

K-MEANS CLUSTERS WITH CALINSKI HARABASZ INDEX EVALUATION TO MAP FOREST DEGRADATION AND DEFORESTATION AREAS

UMMI RAHIMAH^{a*}, SHANTIKA MARTHA^a, NURFITRI IMRO'AH^a

*Mathematics Department, Universitas Tanjungpura, Pontianak, Indonesia
email : h1091211040@student.untan.ac.id, shantika.martha@math.untan.ac.id,
nurfitimroah@math.untan.ac.id*

Received December 24, 2024, Received in revised form November 1, 2025
Accepted January 21, 2026 Available online April 30, 2026

Abstract. *Although Indonesia is home to a rich biodiversity, the country is threatened by forest degradation and deforestation, particularly in West Kalimantan. As a significant contributor to the agricultural sector's gross domestic product (PDRB), the Sanggau Regency is vital for preserving the environment and promoting sustainable development. This research uses the K-Means Cluster to categorize regions in Sanggau that can potentially experience forest degradation. Then, the Calinski Harabasz Index will be used to determine which clusters are the most effective. Two thousand twenty-three, the research findings revealed five ideal clusters, each with a Calinski Harabasz Index value of 3.87. The first cluster consists of one sub-district, the second cluster consists of three sub-districts, the third cluster consists of two sub-districts, the fourth cluster consists of five sub-districts, and the fifth cluster consists of four sub-districts, which are all included in the distribution of clusters. A map illustrating the degree of urgency associated with forest degradation is produced as a result of this study. The map serves as a strategic reference for the government of Sanggau in its efforts to reduce the forest's degradation and develop areas per the peculiarities of each sub-districts.*

Keywords: Clustering, Ecology, Sustainability

1. Introduction

Indonesia has the most significant biodiversity in the world, with 63.1% of its land area being forested [1]. The forests of Kalimantan, know as the world's lungs, function as the primary oxygen producer and provider of abundant natural resources, such as commercial and non-commercial timber, as well as habitat for wildlife. However, this potential is threatened by high deforestation rates, especially in West Kalimantan, which in 2021 and 2022 recorded the highest rates of deforestation in Kalimantan [2].

*Corresponding author

The Sanggau Regency had a population of 533,937 people in 2023 and employment in the agricultural sector accounted for 64.64 % of the total population [3]. The farming industry in West Kalimantan is responsible for a substantial portion of the Gross Regional Domestic Product (PDRB) that is generated in Sanggau Regency, which is an agricultural area. Most of the population is employed in land-clearing operation that still rely on fire, leading to environmental degradation and deforestation. This is because the agricultural sector is highly dependent on the economy, and most of the population works there. The degradation of forests reduces land productivity and increases the chance of ecological caused by erosion, flooding, and climate change ([4], [5]). This situation contributes to the worsening of deforestation and forest degradation, creating a vicious cycle that continues to compound itself.

An investigation into the classification of regions according to the features of the factors that contribute to the deterioration of forests and the loss of forests is required as the first step toward resolving this issue. The K-Means Cluster algorithm in an efficient methodology for categorizing data according to shared criteria, due to its simplicity, ease of implementation, and relatively fast computation for the size of data used ([6], [7]). Despite this, review of the grouping results is required to guarantee that the group formation process is carried out most effectively. Calinski Harabasz Index (CHI) is a well-known method for accurately determining the appropriate number of groups [8]. This method was utilized in the evaluation that was carried out in this study.

The K-Means Cluster algorithm will be used to analyze areas in the Sanggau Regency that have the potential to experience forest degradation and deforestation. The Calinski Harabasz Index will be used to evaluate the grouping results, and the urgency level of the sub-districts in the Sanggau Regency will be mapped. These are the three primary objectives of this exploration. In light of this, the findings of this study will serve as a strategic reference for the government of the Sanggau regency in disaster mitigation and regional development planning, taking into account the peculiarities of each sub-district independently.

2. Theoretical Foundation

2.1. Cluster Analysis

Cluster analysis is a multivariate analysis method utilized for grouping study objects into clusters that exhibit high internal homogeneity and considerable external heterogeneity [9]. In a cluster, homogeneity refers to the degree to which members have similar features, whereas heterogeneity demonstrates to degree to which different clusters are distinct. There are two distinct approaches to cluster analysis: the hierarchical method involves grouping items into stages based on distance or similarity, beginning with small groups and progressing to large groups. The dendrogram is used to depict the results. The method that does not include hierarchical structures involves immediately dividing the data into groups based on the centroid through an iterative process until the best possible results are obtained.

Before carrying out cluster analysis, it is necessary to fulfill a number of presumptions to ensure that the method being utilized can function in the most effective

manner and generate outcomes that align with the data's characteristics. According to Ningrat, Di Asih, and Wuryandi [10], cluster analysis function based on two assumptions, which are as follows.

- (1) The sample is representative because the sample utilized can accurately represent the population as a whole. Because there is no specific value that can be used to determine whether or not the sample is adequate. One can find the KMO value somewhere between 0 and 1, with a value less than 0.5, the sample is deemed insufficient, suggesting that additional samples are required [11].
- (2) Multicollinearity occurs when no strong linear relationship exists between two or more variables being worked with. According to Nugraha, multicollinearity is indicated by a Variance Inflation Factor (VIF) value greater than 10 [12]. The following Eq. (2.1) can be used to calculate the value of the VIF [10].

$$VIF_n = \frac{1}{1 - R_n^2}, \quad (2.1)$$

where R_n^2 is the determination coefficient of the n -th variable.

2.2. Standardization of The Data

The results generated from cluster analysis could be biased due to differences in the measurement scales employed for the variables. As a result, it is essential to standardize the data to generate a Z-score, which is designed to maintain a measuring scale that is consistent across all variables. The following is an example of how Eq. (2.2) is typically used in data standardization calculations [13].

$$Z_{a;n} = \frac{x_{a;n} - \bar{x}_n}{\sigma_n}, \quad (2.2)$$

where $Z_{a;n}$ represents the standardization of data for the a -th object on the n -th variable, $x_{a;n}$ represents the value of the a -th object on the n -th variable, \bar{x}_n represents the mean of the n -th variable, and σ_n represents the standard deviation of the n -th variable.

2.3. Euclidean Distance

The objective of cluster analysis is to categorize objects according to their shared properties, minimizing the distance between clusters while enhancing the distance between clusters. Euclidean distance is a way of measuring distance based on the Pythagorean theorem [14]. This approach determines the geometric distance between two locations in multidimensional space and is one of the methods that may be used to compute distance. As the Euclidean distance decreases, the similarity between the items increases. As a result, things with a tiny distance between them tend to be clustered together. The following Eq. (2.3) is utilized to calculate the Euclidean distance [15].

$$d_{a;b} = \sqrt{\sum_{n=1}^N (x_{a;n} - x_{b;n})^2}, \quad (2.3)$$

where $d_{a;b}$ represents the Euclidean distance between the a -th object and the b -th item, and N denotes the number of variables.

2.4. *K-Means Cluster*

The K-Means Cluster algorithm is a non-hierarchical strategy for grouping numerical data [16]. This method involves splitting items into numerous homogenous clusters depending on their similarities. To produce optimal grouping, this algorithm seeks to reduce variation within clusters while simultaneously maximizing variation across clusters. The following is a list of the processes involved in conducting a K-Means Cluster analysis [17].

- (1) Take the value of K , which represents the number of clusters that will be generated.
- (2) Determine the first centroid by following a random number of K methods.
- (3) In order to determine the Euclidean distance between each object and its centroid, use Eq. (2.3) to perform the calculation.
- (4) Put each object into the clusters with the centroid somewhat close to it.
- (5) Using Eq. (2.4), determine the new centroid from the previous one.

$$u_{k;n} = \frac{\sum_{a=1}^{A_k} x_{a;k;n}}{A_k}, \quad (2.4)$$

where $u_{k;n}$ represents the k -th centroid of the n -th variable, $k = 1, 2, \dots, K$, A_k is the number of objects contained within the k -th clusters, and $x_{a;k;n}$ represents the value of the a -th item in the k -th clusters on the n -th variable, where $a = 1, 2, \dots, A_k$.

- (6) Carry out step three once more if the new centroid and the previous centroid are not identical.

2.5. *Calinski Harabasz Index*

In order to achieve successful clustering results using the K-Means Cluster technique, one of the most important steps is to determine the ideal number of clusters. When evaluating the quality of cluster produced by the K-Means Cluster algorithm, the Calinski Harabasz Index (CHI) is utilized as an internal criterion approach. Cluster cohesion and separation can be evaluated using the CHI function, which computes the Sum of Square Between-cluster (SSB) ratio to the Sum of Square Within-cluster (SSW) ([18], [19]). CHI is a common approach for determining the appropriate number of clusters since the more significant the CHI value, the better the clustering quality [20]. As Saputro outlined, the following are the processes involved in carrying out a K-Means Cluster analysis using the CHI assessment method [21].

- (1) Using Eq. (2.4) to get the centroid of the given area.
- (2) Using Eq. (2.5) to determine the grand mean of the observation.

$$\bar{x} = \frac{\sum_{a=1}^A \sum_{n=1}^N x_{a;n}}{A}, \quad (2.5)$$

where A represents the total number of objects, and \bar{x} represents the average of all objects.

- (3) Using Equation (2.6), the Sum of Square Between-clusters (SSB) is determine as:

$$SSB = \sum_{k=1}^K A_k \sum_{n=1}^N (u_{k;n} - \bar{x})^2, \quad (2.6)$$

where K is the total number of clusters, A_k is the number of objects in the k -th clusters, $u_{k;n}$ is the centroid of the k -th clusters for the n -th variable, and \bar{x} is the grand mean. Because the data used standardized data, the grand mean has a value 0.

- (4) Using Eq. (2.7), the Sum of Squares Within-Cluster (SSW) is given by:

$$SSW = \sum_{k=1}^K \sum_{a=1}^{A_k} \sum_{n=1}^N (x_{a;k;n} - u_{k;n})^2 \quad (2.7)$$

- (5) The Calinski Harabasz Index (CHI) can be calculated using Eq. (2.8):

$$CHI = \frac{SSB}{SSW} \times \frac{A - K}{K - 1}, \quad (2.8)$$

where SSB is the Sum of Squares Between-Clusters, SSW is the Sum of Squares Within-Clusters, A is the total number of objects, and K is the total number of clusters.

3. Result and Discussion

3.1. Descriptive Statistic

In 2023, this study uses ten variables that serve as indicators of the factors contributing to the degradation and deforestation of forest in the Sanggau Regency. The data that was utilized includes slope data X_1 that was obtained from the National Digital Elevation Model (DEMNAS), annual rainfall data X_2 that was obtained from the Climate Hazards Group Infrared Precipitation (CHIRPS), population density data X_3 , and land cover indicators such as rice fields X_4 , dry fields/gardens X_5 , fields/huma X_6 , plantations X_7 , community forests X_8 , state forests X_9 , and non-agriculture X_{10} that were obtained from the publication of the Central Statistics Agency of Sanggau Regency, specifically Sanggau Regency in Figures for the year 2024.

It is necessary to do descriptive statistical analysis to understand the data comprehensively, including characteristics such as the quantity of data, the minimum value, the highest value, and the average. Table 1 contains every piece of information regarding the data that was utilized and the descriptive statistical analysis performed for the indicators of the causes of forest degradation and deforestation in Sanggau Regency in the year 2023.

The number of objects used in this study is fifteen, and they are distributed among several sub-districts in the Sanggau Regency. This information can be found in Table 1. Toba District has the lowest slope percentage of 12.31%, while Entikong

Table 1. Variable Information Summary

Variable	Information (unit)	Objects	Minimum	Maximum	Mean
X_1	Slope (%)	15	12.31	26.67	19.36
X_2	Annual rainfall (m ³)	15	3,186.00	3,936.00	3,405.00
X_3	Population density (population/km ²)	15	15.00	78.00	42.33
X_4	Rice fields (ha)	15	364.00	3,293.00	1,320.30
X_5	Dry fields/gardens (ha)	15	125.00	21,338.00	4,215.00
X_6	Fields/huma (ha)	15	272.00	27,509.00	3,704.00
X_7	Plantations (ha)	15	1,951.00	65,016.00	24,009.67
X_8	Community forests (ha)	15	115.00	10,996.00	7,343.00
X_9	State forests (ha)	15	425.00	14,998.00	4,605.00
X_{10}	Non-agriculture (ha)	15	820.00	77,050.00	28,023.00

District has the highest slope percentage of 26.67%. The average slope X_1 of the region in Sanggau Regency in 2023 is 19.36% with Toba District having the lowest slope percentage and the lowest slope percentage. Because of their accessibility, low slopes, such as those found in the Toba District, are typically less susceptible to deforestation. This is because these slopes support sustainable land use. On the other hand, high slopes are more susceptible to deforestation. Land clearing in steep locations raises the danger of erosion and forest degradation [22].

3.2. The Assumptions Made in Cluster Analysis

In order to begin the process of cluster formation, an assumption test must first be conducted. The first assumption test that must be satisfied is that the sample being utilized can represent the population that can be tested using the KMO test. That said, there is no requirement for the KMO test because all objects being used are sub-districts or, to put it another way, the population of Sanggau Regency. After completing the KMO test, the process is then resumed with a multicollinearity test.

The multicollinearity test examines the linear relationship between the variables being used. If the value of the VIF is greater than 10, it indicates that the variables being utilized are multicollinear. By applying Equation (2.1), one may determine the value of the VIF, which is then displayed in Table 2 in the following manner.

Table 2 presents the correlation matrix among the independent variables along with their respective VIF values. The diagonal elements of the table represent the VIF for each variable, while the off-diagonal elements indicate the correlation coefficients between pairs of variables.

As can be seen Table 2, the VIF values for each variable used in this investigation are still greater than 10, with the exception of variable X_5 , which has a value of 12.32. This results in the elimination of variable X_5 , which will not be utilized for any future investigation. The values of the VIF is recalculated for the remaining variables after they have been removed from consideration. Following the recalculation of the values of the VIF, the results are presented in Table 3.

Table 3 shows that the VIF value of every variable utilized in this investigation is lower than 10. As a result, it is possible to conclude that each variable does not

Table 2. Correlation Matrix and VIF Values for Each Variable

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	2.52	0.27	0.42	-0.43	-1.81	1.37	-0.75	1.35	-1.12	2.04
X2	0.27	7.98	1.17	-2.06	-8.18	2.64	0.76	2.94	-1.08	3.47
X3	0.42	1.17	2.44	-2.16	-0.60	1.28	1.40	0.31	0.22	1.02
X4	-0.43	-2.06	-2.16	4.28	1.25	-1.32	-3.15	0.73	0.54	-1.13
X5	-1.81	-8.18	-0.60	1.25	12.32	-4.99	3.33	-4.75	2.23	-4.56
X6	1.37	2.64	1.28	-1.32	-4.99	4.62	-0.08	1.59	-1.84	1.72
X7	-0.75	0.76	1.40	-3.15	3.33	-0.08	5.65	-3.60	0.48	-1.51
X8	1.35	2.94	0.31	0.73	-4.75	1.59	-3.60	4.92	-1.26	3.42
X9	-1.12	-1.08	0.22	0.54	2.23	-1.84	0.48	-1.26	2.09	-1.55
X10	2.04	3.47	1.02	-1.13	-4.56	1.72	-1.51	3.42	-1.55	4.74

Table 3. VIF without Variable X₅

Variable	X ₁	X ₂	X ₃	X ₄	X ₆	X ₇	X ₈	X ₉	X ₁₀
VIF	2.25	2.54	2.41	4.16	2.60	4.75	3.09	1.69	3.05

exhibit multicollinearity, and the variables can be utilized for additional study.

3.3. The Standardization of Data

Because the variables being used in this study have a different units, it is necessary to perform data standardization. This will allow the data to be transformed into Z-score form, which will prepare the data for the cluster analysis procedure to proceed. Eq. (2.2) may be used to obtain the Z-score value.

3.4. Use of The Calinski Harabasz Index to Determine the Optimal Cluster

Because of the disparities in units discovered in the variables, it is necessary to standardize the data to ensure that the findings produced are not concentrated in particular variables. For this reason, the data utilized for subsequent analysis is normalized and does not exhibit multicollinearity conditions.

In the present investigation, the K-Means approach is used for clustering analysis, and the number of clusters (K) utilized varies, namely 2, 3, 4, and 5. This is a significant challenge because no predetermined guidelines determine the ideal number of clusters. Because of this, the evaluation was carried out using the Calinski Harabasz Index (CHI) approach, which is a method that measures the quality of clusters based on the degree of separation between clusters obtained by finding the SSB value using Eq. (2.6) and the degree of homogeneity within clusters obtained by finding the SSW value using Eq. (2.7). The CHI statistic compares the variation across clusters with the variance within clusters. Higher CHI values indicate superior data division quality, defined by unambiguous distinction between clusters and high uniformity of objects within each clusters. Table 4 presents the CHI values

that were determined by applying Eq. (2.8) to clusters 2 through 5 in the following manner.

Table 4. Calinski Harabasz Index Values

Cluster	Calinski Harabasz Index Values
2	3.18
3	3.40
4	3.62
5	3.87

According to Table 4, cluster 5 has the highest CHI value, 3.87. As a result, the best number of clusters for classifying possible regions of forest degradation and deforestation in Sanggau Regency is five. The outcomes of the grouping process with five clusters are reported.

Table 5. Grouping of Sub-Districts in Sanggau Regency into 5 Clusters

Cluster	Number of Members	Cluster Members
1	1	Jangkang
2	3	Parindu, Kapuas, and Balai
3	3	Sekayam and Toba
4	5	Beduai, Bonti, Kembayan, Entikong, and Noyan
5	4	Tayan Hilir, Mukok, Tayan Hulu, and Meliau

3.5. The Characteristics of Clusters

In the Sanggau Regency, there are five distinct clusters that can be defined as areas with the potential for forest degradation and deforestation. These clusters are as follows very high, high, medium, low, and very low. To determine the features of each cluster, the average of each variable in each cluster is analyzed. The results of this analysis are the presented in Table 6.

Cluster 1 is an area considered favorable for agricultural activities due to its low slope and high yearly rainfall. Fields or swiddens make up most of the land usage, which indicates intense agricultural practices that have been used for generations. The conversion of land for agricultural use, on the other hand, places a significant amount of pressure on the forest. Even if the population density is low, the risk of forest degradation and deforestation is increased due to less productive land management. As a result, sustainable environmental management is required.

Cluster 2 has the highest population density and a considerable dominance of rice fields, which together put a significant amount of pressure on forested land. Additionally, the absence of state and community forest makes this region extremely susceptible to forest degradation and deforestation, Steep slope conditions and moderate rainfall contribute to an increased risk of erosion during this period. There is

Table 6. Mean of clusters

Variable	Mean of Clusters				
	1	2	3	4	5
X_1	17.33	18.38	15.00	23.79	17.25
X_2	3,586.72	3,339.13	3,319.82	3,526.88	3,297.76
X_3	18.00	70.00	33.00	33.00	42.00
X_4	1,785.00	2,298.00	397.00	1,053.40	1,266.25
X_6	27,509.00	1,417.33	911.00	2,960.80	1,794.00
X_7	10,502.00	22,448.33	12,880.50	17,436.20	42,664.50
X_8	2,608.00	1,544.33	2,622.00	6,867.80	15,825.75
X_9	13,335.00	2,424.67	7,243.50	3,025.20	4,712.50
X_{10}	66,921.00	43,895.00	63,558.50	7,630.60	14,117.50

an urgent need for intervention in this region by implementing improved conservation and land management strategies.

Stable features are exhibited by Cluster 3, which has an extremely low environmental pressure. Rice fields, fields, and plantations are examples of agricultural activities that can be successful in areas with gentle slopes and moderate rainfall. This region still has a significant amount of woodland, which offers excellent ecological protection. Even though the risk of forest degradation is low when population pressure is moderate, routine monitoring is nevertheless required to prevent territory from being converted.

Cluster 4 is in a mountainous region that receives much precipitation, making it less ideal for large-scale agriculture. Even though human activity is relatively low, there are few forests in the area. This makes it susceptible to forest degradation, particularly due to the possibility of erosion or operations that involve clearing land. Preventative measures are required to preserve the ecological balance of this region.

Plantation activities based in perennial crops are the predominant agriculture in Cluster 5, characterized by moderate slopes and little rainfall. The population exerts modest pressure on the land, while other activities, such as rice fields, are relatively small. The ecological balance in this region is generally; nonetheless, it is essential to practice sustainable agricultural management to reduce the likelihood of forest degradation occurring in the future.

The results of grouping regions of possible forests degradation and deforestation in Sanggau Regency in 2023 based on their characteristics are presented as a map using QGIS software, as shown in Figure 1. This map was created using the process described above.

Because of its position as the administrative headquarters of Sanggau Regency, Kapuas District has a high potential for forest degradation and deforestation. This is especially true because the large population density contributes to the district's status. The vast majority of the land in this region is utilized for residential and com-

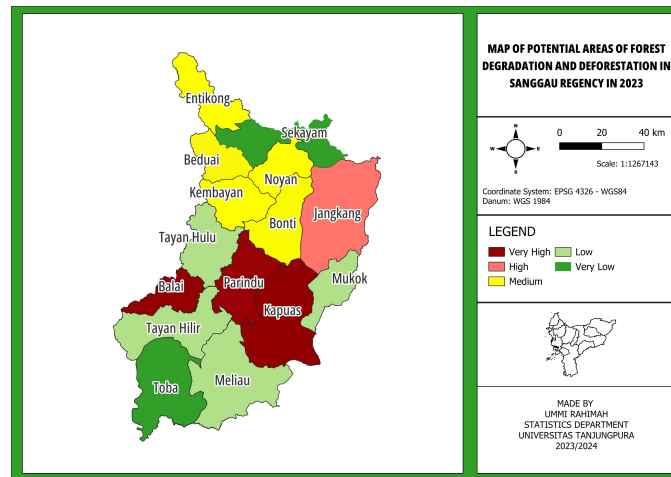


Figure. 1. Map of Potential Forest Degradation and Deforestation

mercial purposes and public infrastructure; consequently, agricultural and forestry activities are limited. Even though these activities are relatively low, it is necessary to foresee the pressure that will be placed on land due to high demand for space among inhabitants. Creating Green Open Space (RTH) is one of the mitigating steps that may be performed to improve environmental quality, supply water catchment areas, and minimize pollution, as well as the impact of urbanization on the environment.

The potential for forest degradation and deforestation in the Toba District is relatively low compared to other districts in Sanggau Regency. This is backed by the fact that there is a minimum amount of land clearing activities in the region, which is reflected in the low amount of land used for traditional agriculture, such as fields or humans. This situation demonstrates that there is very little strain placed on forests due to the seldom occurrence of large-scale land conversion for intensive agriculture or plantations. The Toba District is typically protected from substantial damage since it has a stable ecology. Nevertheless, to preserve this stability, it is necessary to continue monitoring the possibility of land conversion to guarantee the long-term viability of the environment.

4. Conclusion

The K-Means Cluster method with K variations was used to analyze the sub-district grouping in the Sanggau Regency in 2023. This study's results showed that different clusters were obtained, specifically 2, 3, 4, and 5 clusters, with distinct sub-district members. These are the different numbers of clusters that were acquired. Following the completion of the evaluation analysis using the Calinski Harabasz Index, it was determined that the best cluster consisted of five clusters, each of which had a CHI value of 3.87. Five levels of regency can be assigned to the possibility of forest degradation and deforestation in Sanggau Regency. These levels are very

high, high, medium, low, and very low. Parindu, Kapuas, and Balai are examples of sub-districts that have a very high potential; Beduai, Bonti, Kembayan, Entikong, and Noyan have a medium potential; Tayan Hilir, Mukok, Tayan Hulu, and Meliau have a low potential; and Sekayam and Toba have a very low potential.

Bibliography

- [1] Ministry of Environment and Forestry, 2023, *Deforestasi Indonesia Tahun 2021-2022*, Directorate of Forest Resource Inventory and Monitoring, Directorate General of Forestry Planning and Environmental Management, Jakarta.
- [2] Ministry of Environment and Forestry, 2023, *Neraca Sumber Daya Hutan Tahun 2023*, Directorate of Forest Resource Inventory and Monitoring, Directorate General of Forestry Planning and Environmental Management, Jakarta.
- [3] Badan Pusat Statistik Provinsi Kalimantan Barat, 2024, *Provinsi Kalimantan Barat dalam Angka 2024 [Kalimantan Barat Province in Figures 2024]*, Vol. 41, Badan Pusat Statistik Kalimantan Barat.
- [4] T. I. Fauzi and F. M. Hazz, 2020, Optimalisasi Data Satelit Penginderaan Jauh untuk Perhitungan Neraca Sumber Daya Hutan Kota Bukit Tinggi, *JURNAL SWARNABHUMI: Jurnal Geografi dan Pembelajaran Geografi*, vol. 5, no. 2, pp. 67 – 75.
- [5] N. Jainuddin, 2023, Dampak Deforestasi Terhadap Keanekaragaman Hayati dan Ekosistem, *HUMANITIS: Jurnal Homaniora, Sosial dan Bisnis*, vol. 1, no. 2, pp. 131 – 140.
- [6] Ma, Y., Liu, J., Zhang, H., 2022, *An efficient K-Means clustering approach for spatial environmental data analysis*, *Ecological Informatics*, 70, 101693.
- [7] F. Musharyadi and M. Kom, 2017, Tingkat Pemahaman Mahasiswa Terhadap Norma-Norma Agama Islam Menggunakan Algoritme K-Means Clustering, *Menara Ilmu: Jurnal Penelitian dan Kajian Ilmiah*, vol. 11, no. 78, pp. 48 – 54
- [8] S. P. Lima and M. D. Cruz, 2020, A Genetic Algorithm Using Calinski Harabasz Index for Automatic Clustering Problem, *Revista Brasileira de Computação Aplicada*, vol. 12, no. 3, pp. 97 – 106.
- [9] N. Syahfitri, E. Budianta, A. Nazir, and I. Afrianty, 2023, Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid, *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 3, pp. 1668 – 1675.
- [10] D. R. Ningrat, I. M. Di Asih, and T. Wuryandari, 2016, Analisis Cluster dengan Algoritma K-Means dan Fuzzy C-Means Clustering untuk Pengelompokan Data Obligasi Korporasi, *Jurnal Gaussian*, vol. 5, no. 4, pp. 641 – 650.
- [11] I. G. B. R. Utama, 2016, *Analisis Factor Explomatory*, ResearchGate. [Online]. Address: <https://doi.org/10.20527/Klik.V5i2.157> [access: 10 September 2024].
- [12] J. Nugraha, 2019, Penerapan Metode Regresi Ridge dalam Mengatasi Masalah Multikolinearitas Pada Kasus Index Pembangunan Manusia di Indonesia Tahun 2017, *Jurnal Prosiding Sendika*, vol. 5, no. 2, pp. 239 – 248.
- [13] N. A. A. Lestari, M. N. Hayati, and F. D. T. Amijaya, 2021, Analisis Cluster Pada Data Kategorik dan Numerik dengan Pendekatan Cluster Ensemble, *EKSPONENSIAL*, vol. 11, no. 2, pp. 117 – 126.
- [14] S. Saniah and E. V. Haryanto, 2023, Rancang Bangunan Aplikasi Pencarian

Lokasi Vaksin Pada Pukesmas di Kota Medan Menggunakan Metode Euclidean Distance Berbasis Android, *Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer*, vol. 3, no. 1, pp. 1 – 11.

- [15] R. A. Shafira, Y. Yahfizam, and A. M. Harahap, 2023, Menentukan Jarak Terpendek Dalam Pengiriman Barang dengan Perbandingan Euclidean Distance dan Manhattan Distance, *Journal of Science and Research*, vol. 6, no. 3, pp. 678 – 685.
- [16] R. N. Puspita, 2021, Analisis K-Means Clusters Pada Kabupaten/Kota di Provinsi Banten Berdasarkan Indikator Indeks Pembangunan Manusia, *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 2, no. 3, pp. 267 – 281.
- [17] H. Hidayatullah, S. Martha, and S. Aprizkiyandari, 2022, Analisis K-Means Menggunakan Metode Dunn Index dalam Menentukan Jumlah Cluster Optimal (Studi Kasus: Indikator Pendidikan SMA di Indonesia Tahun 2022), *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 13, no. 3, pp. 303 – 310.
- [18] S. K. Mann and S. Chawla, 2023, A proposed hybrid clustering algorithm using K-Means and BIRCH for cluster based cab recommender system (CBCRS), *International Journal of Information Technology*, vol. 15, no. 1, pp. 219 – 227.
- [19] M. N. Zhafar, K. Usman, and F. Akhyar, 2023, Penerapan Metode Clustering dengan Algoritma K-Means untuk Analisis Persebaran Varian Covid-19 (Studi Kasus Kelurahan Antapani Kidul), *eProceedings of Enegineering*, vol. 10, no. 5, pp. 4257 – 4264.
- [20] X. Wang and Y. Xu, 2019, An improved index for clustering validation based on Silhouette index and Calinski Harabasz index, in *IOP Conference Series: Materials Science and Engineering*, vol.569, no. 5, p. 052024, IOP Publishing.
- [21] D. R. S. Saputro, Algoritma Partitioning Around Medoid (PAM) dengan Calinski Harabasz Index untuk Clustering Data Outlier, *UNEJ e-Proceeding*, pp. 22 – 29.
- [22] R. F. Silamon, M. H. Idris, and S. Anggara, 2015, Analisis Degradasi Lahan dan Evaluasi Kesesuaian Lahan Hutan Tanaman Industri di Desa Marga Karya Resort Semamung, Kphp Batulanteh, *Jurnal Sains Teknologi & Lingkungan*, vol. 1, no. 2, pp. 9 – 15.