and climate change. These gases, primarily carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) are produced through various human activities [1]. GHG emissions in Indonesia have become one of the most critical environmental issues in recent decades, considering the country's contribution to global climate change. According to Climate Watch data on its official website, Indonesia ranked 6th as the largest GHG emitter in the world in 2021, following China, the United States, India, Russia, and Brazil. In that year, Indonesia contributed 1,480,000 Gigagrams of CO_2 equivalent [2]. One way to measure and compare GHG emissions from various sources is by using CO_2 equivalent (CO_2 eq). Carbon dioxide equivalent (CO_2 eq) is a metric used to compare emissions of different GHGs based on their global warming potential (GWP). Each type of GHG has a different GWP, and CO_2 eq converts the amount of emissions from each GHG into CO_2 equivalent based on its GWP [3].

This study aims to cluster provinces in Indonesia according to the carbon emissions. The carbon emission data is analyzed at the provincial level in Indonesia from 2000 to 2023, covering various crucial parameters such as emissions from the energy, industry, agriculture, forestry, and waste sectors. The dataset is obtained from the Central Bureau of Statistics or *Badan Pusat Statistik (BPS)* and the Ministry of Environment and Forestry or *Kementerian Lingkungan Hidup Dan Kehutanan (KLHK)*. The K-Means and its variants were applied due to their proven versatility across various fields. Various studies show that K-means and its variants have given a good performance in many fields such as agricultural productivity [4] [5] [6] [7], public health mapping [8] [9] [10] [11] [12], disaster risk assessment [13], organizational distribution [14], environmental mapping [15] [16] [17], and educational performance [18] [19]. This versatile unsupervised learning method has proven effective in grouping and analyzing various regional characteristics, making it particularly suitable for province-level clustering analysis. The silhouette coefficient is used as a measure to assess how well the provinces are grouped according to their emission characteristics. The results of this paper are expected to identify priority areas for efficient and effective policy interventions.

979-8-3315-0861-6/25/$31.00 ©2025 IEEE

II. Methods

A. K-Means Algorithm

K-Means clustering is an unsupervised learning algorithm widely used to divide a dataset into k clusters based on feature similarity. This algorithm works by iteratively assigning data points to clusters defined by centroids, aiming to minimize variance within each cluster. The algorithm starts by randomly selecting k initial centroids, after which each data point is assigned to the nearest centroid. The centroids are then recalculated as the average of all points in the cluster, and this process repeats until the centroids stabilize. Although simple and efficient, K-Means clustering is sensitive to the initial centroid selection and may converge to suboptimal solutions, especially with non-spherical or complex datasets. However, variations and enhancements to the standard K-means algorithm, such as performing multiple trials or integrating dimensionality reduction techniques like Principal Component Analysis (PCA), can help improve robustness and accuracy in identifying reliable clusters [20].

Below are the steps for using the K-Means clustering algorithm [21]:

1) Determine the number of clusters to be formed.
2) Determine the initial cluster center values (centroids).
3) Calculate the distance of each input data to all centroids using the Euclidean Distance formula below until the result with the closest distance to the centroid is obtained.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

Explanation:

- d = distance.
- n = number of attributes.
- q = centroid.
- p = data.

4) Assign each data point based on the closest distance to the centroid.
5) Update the centroid value with a new centroid obtained from the average calculation of the cluster using the formula below:

$$v_i^w = \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^w \tag{2}$$

Explanation:

- i, w = index of the cluster.
- w = index of the variable.
- v_i^w = centroid/average of the i-th cluster for the w-th variable.
- X_k^w = value of the k-th data in the cluster for the w-th variable.
- N_i = number of data points in the i-th cluster.

6) Repeat steps 3 to 5 until no data points change clusters.

B. Silhouette Coefficient

The Silhouette Coefficient is a metric used to evaluate the quality of clusters in a dataset, particularly in the context of community detection in networks. This metric measures the similarity of an object to its own cluster relative to other clusters. The coefficient is calculated for each data point and ranges from -1 to 1, where a value close to 1 indicates that the data point is well-categorized within its cluster, a value close to 0 indicates that the point is on or very near the decision boundary between two neighboring clusters, and a negative value indicates that the point may have been placed in the wrong cluster [22].

The steps for calculating the silhouette coefficient can be explained as follows [23]:

1) Calculate the average distance of the i-th object to all other objects within the same group A.

$$a(i) = \frac{1}{|A|-1}\sum_{j\in A, j\neq i} d(i,j) \tag{3}$$

where j is another object in the same group A and $d(i,j)$ is the distance between objects i and j.

2) Calculate the average distance of the i-th object to all objects in another group.

$$d(i,C) = \frac{1}{|C|}\sum_{j\in C} d(i,j) \tag{4}$$

where $d(i,C)$ is the average distance of object i to all objects in another group C where $A \neq C$.

3) Determine the minimum value $b(i)$ which shows the average difference of object i for the group closest to its neighbor.

$$b(i) = \min_{C\neq A} d(i,C) \tag{5}$$

4) Calculate the silhouette value.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{6}$$

The result of $s(i)$ ranges from -1 to 1. The value $s(i)$ can be interpreted as:

- $s(i) \approx 1$ means object i is located in the correct group (within A),
- $s(i) \approx 0$ means object i is between two groups (A and B),
- $s(i) \approx -1$ means object i is in the wrong group (closer to B than A).

5) Calculate the silhouette coefficient defined as the average of $s(i)$.

$$SC = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{7}$$

where n is the number of observations.
The best clustering is achieved if SC is maximized, meaning minimizing intra-cluster distance $a(i)$ while maximizing inter-cluster distance $b(i)$. The silhouette coefficient value is categorized as shown in Table 2.1.

TABLE I
SILHOUETTE COEFFICIENT CATEGORIES AND INTERPRETATIONS [12].

Silhouette Coefficient	Interpretation
$0.7 < SC \leq 1$	Strong structure between objects and formed groups
$0.5 < SC \leq 0.7$	Medium structure between objects and formed groups
$0.25 < SC \leq 0.5$	Weak structure between objects and formed groups
$SC \leq 0.25$	No structure between objects and formed groups

Visual representation of the silhouette coefficient calculation, used to evaluate the quality of clusters in the K-Means clustering algorithm, is shown in Figure 1 Below is an explanation of the components in Figure 1:

1) x_i (red dot) is the data point being evaluated. This point is a member of cluster 1.
2) a_{x_i} is the average distance of point x_i to all other points in the same cluster, i.e., cluster 1. It is calculated by taking the distance between x_i and every other point in the same cluster, and then averaging it. This value measures how close x_i is to other members in its own cluster. The smaller the value of a_{x_i}, the better x_i is within its cluster.
3) b_{x_i} is the average distance of point x_i to all points in the nearest cluster that is not the cluster of x_i (e.g., Cluster 2 or Cluster 3). In this illustration, Cluster 2 has the smallest average distance to x_i, so it is used to calculate b_{x_i}. This value indicates how far x_i is from the nearest other cluster. The larger the value of b_{x_i}, the better x_i is separated from other clusters.

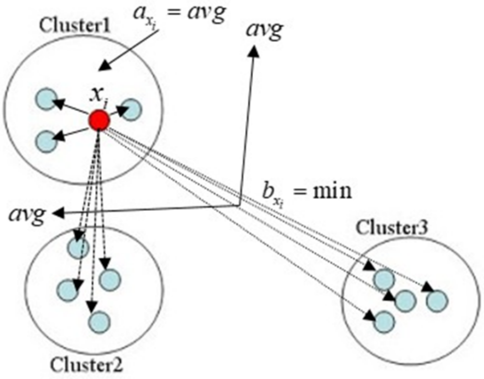

Fig. 1. Silhouette Coefficient Illustration [24]

The research workflow is illustrated in Figure 2. It comprises several stages: data collection, data preprocessing, clustering, evaluation, and visualization. The data, spanning from 2000 to 2023, was obtained from the Indonesia Ministry of Forestry [25]. The dataset contains information on provinces, emission sectors, and emission amounts measured in gigagrams.

During the preprocessing phase, missing values are imputed using forward fill and backward fill methods. Forward fill is a technique that replaces missing values with the most recently observed value, while backward fill fills missing values using the next observed value [16]. The K-Means algorithm is applied for clustering, and the results are evaluated using the Silhouette score. The clustering configuration with the highest Silhouette score is analyzed further. Finally, the clustering results are visualized through province-level mapping.

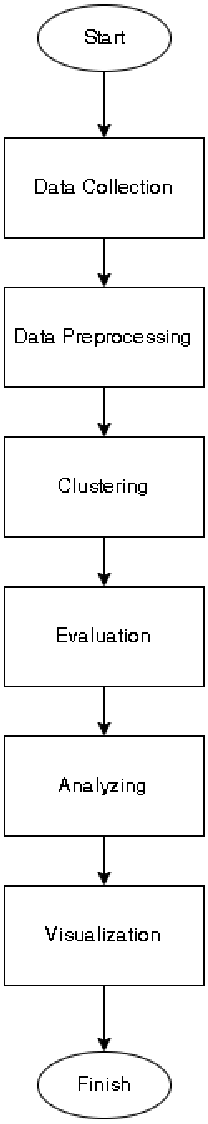

Fig. 2. Research Workflow

III. RESULTS AND DISCUSSIONS

The K-Means algorithms are run to cluster the data. The first experiments cluster 34 provinces. K-Means is run using different numbers of clusters $k = \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The experiments used different cluster numbers to find the best clustering results according to the evaluation metrics. The clustering results were subsequently evaluated using silhouette scores.

The data used in this design focuses on greenhouse gas (GHG) emissions data from 34 provinces in Indonesia, expressed in carbon dioxide equivalent (CO_2 eq) units and measured in Gigagrams (Gg). The time range analyzed covers the period from the year 2000 to 2023 for provincial-level GHG emissions. The dataset comprises comprehensive GHG emission data from 34 provinces in Indonesia, spanning 24 years (2000–2023). Each record includes sector-specific emissions (energy, industry, agriculture, forestry, and waste) in CO_2 equivalent units. The data is collected from the Ministry of Environment and Forestry's database, ensuring reliability and adherence to national reporting standards. The data consists of 3910 data points, comprising 34 provinces in Indonesia, 24 years of GHG emissions data from 2000 to 2023, and 5 GHG emission sectors. The data for the analysis consists of five sector test data and one combined test data from all sectors.

In this section, clustering analysis is performed on data divided into five sectors: waste, energy, industry, agriculture, and forestry, as well as a combined dataset from all sectors. The goal is to evaluate the clustering results with a variation in the number of clusters from two to ten clusters per test data, using the K-Means algorithm. The Silhouette Score is chosen as the evaluation criterion to assess the quality of the resulting clusters, indicating how well objects are grouped within the same cluster compared to other clusters. Table II, which evaluates the combined test data from all sectors, reveals that two clusters provide the highest average silhouette score of 0.7. Figure 3 shows provinces in Indonesia divided into two clusters, where Riau is the only province as a member in one cluster. Hence, it is more interesting to analyze the results of the three clusters.

TABLE II
EVALUATION OF CLUSTERING RESULTS FOR COMBINED SECTOR TEST DATA

Cluster	Average Silhouette Score
2	0.70
3	0.62
4	0.58
5	0.55
6	0.57
7	0.39
8	0.38
9	0.37
10	0.35

Emissions across Indonesia's provinces can be categorized into three distinct classifications. Figure 4 shows a mapping of provinces in three clusters. Riau province has consistently recorded the highest emission (”High” category) for 24 years, demonstrating substantial emissions exceeding 300,000 tonnes annually and reaching its peak at 452,908 tonnes in 2013, reflecting its intensive industrial and agricultural activities. In contrast, Papua and West Papua consistently maintained the lowest emissions (”Low” category), often recording negative values (ranging from -133,097 to around -30,000 tonnes), indicating their role as carbon sinks due to extensive forest coverage. The ”Medium” classification encompasses all other provinces across the archipelago, showing moderate emissions typically ranging between -10,000 to 200,000 tonnes, such as East Java, Central Kalimantan, and South Sumatra, representing areas with balanced industrial development and varying degrees of urbanization, as shown in Figures 5, 6, and 7. The decision to adopt a three-cluster configuration was informed by its ability to provide nuanced insights, despite the two-cluster solution having a slightly higher silhouette score. The three-cluster model enables differentiation between high-emission provinces like Riau and low-emission regions like Papua, capturing subtleties in medium-emission provinces that the two-cluster solution misses. This configuration facilitates more actionable regional strategies.

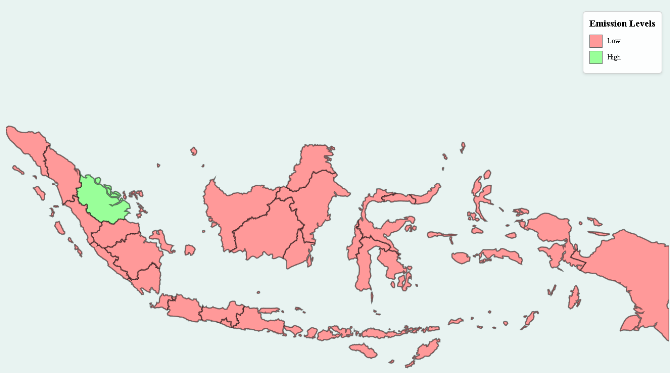

Fig. 3. Two Cluster Map of Combined Sectors in Indonesia

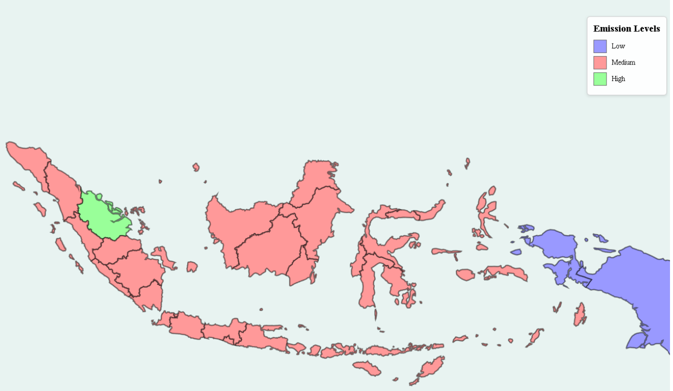

Fig. 4. Three Cluster Map of Combined Sectors in Indonesia

IV. Conclusion

The application successfully identified greenhouse gas (GHG) emission patterns at the provincial level in Indonesia from 2000 to 2023 using the K-Means Clustering algorithm and evaluation with the Silhouette Coefficient. The study successfully demonstrates the K-Means algorithm's capability to identify patterns in GHG emissions, achieving its objective of classifying provinces based on their emission profiles. This classification provides a foundation for prioritizing regions in future emission reduction initiatives. The analysis revealed that Riau Province was identified as the highest emitter, while

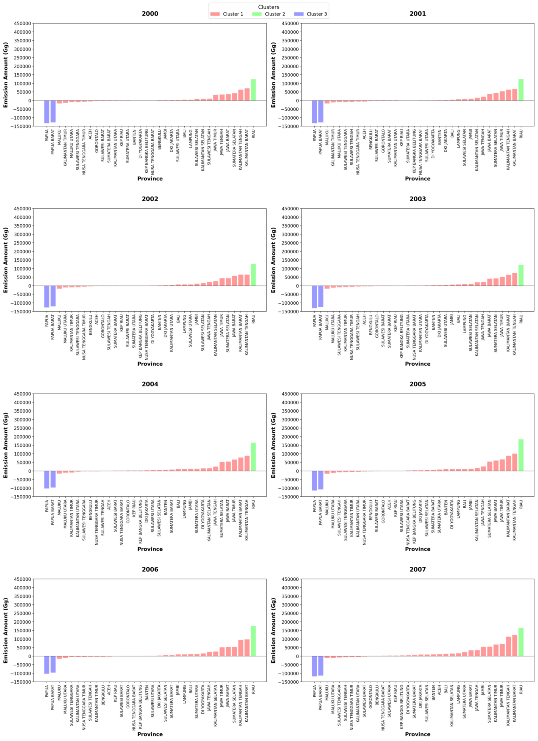

Fig. 5. GHG Emissions Combined Test Data of Provinces per Cluster for the Year 2000 - 2007

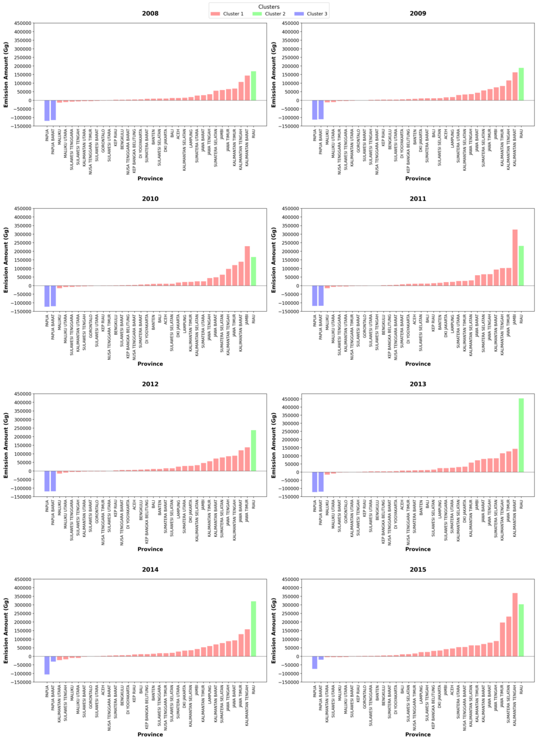

Fig. 6. GHG Emissions Combined Test Data of Provinces per Cluster for the Year 2008 - 2015

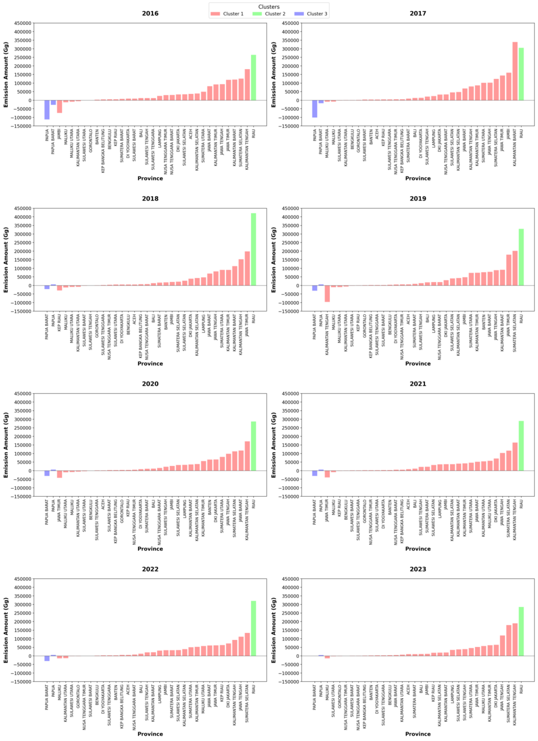

Fig. 7. GHG Emissions Combined Test Data of Provinces per Cluster for the Year 2016 - 2023

Papua and West Papua were recognized as the provinces acting as the highest absorbers. Evaluation using the Silhouette Coefficient showed that clustering with three clusters produces an average silhouette score of 0.62, indicating a medium level of similarity among provinces within a cluster. The interactive map not only illustrates the spatial distribution of emissions but also provides a deeper understanding of emission patterns and intensities across various regions of Indonesia.

References

[1] W. Wardoyo, “Perubahan Iklim Dan Perdagangan Karbon Dari Penurunan Emisi Gas Rumah Kaca (GRK),” JMB: Jurnal Manajemen dan Bisnis, vol. 5, no. 1, Oct. 2019, doi: 10.31000/JMB.V5I1.1993.

[2] “Climate Watch Historical GHG Emissions,” 2022, Washington, DC. Accessed: Aug. 30, 2024. [Online]. Available: https://www.climatewatchdata.org/ghg-emissions

[3] J. Self, “Calculating the carbon dioxide equivalent produced by vaporising a bottle of desflurane,” Anaesthesia, vol. 74, no. 11, pp. 1479–1479, Nov. 2019, doi: 10.1111/anae.14802.

[4] Y. A. Ishak, T. Handhayani, M. D. L. Sitorus, William, J. Pragantha, I. Lewenusa, ”Advanced Clustering Approach For Mapping Regions of Paddy Productivity In Indonesia Using Intelligent K-Means”, in 2024 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), Manado, 2024.

[5] Y. Febriani, Y. P. Sari, and D. Octaria, “Metode K-Means Cluster Untuk Mengelompokkan Kota/Kabupaten di Sumatera Selatan Berdasarkan Produksi Ikan Air Tawar,” Sainmatika: Jurnal Ilmiah Matematika dan Ilmu Pengetahuan Alam, vol. 18, no. 2, pp. 175–182, Dec. 2021, doi: 10.31851/SAINMATIKA.V18I2.6722.

[6] E. Lianita, A. Pratama, and A. F. Ulva, “Penerapan Metode K-Means Clustering Untuk Pemetaan Produktivitas Sayur-Sayuran Berbasis Sistem Informasi Geografis Di Provinsi Sumatera Utara,” JUSTIN (Jurnal Sistem dan Teknologi Informasi), vol. 12, no. 2, pp. 232–238, Apr. 2024, doi: 10.26418/JUSTIN.V12I2.72934.

[7] A. Al Masykur et al., “Penerapan Metode K-Means Clustering untuk Pemetaan Pengelompokan Lahan Produksi Tandan Buah Segar,” Jurnal Informatika, vol. 10, no. 1, pp. 92–100, Apr. 2023, doi: 10.31294/INF.V10I1.15621.

[8] T. Handhayani and I. Lewenusa, ”An Intelligent Clustering Approach For Analyzing A Multivariate Time Series Dataset, Case Study COVID-19 Outbreak in Indonesia,” 2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA), Lombok, Indonesia, 2023, pp. 1-6, doi: 10.1109/TSSA59948.2023.10367007.

[9] S. Suprihatin, Y. R. W. Utami, and D. Nugroho, “K-Means Clustering Untuk Pemetaan Daerah Rawan Demam Berdarah,” Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN), vol. 7, no. 1, Jul. 2019, doi: 10.30646/TIKOMSIN.V7I1.408.

[10] F. R. Tanjung, R. Efendi, and F. F. Coastera, “Pengelompokkan Dan Pemetaan Derajat Kesehatan Kota Bengkulu Dengan Metode K-Means Clustering,” 2019.

[11] I. Amal and R. A. Putri, “Clustering Pecandu Narkoba Menggunakan Algoritma K-Means Clustering,” Jurnal Sistem Komputer dan Informatika (JSON), vol. 5, no. 2, pp. 434–443, Dec. 2023, doi: 10.30865/JSON.V5I2.7009.

[12] R. A. Farissa, R. Mayasari, and Y. Umaidah, “Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung,” Journal of Applied Informatics and Computing, vol. 5, no. 2, pp. 109–116, Oct. 2021, doi: 10.30871/jaic.v5i1.3237.

[13] D. A. Wicaksono and Y. A. Susetyo, “Clustering Zonasi Daerah Rawan Bencana Alam Di Provinsi Sumatera Barat Menggunakan Algoritma K-Means Dan Library Geopandas,” Jurnal Indonesia: Manajemen Informatika dan Komunikasi, vol. 4, no. 2, pp. 426–438, May 2023, doi: 10.35870/JIMIK.V4I2.225.

[14] E. Jati Ramadhan, U. H. Muhammadiyah Kalimantan Timur Jl Ir Juanda No, K. SamarindaUlu, and K. Samarinda, “Pemetaan Penyebaran Anggota Muhammadiyah Berdasarkan Tingkat Kepadatan Menggunakan Metode K-Means Clustering,” Jurnal Rekayasa Teknologi Informasi (JURTI), vol. 6, no. 2, pp. 117–123, Nov. 2022, doi: 10.30872/JURTI.V6I2.8946.

[15] N. Kholila et al., “Pemetaan Kondisi Lingkungan Tanam menggunakan K-Means Clustering,” JSITIK: Jurnal Sistem Informasi dan Teknologi Informasi Komputer, vol. 1, no. 2, pp. 137–147, Mar. 2023, doi: 10.53624/JSITIK.V1I2.182.

[16] T. Handhayani and Z. Rusdi, “K-Means Using Dynamic Time Warping For Clustering Cities in Java Island According to Meteorological Conditions“, in 2023 Eighth International Conference on Informatics and Computing (ICIC), Manado, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICIC60109.2023.10381899.

[17] G.Andrian, D. Arisandi, and T. Handhayani, ”Clustering Data Meteorologi Wilayah Indonesia Timur Dengan Metode K-Means Dan Fuzzy C-Means, ” Jurnal Inti Nusa Mandiri, vol. 18, no. 2, pp. 100-106, 2024, doi:https://doi.org/10.33480/inti.v18i2.5039.

[18] N. D. Rahayu, A. H. Anshor, I. Afriantoro, and A. Halim Anshor, “Penerapan Data Mining untuk Pemetaan Siswa Berprestasi menggunakan Metode Clustering K-Means,” JUKI: Jurnal Komputer dan Informatika, vol. 6, no. 1, pp. 71–83, May 2024, doi: 10.53842/JUKI.V6I1.474.

[19] L. Cahaya, L. Hiryanto and T. Handhayani, ”Student graduation time prediction using intelligent K-Medoids Algorithm, ” in 3rd International Conference on Science in Information Technology (ICSITech), Bandung, 2017, pp. 263-266, doi: 10.1109/ICSITech.2017.8257122.

[20] E. Xiao and E. Xiao, “Comprehensive K-Means Clustering,” Journal of Computer and Communications, vol. 12, no. 3, pp. 146–159, Mar. 2024, doi: 10.4236/JCC.2024.123009.

[21] A. Saputra, B. Mulyawan, and T. J. Sutrisno, “Rekomendasi Lokasi Wisata Kuliner Di Jakarta Menggunakan Metode K-Means Clustering Dan Simple Additive Weighting,” 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208109865

[22] B. Škrlj, J. Kralj, and N. Lavrač, “Embedding-based Silhouette community detection,” Mach Learn, vol. 109, no. 11, pp. 2161–2193, Nov. 2020, doi: 10.1007/s10994-020-05882-8.

[23] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J Comput Appl Math, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[24] I. Lasri, A. RiadSolh, and M. El Belkacemi, “Toward an Effective Analysis of COVID-19 Moroccan Business Survey Data using Machine Learning Techniques,” in 2021 13th International Conference on Machine Learning and Computing, New York, NY, USA: ACM, Feb. 2021, pp. 50–58. doi: 10.1145/3457682.3457690.

[25] “Sign Smart - KLHK,” 2024. [Online]. Available:https://signsmart.menlhk.go.id/v2.1/app/chart/emisi_m.