CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 1-15
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Cluster Analysis of Cities/Districts in West Kalimantan based on Stunting Response Indicators using the Calinski Harabasz Index

Tegar Rama Priyatna, Yundari, and Nur'ainul Miftahul Huda*

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura, Indonesia

Abstract

The stunting rate in West Kalimantan has reached 27% , mainly due to the government's inability to prioritise regions for essential services and education, especially for adolescents and pregnant women. This study aims to explain the role of modified K-Means and CHI methods in forming optimal clusters and interpreting their conditions. Eight research variables, sourced from BPS and SIGA in 2023, were used: number of adolescents receiving counselling, informed consents, complication cases, aslokon expenditure, aslokon stock, population growth rate, population density, and life expectancy. Clustering was done by analysing the data for each variable and the characteristics of the objects using the Euclidean distance, determining the centroid values, and iterating until the results stabilised. The clusters were evaluated from one to seven to find the optimal amount using CHI. The results identified five clusters: cluster 1 (relatively poor, three objects), cluster 2 (inferior, four objects), cluster 3 (good, three objects), cluster 4 (exquisite, three objects) and cluster 5 (good, one object).

Keywords: Centroid, Modified K-Mean Cluster, and Euclidean Distance.

1 Introduction

Sustainable Development Goals (SDGs) are a program for the sustainable development of human quality, with health and humanitarian issues addressed in the third leading goal, namely 'good health and well-being. The problem faced in implementing the third point of the SDGs is regarding the prevalence of stunting. Stunting is a condition related to an individual's height not in accordance with the expected growth conditions based on age and gender, indicating prolonged under nutrition when stunting occurs, indicating a long-term malnutrition status [1].

The prevalence of stunting in Indonesia remains relatively high, reaching 21.6% in 2022, which is 1.6% higher than the threshold set by the World Health Organisation (WHO). One of the provinces prioritized for accelerating stunting reduction is West Kalimantan. The stunting rate in West Kalimantan is higher than the national average, at 27%, making it crucial to address this issue. Apart from its potential to lower the quality of human resources, stunting also contributes to a loss of 2%–3% of the Gross Domestic Product (GDP) each year [2].

Previous studies have predominantly utilized single-distribution probability models or classical clustering methods such as k-means and hierarchical clustering. Conventional approaches, such as

*Corresponding author. E-mail: nurainul@fmipa.untan.ac.id

Submitted: June 19, 2025 Reviewed: October 10, 2025 Accepted: December 2, 2025
DOI: https://doi.org/10.18860/cauchy.v11i1.34510

K-means, suffer from limitations in the initialization of centroids, as they are chosen randomly by researchers, making the method susceptible to the determination of the initial cluster partitions. Furthermore, the k-means algorithm is vulnerable to the presence of outliers, which can result in a large sum of squared errors and relatively low accuracy [3].

To address these limitations, this study modifies the process of initial centroid selection. The formation of initial centroids in the modified k-means clustering emphasises inter-cluster heterogeneity and intra-cluster member similarity by using distance calculations for each data point. This approach enables more consistent clustering results and improves clustering accuracy [4].

The study aims to explain the role of the modified K-Means and CHI methods in forming the optimal clusters and interpreting the conditions of each cluster. In this study, the modification of the k-means method involves selecting the initial centroids based on the characteristics of the research objects. This modification has also been carried out [4]. The primary difference between this study and previous research is that it aims to enhance the quality of cluster formation using a modified K-means clustering method and applies the Euclidean method to determine the level of similarity between objects. Furthermore, the results of each cluster formation are analysed using the Chi-square method.

2 Methodology

The following section describes all the methods used in this study, from the process of analysing the data units to determining the optimal number of clusters. All methods employed are presented as follows: The methods section explains all the techniques used in this study, starting from the process of data unit analysis to determining the optimal number of clusters. All employed methods are presented as follows.

2.1 Standardization of Data

Data standardisation using the Z-score method was chosen over other common approaches, such as the Min-Max method, because, during dataset standardisation, Z-score measures data deviation based on the mean and evaluates data centrality using the standard deviation to determine the dispersion of the dataset. This process reduces the influence of outliers, which in turn leads to improved cluster quality [5]. The following is the formula for finding the value of data standardisation [5]:

$$Z_{i,j} = \frac{x_{i,j} - (\overline{x}_j)}{s_j}; \tag{1}$$

where $Z_{i,j}$ is the standardization of data for object to on-i variable to-j for i=$1, 2, \ldots, 14$ and $j = 1, 2, \ldots, 8$, $x_{i,j}$ represent object to on-i variable to-j, $\overline{x}_j$ represent the average value of variable to-j, s_j is standard deviation value of the variable to-j.

2.2 Distance Calculations

The calculation employs the Euclidean distance method to address the issue of random initialisation at the centre of the data, thereby enhancing accuracy in the cluster formation process [6]. Euclidean distance calculation is divided into two calculations: (1) calculating the distance between objects and (2) calculating the distance between centroids and objects. The following is the formula for finding the distance value using the Euclidean method [7].

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{n}(x_{i,j} - x_{k,j})^2} \tag{2}$$

$$d(c_l, x_i) = \sqrt{\sum_{j=1}^{n}(c_{l,j} - x_{i,j})^2} \tag{3}$$

where $d(x_i, x_k)$ state for distance between object to-i and object to-k, $d(c_l, x_i)$ is distance between object to-l and object to-i, n represent for total number of variables, meanwhile j represent for data from variable to-j for $j = 1$, x_i is object to-i meanwhile x_k is object to-k, and $c_{l,j}$ is centroid to-l on variable to-j.

2.3 Modified K-Means Cluster

The Modified K-Means Clustering method is a modification designed to address the shortcomings of the K-Means Clustering method. The initial cluster centre value (centroid) is randomly selected, resulting in inconsistent cluster results. Modifications are made to the initial cluster formation by adding a condition that the number of members of each cluster must meet $\frac{n(N)}{n(K)}$. This aims to improve the quality of cluster formation and time efficiency at the iteration stage [8]. Cluster formation is carried out in stages by first completing the formation at the minimum number of clusters (K_{min}), namely one after finishing the formation at the number of clusters of two and so on, then the $K = K + 1$ process can be carried out up to the maximum number of clusters(K_{max})), namely up to seven clusters. The stages with the Modified K-Means Clustering method are as follows:

1. **Determine the number of clusters to be formed (K).** Determining the number of clusters starts with $(K_{max}) = 1$, aiming to see the effect in the clustering process, then sets for $(K_{max}) = \frac{N}{2} = \frac{14}{2} = 7$. The value of $(K_{max}) = 7$ because it meets the requirements of nonhierarchical clusters, namely $K < N$. In addition, the value of $K > \frac{N}{2}$ or the value of K in the range of 8 to 13 will cause a bias that does not represent the actual conditions, namely overfitting, because the number of clusters is too much than necessary, causing the model to be too complex [9].
2. **Count the number of members in each cluster.** The number of members in each cluster is calculated by $\frac{N}{K}$. This process aims to ensure that each cluster has the same number of members, making the iteration process more efficient, and optimising each object's role in determining the initial centroid value.
3. **Places objects in the cluster.** Next, find the Euclidean distance value used to analyse the level of object similarity; the smaller the Euclidean value of the two objects, the more similar they are. Once obtained, the two objects enter the first two members in the first cluster. Compare the object distance value of the remaining 12 objects with the two objects at the beginning, then select the smallest object value until it meets $\frac{N}{K}$. Repeat the same steps to place objects in the second cluster until you reach the next cluster.
4. **Evaluation of initial cluster members.** The results in Step 3 are then evaluated by determining whether the $\frac{N}{K}$ requirement has been met and whether the threshold value is appropriate based on the closest distance. If it has not met the $\frac{N}{K}$ requirement or there is an error in placing the smallest value as a cluster member, then do Step 3 again.
5. **Obtaining the initial cluster.** The final result in Step 4 is the initial cluster, which is the cluster used to obtain the initial centroid value based on the members of each cluster. This value is then used as the initial capital in the iteration process in the next stage.
6. **Initial centroid formation.** Initial centroid formation is used for the iteration process in cluster formation. The following is the formula for finding the centroid value [4].

$$c_{l,j} = \left(\frac{1}{m_k}\right) \sum_{x \in C_l} z_{ij} \tag{4}$$

where m_k is the number of observations in the cluster to -k and $c_{l,j}$ is the j the centroid value to l variable, where t=$1, 2, \ldots, K$ and j=$1, 2, \ldots, 8$.

7. **Calculate the distance value of each object with each centroid.** Calculating the distance between the $c_{l,j}$ value and $x_{l,j}$ aims to place the object in the cluster. The calculation uses Equation (3).
8. **Place objects into clusters with the closest centroid distance.** The results in Step 7 have been obtained, specifically the value of the Euclidean distance between each object and the centroid, and then identifying the minimum value among all objects and the centroid. If the minimum value of the first object is located at the first centroid, then the first object is placed in the first cluster; if the minimum value of the second object is located at the fourth centroid, then the second object is placed in cluster four, and so on until the 14th object. The next step is to iterate, i.e. the result of placing objects in clusters are used to re-conduct Stage 6, then repeat Stage 6 to Stage 8 to iterate, the iteration stops when the centroid value at iteration $n + 1$ is the same as the centroid value at iteration n, and the object placed into the cluster at iteration $n + 1$ does not change position from iteration n, so the centroid and object placement are declared consistent.

2.4 Calinski Harabasz Index

The determination of the optimal number of clusters to facilitate grouping with the right amount of use, based on the level of similarity among cluster members, is achieved through the CHI method [10]. CHI is used to assess the cluster model that has been formed; this method refers to the ratio of the amount of dispersion (sum of squared distances) between clusters and within clusters for all clusters [11]. The stages with the Calinski-Harabasz Index method are as follows:

1. **Calculating SSW and SSB values.** The Sum of Squared Within Cluster (SSW) process is used to evaluate the level of similarity between each member within the same cluster. In contrast, the Sum of Squared Between cluster (SSB) process is carried out to evaluate the level of similarity between each cluster. The following is the formula for finding the SSW and SSB values:

   Equation (2) is used to calculate the Euclidean distance between two objects, whose results are used for initial cluster formation to determine the initial centroid value. In contrast, Equation (3) finds the distance value between the centroid and the object to-i the variable to-j the iteration process.

$$SSW = \sum_{i=1}^{K} \sum_{x_i \in C_l} d^2(c_l, Z_{i,j}) \tag{5}$$

   and

$$SSB = \sum_{i=1}^{k} N_k * d^2(c_l, \overline{x}_l) \tag{6}$$

   where $d^2(c_l, z_{i,j})$ is square of the distance between centroids to-l with object to-i, $d^2(c_l, \overline{x}_l)$ is square of the distance between centroids to-l with average of each variable to-i, N is number of objects researched, and N_k is The number of objects belonging to the cluster-k.
2. **Finding the CHI value.** Finding the CHI value aims to evaluate the number of clusters formed based on a review of the SSW and SSB, considering values that involve the number of objects (N) and the number of clusters formed (K). The following is the formula for finding the CHI value [12].

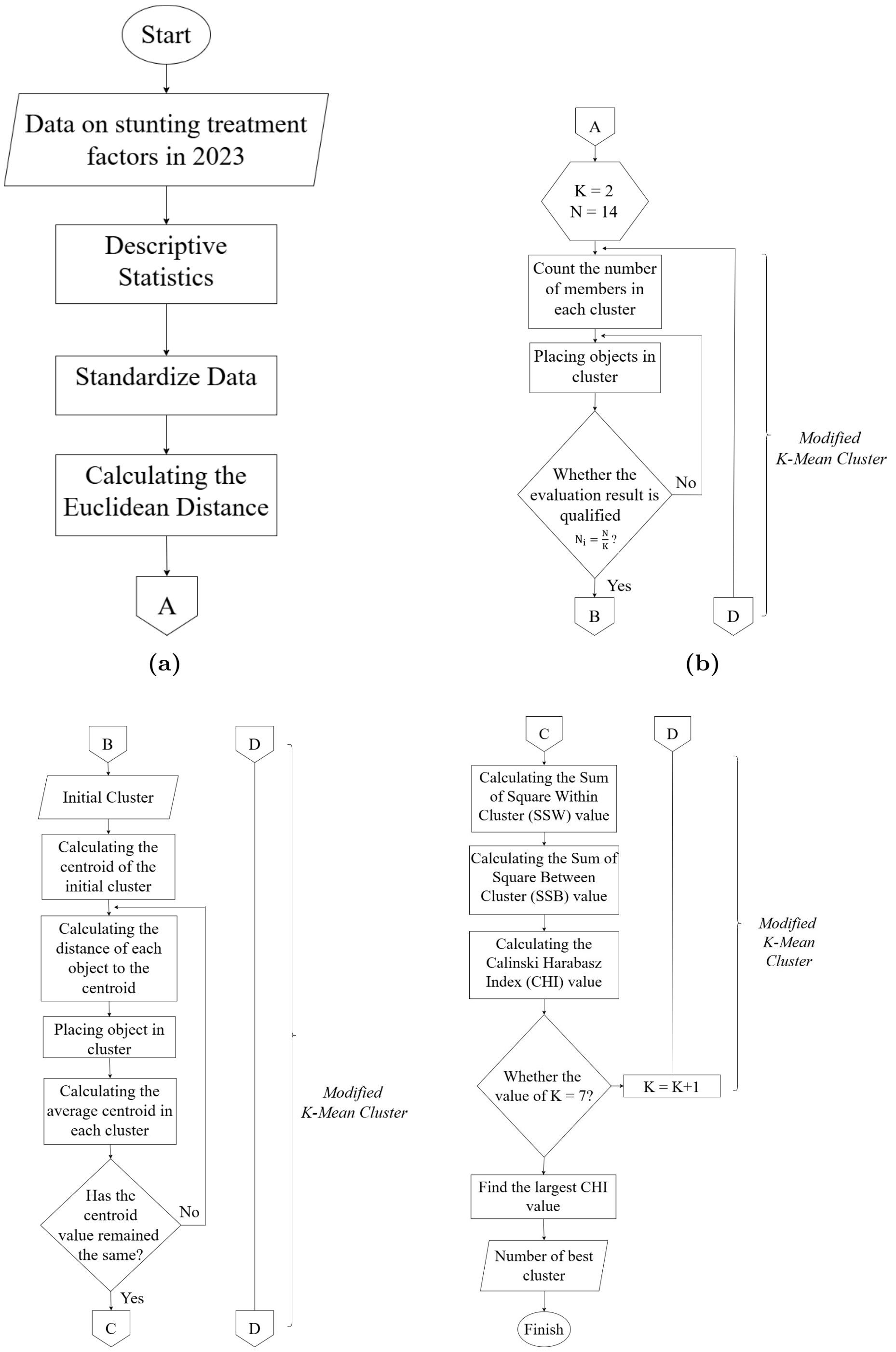
$$CHI = \frac{SSB}{SSW} \times \frac{N - K}{K - 1} \tag{7}$$

   The next step is to label the cluster obtained. Cluster labelling is used to facilitate understanding of the condition of each cluster that has formed, which in turn helps to identify priority areas. The following is the formula for finding values for cluster labelling [13].

$$C_k x_j = (\overline{x}_j) + c_{l,j} \times s_j \tag{8}$$

where $C_k X_j$ is the labeling of the cluster to-k and variable to-j.

More details of the research stages are presented in the research flow, as shown in Figure 1.


Start
Data on stunting treatment factors in 2023
Descriptive Statistics
Standardize Data
Calculating the Euclidean Distance
A
A
K = 2
N = 14
Count the number of members in each cluster
Placing objects in cluster
Whether the evaluation result is qualified $N_i = \frac{N}{K}$?
No
Yes
B
D
Modified K-Mean Cluster
B
D
Initial Cluster
Calculating the centroid of the initial cluster
Calculating the distance of each object to the centroid
Placing object in cluster
Calculating the average centroid in each cluster
Has the centroid value remained the same?
No
Yes
C
D
Modified K-Mean Cluster
C
D
Calculating the Sum of Square Within Cluster (SSW) value
Calculating the Sum of Square Between Cluster (SSB) value
Calculating the Calinski Harabasz Index (CHI) value
Whether the value of K = 7?
K = K+1
Modified K-Mean Cluster
Find the largest CHI value
Number of best cluster
Finish

(a) (b)

(c) (d)

Figure 1: Flowchart of cluster analysis using the Calinski Harabasz Index

In Figure 1, the evaluation results section highlights the necessity of satisfying the $\frac{N}{K}$ criterion during the cluster formation process. The purpose of using $\frac{N}{K}$ is to ensure that data in each initial cluster is distributed more evenly, thus allowing the centroid, calculated as the mean of cluster members, to accurately represent the overall data distribution, rather than arbitrarily chosen regions, as in the traditional k-means method. By employing the $\frac{N}{K}$ approach, the algorithm can avoid the scenario in which all initial centroids are placed very close to each other or at outlier points [14].

At this stage, an evaluation is also performed to re-examine the placement of objects within clusters to prevent misplacement. Such misplacement may occur when, due to differences in the computed distances between two objects, object A should correctly occupy a position in the cluster. However, that position is instead assigned to object B. This checking process ensures the accuracy and appropriateness of object assignments within their respective clusters.

3 Results and Discussion

This section presents the empirical findings of the study based on the stunting management indicators in cities and districts across West Kalimantan. First, the dataset and the eight indicator variables are described to provide an overview of the research objects and the information captured in each variable. The data are then summarised through descriptive statistics to highlight central tendencies, variation, and potential outliers that justify the use of standardisation. After that, the modified K-Means clustering procedure is applied, and alternative numbers of clusters are evaluated using the Calinski–Harabasz Index (CHI) to determine the optimal partition. Finally, the resulting clusters are labelled and interpreted to characterise regional stunting conditions and to identify priority areas for policy intervention.

3.1 Data

The data used in this study consists of stunting management indicators, comprising eight indicators: the number of informed consent forms given during services to new and repeat family planning participants (X_1), the number of severe complication and failure cases (X_2), the availability of contraceptive devices and drugs (X_3), the distribution of alokon (X_4), the number of adolescents attending counseling at youth and student information and counseling centers (X_5), life expectancy (X_6), population growth rate (X_7), and population density (X_8) for each city or regency in West Kalimantan Province in 2023, which were obtained from "*Sistem Informasi Keluarga*" (SIGA), "*Badan Pusat Statistik*" (BPS), and "*Laporan Semester I Penyelenggaraan Percepatan Penurunan Stunting Kalimantan Barat Tahun 2023*". The variables used represent several conditions, namely family planning services, demographics, social aspects, communication, and access, with the data presented in Table 1. The research data in Table 1 will be further analyzed to perform clustering and obtain the optimal number of clusters.

Table 1: Data on Stunting Handling Indicator Variables

Cities/Districts	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Sambas	15	6535	1	41687	64899	1.20	109	69.81
Mempawah	24	9273	2	19132	40234	1.32	162	71.79
Sanggau	437	16556	3	26534	42369	1.39	40	71.82
Ketapang	583	28922	15	39214	99373	1.35	20	71.51
Sintang	0	16974	8	29201	71327	1.40	20	72.46
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Singkawang	930	9741	4	29029	45803	1.67	447	72.87

3.2 Descriptive Statistic

Table 2 shows the abbreviations for the names of cities/districts. This was done to facilitate the research process, so the names of each district/city were abbreviated. The data in Table

Table 2: Abbreviation of Each Object

Districts/Cities	Abbreviation	Districts/Cities	Abbreviation
Sambas	SBS	Landak	LDK
Mempawah	MPW	Sekadau	SKD
Sanggau	SGU	Sekadau	SKD
Ketapang	KTP	Kayong Utara	KYU
Sintang	STG	Kubu Raya	KBR
Kapuas Hulu	KPH	Pontianak	PTK
Bengkayang	BKY	Singkawang	SKW

1 are then presented as descriptive statistics to facilitate understanding of the data for each variable. Descriptive statistics show the mean and standard deviation values that will be used for data standardisation and interpretation of the condition of each cluster at the best number of clusters. Descriptive statistics are shown in Table 3. The data in Table 3 represents good enough variables. The variables X_1, X_3, and X_6 recorded good values because the difference between the mean value and the standard deviation is still within the unit value, meaning the values are spread around the average. Other variables must be improved to address stunting conditions in the community. The standard deviation of each variable is still relatively high, meaning that the value is spread away from the average, so that the data has outliers; therefore, the data needs to be standardised. Based on Table 3, the data units are quite different, so they are more susceptible to the influence of outliers, so the units need to be equalised with the *Z-Score* method using Equation (1).

Table 3: Descriptive Statistics of Data

Variable	N	Minimum		Maximum		Average	Standard
		Value	Districts/Cities	Value	Districts/Cities		Deviation
X_1	14	0	STG, BKY, SKD, MLW	930	SKW	213.143	281.690
X_2	14	3623	MLW	28922	KTP	12229.785	7848.068
X_3	14	1	LDK, SBS	27	KBR	7.500	6.832
X_4	14	10684	KYU	47519	KBR	26271.929	10601.631
X_5	14	22680	MLW	196707	SKD	66778.286	44240.906
X_6	14	0.87	PTK	1770	KYU	1.436	0.231
X_7	14	8	KPH	5714	PTK	485.071	1454.304
X_8	14	69270	KYU	74.575	BKY	72.302	1.460

3.3 Modified K-Means Cluster

Data on the quality of stunting handling that has been standardized, then the calculation of Euclidean distance is carried out based on Equation (2) with $n = 8$ because there are eight indicator variables, it should be noted that for the results of the Euclidean distance, the characteristics between regions are said to be more similar if the distance value is smaller. The next stage for the value of N obtained in Table 3 is 14 and K is the number of clusters to be formed. The process of determining the value of K starts for $K_{min} = 1$, because to prove homogeneity and heterogeneity in cluster formation, then choose $K_{max} = \frac{N}{2} = \frac{14}{2} = 7$, because it meets the requirements of nonhierarchical clusters and avoids overfitting due to the number of clusters being too much than necessary [9].

Another important reason is that when the value of $K_{max} = \frac{N}{2}$, it significantly affects the calculation of the Calinski-Harabasz Index (CHI), resulting in bias during cluster evaluation. This situation influences the computation of both the Within-Cluster Sum of Squares (SSW) and the Between-Cluster Sum of Squares (SSB). Specifically, the SSW values become increasingly smaller because each cluster contains only a single member, meaning the distance calculation between two objects within a cluster cannot be performed. On the other hand, SSB values tend to rise because the initial centroids are spread across many clusters, most of which consist of just one member, thereby inflating the centroid distance used in SSB calculations. As a result, cluster quality assessment based on these metrics becomes distorted, making it unreliable for identifying optimal cluster solutions.

Initial cluster formation is based on the shortest distance obtained from the Euclidean distance calculation as a threshold value. Then, objects are placed in clusters by assigning each object to the cluster with the fewest number, up to N/K. The value of K, starting from $K = 1$ to $K = 7$, is processed until it is determined that the initial cluster has no change in the value of the objective function [15]. The results of the initial cluster formation are presented in Figure 2.

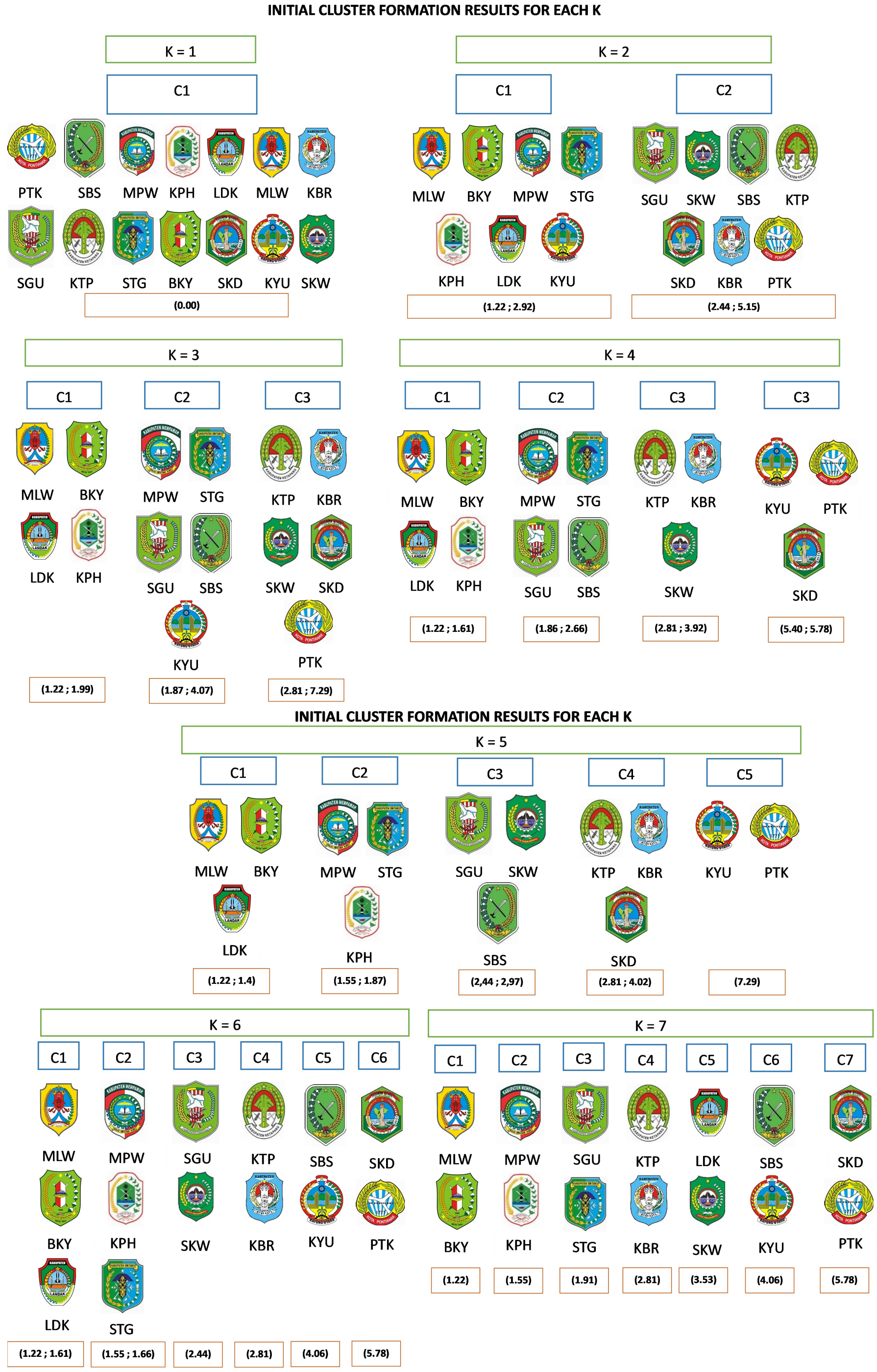


Figure 2: Initial Cluster Formation Results for Each K

Based on Figure 2, the number of members in each cluster has met N/K, then it has met the objective function. The value located at the bottom position of each cluster is a range of threshold values that indicate the similarity between members in each cluster. This stage distinguishes the Modified K-Means Cluster method from the K-Means Cluster method, as the placement of objects in the initial cluster is used to form the initial centroid, thereby streamlining the iteration process.

Based on Figure 1, the average value of each variable for each cluster is calculated based on Equation (4) using standardised values to obtain the initial centroid value [16]. The centroid values are presented in Table 4. The calculation uses Equation (3) to obtain the distance value of each object from the initial centroid. The calculation results are presented in Table (5). The minimum value of the initial centroid value is bolded in orange in Table 5 which aims to place objects in clusters, this result is temporary because the iteration process continues. The members in each cluster for each K are presented in Figure 2.

Table 4: Initial Centroid Value

		Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
$K = 1$	C1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$K = 2$	C1	-0.515	-0.610	-0.282	-0.780	-0.554	0.287	-0.299	0.207
	C2	0.515	0.610	0.282	0.780	0.554	-0.287	0.299	-0.207
					...				
$K = 7$	C1	-0.757	-1.011	-0.146	-0.980	-0.792	0.816	-0.307	1.148
	C2	-0.448	-0.474	-0.220	-0.630	-0.435	-0.395	-0.275	0.036
	C3	0.019	0.578	-0.293	0.151	-0.224	-0.179	-0.313	-0.113
	C4	1.215	1.592	1.976	1.612	0.922	0.470	-0.301	-0.615
	C5	1.095	-0.606	-0.732	-0.393	-0.509	0.448	-0.163	0.716
	C6	-0.393	-0.868	-0.659	-0.008	-0.513	0.210	-0.285	-1.894
	C7	-0.732	0.789	0.073	0.248	1.552	-1.369	1.644	0.721

To show how each region relates to the centroids obtained above, the Euclidean distances between all objects and all centroids are then computed. These distance values form the basis for the provisional placement of objects into clusters, as presented in Table 5.

Table 5: Value of Object-Centroid Distance. Orange values represent the minimum value

	$K = 2$		$K = 7$						
	C1	C2	C1	C2	C3	C4	C5	C6	C7
SBS	3.338	2.952	4.333	2.934	2.767	4.691	3.871	2.029	4.123
MPW	1.145	2.711	2.231	0.774	1.626	4.872	2.306	1.917	3.606
SGU	2.094	1.775	3.067	1.846	0.953	3.664	1.776	2.456	3.698
KTP	4.406	2.082	5.210	4.092	2.881	1.404	4.161	4.494	3.884
STG	1.848	1.618	2.658	1.587	0.953	3.443	2.669	2.728	2.799
KPH	1.002	2.301	1.704	0.774	1.632	4.155	1.922	2.683	3.370
BKY	1.474	3.465	0.609	1.903	2.713	5.171	2.257	3.653	4.091
LDK	1.221	3.323	1.328	1.416	2.371	5.435	1.762	3.149	3.943
SKD	3.840	3.038	4.316	3.637	3.526	4.000	4.342	4.475	2.891
MLW	1.253	3.774	0.609	1.951	2.835	5.487	2.247	3.084	4.702
KYU	2.751	4.208	3.403	3.049	3.498	5.493	3.448	2.029	5.727
KBR	5.257	3.480	5.736	5.196	4.371	1.404	5.168	5.262	5.233
PTK	5.549	4.581	6.016	5.162	4.897	6.264	5.672	6.186	2.891
SKW	3.349	3.013	3.722	3.467	3.008	4.093	1.762	3.864	4.997

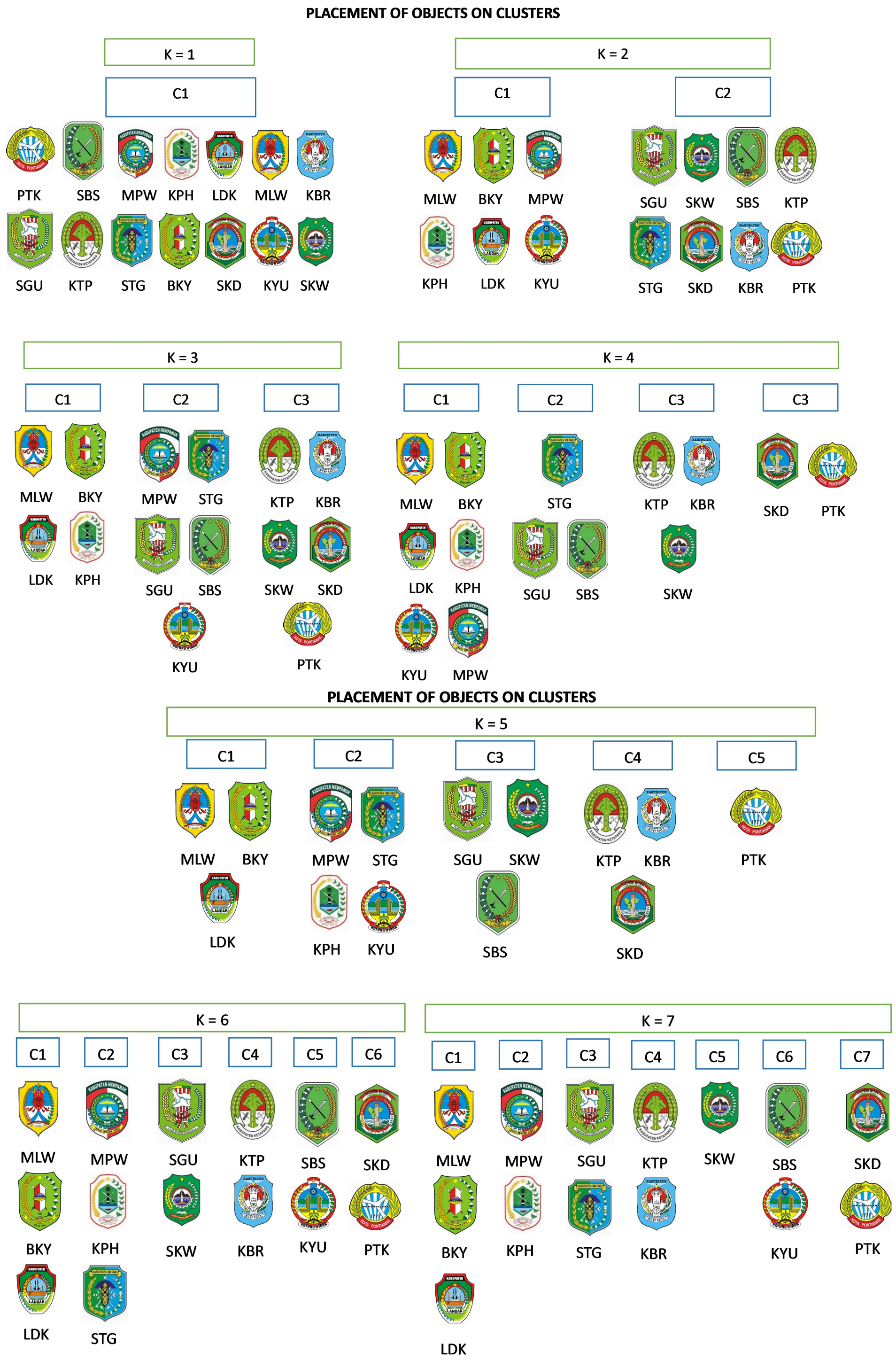


Figure 3: Placement of Objects on Clusters

The next stage involves iterating by calculating the average centroid to form a new one. The purpose of forming a new centroid is to see changes in the centroid value. If there is a change, the iteration process is carried out. If the average centroid value is the new centroid value, this value is the final centroid value with the object's location in the cluster being final or not needing change. Then, based on Figure 3, the average centroid value is calculated using Equation (4). The new centroid value obtained in the first process matched the previous centroid value, indicating that the initial centroid value is also the final centroid value. Consequently, the iteration process stops because it has reached consistency.

Table 8: Final Centroid Value and Status of Each Cluster. Red indicates nasty cluster, orange indicates bad cluster, yellow indicates fair cluster, green indicates good cluster, and blue indicates perfect cluster

Indicator Variable	C1	C2	C3	C4	C5
X_1	-0.623	-0.434	0.879	0.558	-0.707
X_2	-0.972	-0.339	-0.164	1.039	1.645
X_3	-0.415	-0.183	-0.707	1.586	-0.659
X_4	-1.002	-0.614	0.580	1.089	0.454
X_5	-0.710	-0.438	-0.356	1.594	0.167
X_6	0.506	0.124	-0.071	0.217	-2.450
X_7	-0.304	-0.295	-0.197	-0.305	3.595
X_8	1.114	-0.474	-0.553	-0.300	-0.112
Average	-0.301	-0.332	-0.074	0.685	0.395
Status	Bad	Very Bad	Enough	Very Good	Good

3.4 CHI Value

The final centroid value has been obtained. The next step is to find the SSW value by referring to Equation (5). The formation of clusters is better when the SSW value is smaller. Furthermore, calculating the SSB value using Equation (6), the optimal number of clusters is obtained when the SSB value increases. Table 6 presents the SSB and SSW values. Based on Table 6, the best number of clusters is $K = 5$. It means that there are five categories of regional groups based on the quality of stunting handling, which include excellent, good, sufficient, less, and significantly less handling groups.

Table 6: Determination of the Best Number of Clusters. Boalded values represent the highest value for each indicator.

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Value of SSW	84.442	73.714	53.298	39.311	36.671	**33.827**
Value of SSB	27.558	38.286	58.881	72.689	75.330	**78.173**
Value of CHI	3.916	2.857	3.682	**4.160**	3.287	2.696

3.5 Cluster Labeling

The labelling of each cluster is used to sort the cluster positions from best to worst. This aims to determine the cluster priority standards that must be focused on so that the 2024 stunting index target can be achieved. Based on Table 6, the best number of clusters is obtained when $K = 5$ with a value of 4,160. The members of each cluster at $K = 5$ are presented in Table 7. Each cluster will be labelled to identify the priority areas. This is important because it is used to analyse the characteristics of each cluster based on the problem of handling stunting by referring to indicator variables. Labelling begins by calculating the average value of the final centroid, which will be presented in Table 8, using Equation (4). Then, calculations are performed using Equation (8). Based on Table 8 for the status of the labelling results, the higher or positive value indicates a better cluster status. The next step is to perform calculations with Equation (8) to obtain the characteristics of each indicator variable in each cluster. The calculation results are presented in Table 9.

Table 7: Final results of each cluster member

C1	C2	C3	C4	C5
BKY	MPW	SBS	KTP	PTK
LDK	STG	SGU	SKD	
MLW	KPH	SKW	KBR	
	KYU			

Table 9: Cluster Labelling. Bolded values represent the values above average

Indicator Variable	C1	C2	C3	C4	C5
X_1	37.667	91.000	**460.667**	**370.333**	14.000
X_2	4601.000	9570.250	10944.000	**20386.330**	**25142.000**
X_3	4.667	6.250	2.667	**18.333**	3.000
X_4	15653.330	19765.750	**32416.670**	**37817.000**	**31083.000**
X_5	35388.670	47411.000	51023.670	**137280.700**	**74173.000**
X_6	**1.553**	**1.465**	1.420	**1.487**	0.870
X_7	42.333	55.500	198.667	44.000	**5714.000**
X_8	**73.928**	71.610	71.495	71.865	**73.925**

Table 9 is the cluster labeling result of two iterations. These results are compared with the initial average value data in Table 3. Several interpretations for black values are bolded, meaning above average, and not bolded, meaning below average. For bolded indicator variables, the meaning is values above average, which is less good. The results in Table 7 to Table 9 are presented visually through Figure 4 to explain the condition of stunting management in each region.

Based on Table 7 and Figure 4, it is further interpreted that cluster 1 is a cluster with poor stunting conditions. Some positive notes have been observed, notably the value of life expectancy, which is above average, indicating that public awareness about health is relatively high. This condition is supported by field evidence showing that one of the cluster 1 members, Melawi Regency, highlighted similar issues during a workshop on accelerating stunting reduction held on May 23, 2023. The discussions revealed several critical problems, including high school dropout rates driven by migration trends among residents seeking better economic opportunities. The situation is further exacerbated by the prevalence of early marriage within the community, the lack of parental awareness of the importance of participating in posyandu (community health posts) for child check-ups, and the low rate of exclusive breastfeeding during the first 1,000 days of life accompanied by inadequate nutritional intake [17].

Cluster 2 illustrates the complex problems that occur, indicating that it is a cluster with inferior quality in handling stunting, where AHH is under average. Public awareness of utilising health facilities is low [18]. One of the cluster 2 members represents the issues in the quality of stunting management is Kapuas Hulu Regency. According to the Head of the Health, Population Control, and Family Planning Office, Kapuas Hulu recorded a very high stunting rate based on nutritional status monitoring 30.3% in 2022 and 29.94% in 2023 indicating that the region remains classified under a high stunting prevalence category [19]. The primary barrier to reducing stunting in this area is the persistent practice of open defecation among residents due to inadequate sanitation facilities, as only 42 out of 278 villages currently have access to proper sanitation infrastructure [20].

Cluster 3 is a moderate condition cluster that has problems with health workers' low awareness of providing informed consent when serving family planning participants. The next issue is that although the alocon expenditure is above average, the alocon supply is below average. Hence, increasing public awareness of the importance of using alocon consistently is necessary. The last problem is that AHH is below average. One of the members of cluster 3 representing issues in the quality of stunting management is Sambas Regency. Stunting in this area is primarily caused by the low economic capacity of parents, which prevents them from providing adequate nutrition according to World Health Organization (WHO) standards. Additionally, low public awareness in utilizing available health facilities further contributes to this problem. These factors also explain the low AHH values observed in cluster 3 [21]. While clusters 4 and 5 are clusters with perfect conditions.

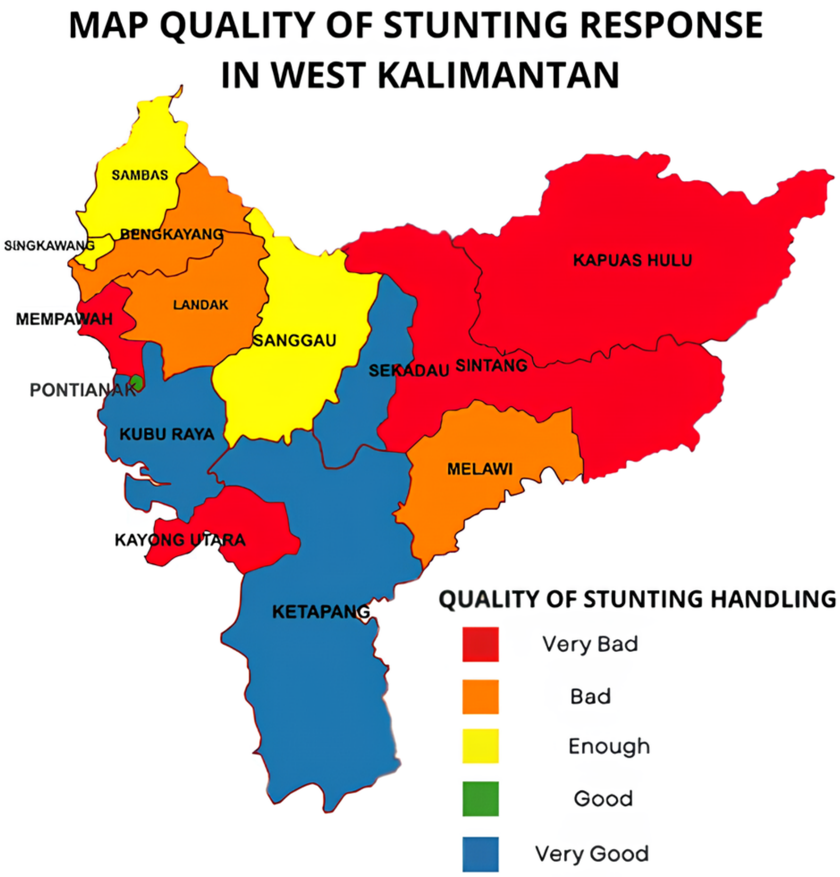


Figure 4: Mapping of the best cluster. Red area indicates the terrible areas, orange area indicates bad areas, yellow area indicates fair areas, green area indicates good areas, and blue area indicates excellent areas.

4 Conclusion

The modified method's role is to form clusters by analysing the data units of each variable, determining the level of characteristics for each object through Euclidean distance, and finding the value of the cluster centre (centroid). It then iterates until the results are consistent due to the formation of clusters. In contrast, the role of CHI is to evaluate several clusters from a minimum of one to a maximum of seven to obtain the best number of clusters. The best clusters from the modified K-Mean cluster process, ranging from K_{min} to K_{max}, are then reviewed using the SSW and SSB methods, and the CHI value is subsequently calculated. The best cluster formed is at $K = 5$ with a CHI value of 4.160, the most considerable CHI value among the other K values. Cluster 1 is a cluster with poor stunting handling quality, consisting of three cities or districts: Bengkayang, Landak, and Melawi. Cluster 2 is a cluster with poor stunting management quality, composed of four towns or districts: Mempawah, Sintang, Kapuas Hulu, and North Kayong. Cluster 3 is a cluster with a sufficient quality of stunting management, comprising three cities or districts: Sambas, Sanggau, and Singkawang. Cluster 4 has excellent stunting management, comprising three towns or districts: Ketapang, Sekadau, and Kubu Raya. Cluster 5 is a cluster with good stunting handling quality, which is only filled by Pontianak.

CRediT Authorship Contribution Statement

Tegar Rama Priyatna: Conceptualisation, Methodology, Formal Analysis, Writing-Original draft. **Nur'ainul Miftahul Huda:** Validation, Writing-review editing, Visualisation. **Yundari:**

Supervision, Funding Acquisition, Project administration, Writing-review editing

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (specifically ChatGPT) were utilised exclusively for language refinement and grammar editing during the preparation of this work. The analysis, interpretation, and core content were not generated by AI. All substantive results and discussions were entirely developed by the authors themselves.

Declaration of Competing Interest

The authors declare that no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data Availability

The data analyzed in this study is not publicly available and was obtained through the West Kalimantan Provincial Family Planning Agency (BKKBN).

References

[1] C. Aryu, *Buku epidemiologi stunting*. Fakultas Kedokteran Universitas Diponegoro, 2020. Available online.

[2] H. Wardoyo, "Laporan semester i penyelenggaraan percepatan penurunan stunting tahun 2023," Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN), Tech. Rep., 2023. Available online.

[3] P. A. Ariawan, N. P. Sastra, and I. M. Sudarma, "Clustering data remunerasi pns menggunakan metode k-means clustering dan local outlier factor," Indonesian, *Majalah Ilmiah Teknologi Elektro*, vol. 19, no. 1, pp. 33–39, Jan. 2020. DOI: 10.24843/MITE.2020.v19i01.P05.

[4] R. Cahyanto, A. R. Chrismanto, and D. D. Sebastian, "Pengelompokan komentar dataset sentipol dengan modified k-means clustering," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 6, no. 3, pp. 531–540, 2020. DOI: 10.28932/jutisi.v6i3.3006. Available online.

[5] I. M. K. Karo and H. Hendriyana, "Klasifikasi penderita diabetes menggunakan algoritma machine learning dan z-score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, 2022. DOI: 10.54914/jtt.v8i2.564. Available online.

[6] S. Setyaningtyas, B. I. Nugroho, and H. Arif Z, "Tinjauan pustaka sistematis: Penerapan data mining teknik clustering algoritma k-means," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 10, no. 2, pp. 52–61, 2022. DOI: 10.21063/jtif.2022.V10.2.52-61. Available online.

[7] R. T. D. Kurniawati, R. Rahmawati, and Y. Wilandari, "Pengelompokan kualitas udara ambien menurut kabupaten/kota di jawa tengah menggunakan analisis klaster," *Jurnal Gaussian*, vol. 4, no. 2, pp. 393–402, 2015. Available online.

[8] H. Aulawi, W. A. Kurniawan, and F. A. Rachman, “Analisis sentimen kepuasan driver terhadap kebijakan baru sistem order gojek,” *Jurnal Sains dan Teknologi ISTP*, vol. 14, no. 1, pp. 86–94, 2020. DOI: DOI:10.59637/jsti.v14i1.55.

[9] M. I. Hutagalung and S. Sriani, “Pengelompokan data penyakit tht menggunakan algoritma k-means clustering: Grouping of ent disease data using k-means clustering algorithm,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1568–1577, 2024. Available online.

[10] D. R. S. Saputro, “Algoritme partitioning around medoid (pam) dengan calinski-harabasz index untuk clustering data outlier,” *UNEJ e-Proceeding*, pp. 22–29, 2022. Available online.

[11] A. M. Sikana and A. M. Wijayanto, “Analisis perbandingan pengelompokan indeks pembangunan manusia indonesia tahun 2019 dengan metode partitioning dan hierarchical clustering,” *J. Ilmu Komput*, vol. 14, no. 2, pp. 66–78, 2021. Available online.

[12] X. Wang and Y. Xu, “An improved index for clustering validation based on silhouette index and calinski-harabasz index,” *IOP Conference Series: Materials Science and Engineering*, pp. 1–6, 2024. DOI: 10.1088/1757-899X/569/5/052024.

[13] D. A. Rahmah, “Analisis klaster berdasarkan indikator kesejahteraan rakyat menggunakan metode self organizing maps (som),” -, 2022. Available online.

[14] S. Sujatha and A. S. Sona, “New fast k-means clustering algorithm using modified centroid selection method,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 2, pp. 1–9, 2013. Available online.

[15] D. Arkham and D. Swanjaya, “K-means method for clustering public service assessment of goverment organization in kediri city,” *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi*, pp. 155–160, 2020. Available online.

[16] M. R. Nugroho, I. E. Hendrawan, and P. P. Purwantoro, “Penerapan algoritma k-means untuk klasterisasi data obat pada rumah sakit asri,” *Nuansa Informatika*, vol. 16, no. 1, pp. 125–133, 2022. Available online.

[17] Lan, *Berbagai penyebab stunting di melawi*, Indonesian, Dokumen tidak dipublikasikan, 2023. Available online.

[18] E. Ramadhani, N. Salwa, and M. S. Mazaya, “Identifikasi faktor-faktor yang memengaruhi angka harapan hidup di sumatera tahun 2018 menggunakan analisis regresi spasial pendekatan area,” *Journal of Data Analysis*, vol. 3, no. 2, pp. 62–75, 2020. Available online.

[19] S. Hakim. “Angka stunting masih tinggi di lima kecamatan di kapuas hulu.” Indonesian. Diakses dari forum online.

[20] D. K. Hulu, “Kasus stunting di kapuas hulu menurun,” Indonesian, *Artikel Surat Kabar*, 2022. Available online.

[21] D. Hardiyanti and Y. Yuniarti, “Analisis sosial ekonomi masyarakat yang memiliki bayi stunting di desa sebayan kabupaten sambas,” Indonesian, *Ekodestinasi*, vol. 2, no. 2, pp. 85–92, 2024. DOI: 10.59996/ekodestinasi.v2i2.133. Available online.