Journal of Informatics and Web Engineering
Vol. 4 No. 2 (June 2025) eISSN: 2821-370X

K-Means Clustering Optimization of Toddler Malnutrition Status Using Elbow Method

Femmi Widyawati[1*], Ahmad Yahya Dawod[2], Heru Agus Santoso[3]
[1]Department of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Imam Bonjol Street No. 207 (H Building) Semarang, Indonesia
[2]International College of Digital Innovation, Chiang Mai University, 239 Nimmanahaeminda Road, Suthep, Muang, Chiang Mai 50200, Thailand
[3]Faculty of Engineering, Dian Nuswantoro University, Nakula I Street No. 5 - 11 (I Building) Semarang, Indonesia
**corresponding author: (femmiwdywti@gmail.com; ORCiD: 0009-0003-5436-0669)*

Abstract - The problem of nutritional status is still a major challenge in the health sector in developing countries, including Indonesia. Malnutrition in toddlers can have serious long-term impacts on children's growth and development, including increased risk of disease, impaired cognitive function, and low productivity in the future. To overcome this problem, an in-depth analysis is needed to determine the distribution of nutritional status of toddlers in one of the provincial capitals in Indonesia, which can be used as a basis for planning more effective interventions. This study uses the K-Means algorithm to classify areas based on the prevalence of malnutrition in toddlers across all sub districts in the city. Determination of the optimal number of clusters was carried out using the Elbow method, which showed that the most appropriate clusters were two clusters. To assess the quality of the cluster, the Davies Bouldin Index (DBI) was used which produced a score of 0.361, while the Silhouette Score was 0.799, indicating that the cluster results were of high quality. The clustering results showed significant variations in the prevalence of malnutrition in various sub districts. Cluster 0 represents areas with low prevalence of malnutrition, comprising six sub districts, while Cluster 1 includes ten sub districts with high prevalence of malnutrition. By identifying these high-risk areas more clearly, health authorities and practitioners can develop more targeted and effective nutrition interventions. This research highlights the importance of data driven decision making in public health, supporting augmented intelligence in identifying and addressing nutritional problems in urban areas. The insights provided by this clustering approach contribute to more efficient and strategic health intervention planning.

Keywords—Malnutrition, Clustering, K-Means, Elbow Method, Davies Bouldin Index, Silhouette Score, Sub-district

Received: 07 November 2024; Accepted: 15 January 2025; Published: 16 June 2025

1. INTRODUCTION

Toddler nutritional status is one of the indicators used to describe the quality of health in toddlers [1]. Dietary problems are still one of the main challenges in the Health sector in Indonesia [2], one of which is the problem of malnutrition.

Malnutrition in toddlers can result in disruption of linear growth and intellectual development [3]. Malnutrition can be identified through nutritional status assessment. Assessment of toddler nutritional status is carried out using several indices, namely weight according to age (BB/A), length according to age (PB/A), and weight according to size (BB/PB) [4].

Nutritional status problems occur in various regions in Indonesia. Based on data from the 2022 Indonesian Nutritional Status Survey (SSGI), the prevalence of stunting as measured by the weight index for age (BB/A) in Central Java reached 20.8%, the prevalence of wasting (BB/A) was 7.9%, the prevalence of underweight (BB/A) was recorded at 15.8%, and the overweight figure was 3.2% [5]. These data show a high prevalence rate, which reflects serious problems related to the nutritional status of toddlers in Central Java. Similar conditions were also found in the capital city of Central Java, where the issue of nutritional status in toddlers is still a significant challenge that requires special attention from the government and the community.

To overcome these problems, a more in-depth analysis is needed to determine the distribution of malnutrition status. One practical approach is using a clustering model to group toddler nutritional status data based on location. In this study, the K-means clustering algorithm was applied to group toddler nutritional status data into several clusters based on similarities in dietary characteristics. The data used to build this clustering model was obtained from one of the health institutions. The K-means algorithm was chosen because it is the most popular and widely used algorithm in data mining, mainly because it has flexibility, efficiency, and is easy to interpret [6].

Previous research on toddler nutrition clustering was conducted by Mutammimul et al. [7]. This research was also conducted using an agglomerative algorithm with the clustering process carried out through data standardization, calculating Euclidean distances, and forming dendrograms to produce the final cluster. Meanwhile, research conducted by Syfriza [8] using the hierarchical clustering method to group provinces in Indonesia based on indicators of the nutritional status of toddlers (under two years old) in 2023 by comparing the Agglomerative Nesting (AGNES) and Divisive Analysis (DIANA) algorithms. The results showed that the AGNES algorithm produced the best cluster with two groups: cluster 1 includes 16 provinces with a high prevalence of underweight, stunting, and wasting, categorised as areas with poor nutritional status, while cluster 2 contains 22 provinces with a lower prevalence of dietary problems, classified as areas with good nutritional status.

The next study, Hadikurniawati et al. [9], discusses the severe challenge of malnutrition in children in Central Java, which has the highest prevalence of stunting in Java. This study uses clustering techniques to identify spatial patterns and distribution of stunted children in 35 districts/cities in Central Java. Two methods, namely K-Means and Density-based spatial clustering (DBSCAN), were applied to child nutritional status data to group areas based on stunting prevalence. The results showed that K-Means produced three clusters with low prevalence categories (11 districts/cities), medium (18 districts/cities), and high (6 districts/cities). At the same time, DBSCAN grouped 21 districts/towns into one central cluster and identified 14 districts/cities as outliers. K-Means' performance was proven to be superior to DBSCAN, indicated by a higher Silhouette score (0.403) and a lower Davies-Bouldin Index (0.785). This study provides important insights for prioritizing areas in planning health interventions to reduce stunting in Central Java.

Another study, Ozzi et al. [10], used the Hierarchical Agglomerative Clustering (AHC) algorithm with the Single Linkage method to group the nutritional status of toddlers in Bekasi City based on weight, height, and age data. The results showed five clusters: malnutrition, undernutrition, good nutrition, overnutrition, and obesity, with most toddlers in the excellent nutrition cluster. From the analysis of gender and posyandu area, a varied distribution of nutritional status was found, such as obesity being more dominant in boys. This study concluded that AHC effectively understands nutritional distribution despite limited data coverage.

This study will only focus on clustering the nutritional status of toddlers to determine the distribution of malnutrition in several sub-districts in the city, where there has never been any research on clustering the nutritional status of toddlers before. By using the K-Means Clustering algorithm and evaluating the optimal number of clusters using the Elbow and Silhouette Score methods using the Python programming language, it will help planning more targeted and effective nutritional interventions in overcoming nutritional problems in the city, as an effort for augmented intelligence implementation.

Augmented intelligence refers to a human-cantered approach to artificial intelligence (AI) that enhances human capabilities rather than replacing them. It combines advanced machine learning algorithms, data analytics, and automation to assist humans in decision-making, problem-solving, and productivity [11]. Unlike traditional AI, which aims to operate independently, augmented intelligence is designed to work collaboratively with humans, emphasizing transparency and interpretability. This concept prioritises leveraging AI to amplify human expertise and intuition, especially in complex or dynamic scenarios. By providing actionable insights, augmented intelligence enables more informed and accurate decisions across various domains. For example, identifying patterns and subgroups that help nutritionists design targeted interventions while retaining their expert judgment in interpreting and validating the results. It bridges the gap between human creativity and machine efficiency, fostering mutual empowerment for optimal outcomes.

2. LITERATURE REVIEW

The K-Means method has been widely used for clustering analysis in various contexts, including health, society, and especially the nutritional status of toddlers. Research conducted by Fadzly et al. [12] and Muhammad Dwi et al. [13] used the K-Means algorithm to group regions based on the prevalence of malnutrition in toddlers. Fadzly et al. grouped toddler data in West Java into three clusters (high, medium, low) using a dataset of 324 toddler data with nine attributes from the West Java Health Office in 2019–2022. Evaluation with the Silhouette Coefficient showed that K-Means was more optimal than K-Medoids, with scores of 0.617 and 0.491, respectively. Research by Dwi et al. used secondary data from BPS in 2016–2018 to group provinces in Indonesia into two clusters (high and low) with consistent validation results through RapidMiner 5.3. K-Means was chosen because it easily handles large and small-scale datasets accurately. The second study showed the distribution of regions with a prevalence of malnutrition as a reference for the government for nutritional improvement interventions.

Research by Karunia et al. [14] and Sri et al. [15] respectively combined clustering methods with determinant analysis to understand the prevalence of malnutrition in toddlers. Kurnia et al. used Canonical Correspondence Analysis (CCA) to reveal the relationship between the prevalence of malnutrition and determinant factors. Predictive Clustering combines the K-Means and Linear Discriminant Analysis (LDA) algorithms. This study successfully predicted the emergency status of malnutrition in three clusters (critical, grave, and warning) with an accuracy of 83.3%. On the other hand, Sri et al. used the K-Medoids algorithm to group provinces based on the prevalence of malnutrition with validation of the results using RapidMiner 5.3. The clustering results placed 13 provinces in the high cluster and 21 provinces in the low cluster. Both provide priority directions for nutritional interventions in certain areas.

The following research was conducted by Sari et al. [16], conducted regional mapping in Bangkalan Regency using the K-Modes algorithm based on secondary data from the 2021 BKKBN. This study focuses on age at marriage, family welfare, nutrition, type of residence, water sources, and sanitation. The analysis was carried out through factor analysis for variable smoothness, followed by cluster analysis with an optimal number of clusters $k = 4$. The study results show each cluster's specific characteristics, such as sanitation problems, understanding of ideal marriage, and sources of income. This study produces policy recommendations such as job training and community education to address the causes of stunting effectively.

Based on previous research provides significant insights into grouping areas based on the prevalence of malnutrition with various algorithms such as K-Means, K-Medoids, and K-Mode. However, this study provides added value in the method of determining clusters using elbow, evaluation using Davies-Bouldin Index (DBI) and Silhouette Score, and clustering the status of malnutrition in toddlers in the districts. These results provide more granular and relevant data for planning nutritional health policies at the local level.

The integration of augmented intelligence with clustering enhances the ability to uncover meaningful patterns in complex datasets while maintaining human oversight and expertise [17]. Clustering algorithms group data based on similarities, and augmented intelligence refines this process by providing interpretable insights and facilitating real-time collaboration with domain experts. This integration empowers users to validate and adjust clusters, ensuring the results align with practical applications and context-specific requirements. By combining machine efficiency with human intuition, this approach enables more accurate, transparent, and actionable clustering outcomes, particularly in dynamic fields like health and nutrition.

3. RESEARCH METHODOLOGY

The research method used is Knowledge Discovery in Database (KDD). This study uses the K-means clustering algorithm to cluster malnutrition status in toddlers. The research steps using the KDD method can be illustrated in Figure 1.

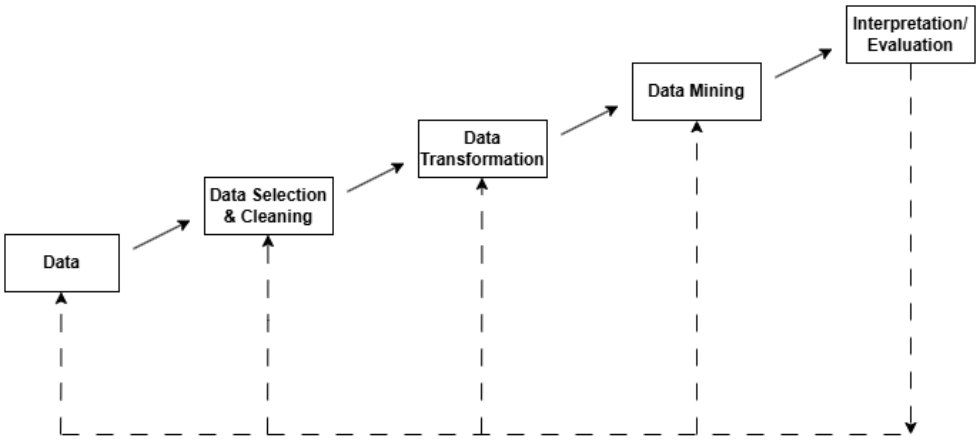


Figure 1. Knowledge Discovery in Database (KDD)

The research steps using the KDD method can be described as follows.

- Data Selection. This stage involves data collection. The data used is the 2023 toddler dataset of 9,594, obtained from one of the Health Department offices of the city.
- Preprocessing. After the data is collected, the next stage is data cleaning. Data cleaning involves identifying and fixing problems in the data, such as addressing missing values, removing duplicate data, and removing outliers. This stage aims to ensure that the data used in the analysis is high-quality and reliable. Next, normalise or standardise the data to ensure that the variables have the same scale to optimise the clustering results.
- Transformation. The next stage is data transformation. This involves converting data into a format suitable for analysis, such as converting categorical data to numeric data using the encoding method. In addition, dimension reduction is performed if there are many irrelevant attributes.
- Data mining involves performing clustering using the K-means algorithm, determining the optimum number of clusters, and using evaluation metrics such as the elbow method or silhouette score. Toddler nutrition data can be grouped based on similar characteristics, such as malnutrition, normal, or overnutrition.
- Evaluation. Evaluate the cluster results to ensure that the clusters formed have high homogeneity within the cluster and high heterogeneity between clusters. The silhouette score and DBI evaluation metrics are used.
- Interpretation. Interpret the cluster results to identify areas with a high prevalence of malnutrition. These results can be evaluated using a bar chart.

3.1 Data Selection

This stage is data collection. The data used is a toddler dataset from 2023 totaling 9,594 data and seven attributes, namely Toddler_Code, District, Age, Height, Weight, Head Circumference, and Upper Arm Circumference (UAC). This data will be selected again using attribute correlation to determine the most relevant attributes for clustering the nutritional status of toddlers based on sub-districts in the City.

3.2. Data Preprocessing

Data preprocessing is an essential stage in machine learning, which includes modifying or encoding data so computers can understand it [18]. Data used in the data mining process often needs to be in optimal condition for processing. Several problems can affect the results of the data mining process, such as missing values, data redundancy, outliers, or data formats incompatible with the system [19]. To overcome this, data pre-processing steps are needed. At the data

cleaning stage, it is carried out if there is empty data in the attributes in the dataset. The first way to overcome this problem is to replace empty data of a numeric type using the average value of a similar class. The second way is for categorical data, which is to fill in empty data using the backward and forward fill methods. To eliminate outliers using the z-scores method. Then, to overcome duplicate data, it includes identifying entries that have similarities in one or more columns, depending on the needs of the analysis. After duplicate entries are identified, verification is carried out to ensure that the entries are redundant and not valid but similar data. Furthermore, duplicate entries are removed to maintain the uniqueness of each observation in the dataset. The next step is feature selection. In clustering, feature selection is very important, especially because it helps balance clusters and can reveal intrinsically balanced data structures [20].

3.3. Data Transformation

Data transformation aims to adjust the dataset for modeling [21]. In the transformation stage in this study, feature standardization and feature encoding are carried out to ensure that the features used in modeling have a consistent scale and format that the machine learning algorithm can accept. For feature standardization, the RobustScaler method is used to standardise the features in the dataset by eliminating the influence of outliers. Unlike StandardScaler or MinMaxScaler, which are sensitive to extreme values, RobustScaler uses the median and interquartile range (IQR) as the basis for its calculations.

The RobustScaler equation is depicted in Equation (1).

$$X_{scaled} = \frac{X - median(X)}{IQR(X)} \quad (1)$$

where:
X = Original value of the feature
$median(X)$ = Median of the data on feature X
$IQR(X)$ = Interquartile Range (IQR) of data on the feature X

The next transformation stage is encoding categorical data. At this stage, the label encoding method is used to convert categorical features initially strings into numbers because the machine learning algorithm only accepts input in numeric form. Therefore, LabelEncoder is used to convert a data column containing text labels (categories) into corresponding numbers [22].

The equation for LabelEncoder is depicted in Equation (2).

$$y_i = f(x_i) \quad (2)$$

Where f is a mapping function that assigns an integer label to each unique category x_i, this mapping is based on the frequency or occurrence order of the categories in the data.

3.4. Data Mining

Data mining is extracting information and patterns hidden in large datasets to draw valuable conclusions for decision-making or prediction [23],[24]. One of the techniques commonly used in data mining is K-Means clustering, which aims to group data into several clusters based on the similarity of characteristics between elements in the dataset. This process begins by determining the optimal number of clusters using the Elbow method; then, the K-Means algorithm will iteratively adjust the position of the centroid of each cluster until convergence is achieved. Testing and implementing the K-Means clustering technique in this study were carried out using Python.

3.4.1 Elbow Method

The elbow method to determine the most optimal number of clusters by calculating the SSE (Sum of Squares Error) value for each cluster [25]. The equation (3) for elbow method is as follows:

$$SSE = \sum_{K-1}^{K} \sum_{xiesk} ||Xi - Ck||_2^2 \quad (3)$$

where:
Xi= The value attribute of i th data
Ck= Value of i th primary cluster attribute

3.4.2 K-Means Algorithm

The K-Means method is used to find clustering data. It starts by determining the number of clusters (C) and the initial centroids that are randomly selected. The centroid is the average of observations in one cluster [26]. The equation (4) for K-Means is as follows:

$$Ck = 1\, nk \sum dt \quad (4)$$

where:
Ck = Cluster
nk = Amount of data in cluster
dt = The number of each object included in each cluster

3.5. Evaluation

The final step in the data mining process involves assessing the results of the tests using the K-Means clustering method on the toddler dataset. This assessment aims to evaluate the extent to which the formed clusters reflect differences in nutritional status among toddlers in sub-districts. After the clustering process, the clusters are evaluated using the Silhouette Score evaluation metric to measure the quality of the resulting clusters and the DBI to determine how well the resulting clusters separate data and maintain uniformity within each cluster.

3.5.1 Silhouette Score

Silhouette Score is an evaluation metric that measures the extent to which each sample fits into its cluster and the extent to which the cluster is separated from other clusters. The higher the Silhouette Score, the better the clustering quality [27].

The Silhouette Score equation is shown in Equation (5).

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (5)$$

$s(i)$ = Silhouette value for data i, which has a value between -1 and 1
$a(i)$ = The average distance between data i and all other data in the same cluster (internal density)
$b(i)$ = The average distance between data i and all data in the nearest cluster (the most similar other cluster)

3.5.2 DBI

Testing using the DBI is one of the tests that can help evaluate the cluster's accuracy in this study. DBI testing is done by measuring the strength or accuracy of the Number of K (clusters) formed in the K-Means and K-Medoids algorithms [28],[29].

The DBI can be seen in the Equation (6).

$$DBI = \frac{1}{N}\sum_{i=1}^{N} Max\ j \neq i \left(\frac{S_i+S_j}{M_{i,j}}\right) \quad (6)$$

where :
N = Total number of clusters
S_i= The average distance between each data in cluster i and the centroid of cluster i
$M_{i,j}$ = The distance between the centroid of cluster i and the centroid j

4. RESULTS AND DISCUSSIONS

In this study, K-Means clustering was used. This was carried out using the KDD method, as well as evaluation using the silhouette score and DBI evaluation metrics. This study was conducted using Python programming.

4.1 Data Selection

The initial stage in the KDD process begins with the data selection. The data used in this study were obtained from the health office. The dataset consists of 9,594 entries with seven attributes, namely Toddler_Code, Subdistrict, Age, Height, Weight, Head Circumference, and UAC, as shown in Table 1. This data selection stage is essential in the study because the selected dataset will be used as the primary material in the data processing and grouping process using the clustering.

Table 1. Dataset

Toddler_Code	Subdistrict	Age	Height	Weight	UAC	Head_Circumference
9931202312	0	31,1	10,1	84,4	14,6	47,5
8751202322	9	20,1	8,25	75,1	14,2	45,5
8487282345	5	23,2	9,9	80,3	14	45
6971202312	8	9,4	5,5	63	11,5	40
9832403276	7	22,1	9,9	75,8	15	47,1
8062402212	2	28,5	10,2	83	9,9	47,1
...	...	...	...	...	...	...
...	...	...	...	...	...	...
3509202391	13	26,8	9,1	81,4	13,3	46,4

4.2 Data Preprocessing

After carrying out the data selection stage, the next stage is the data preprocessing stage, which aims to prepare the dataset so that it is ready to be used in further analysis processes. This stage involves various essential processes to ensure data quality, such as handling missing values to avoid interfering with the data analysis process. In addition, duplicate data is removed to prevent bias affecting clustering results. As well as outlier handling, namely identifying and managing data with extreme values that can damage the pattern in the dataset.

The next stage is feature selection, which aims to select the most relevant parameters in grouping the nutritional status of toddlers. Feature selection is done by considering the correlation between features, where features with high correlation are prioritised for further analysis. Based on the correlation analysis visualised in the Figure 2, it can be seen that the Usia_Ukur (Age), Tinggi_Badan (Height), and Berat_Badan (Weight) features have high correlations, so they are selected as the main features for the grouping process.

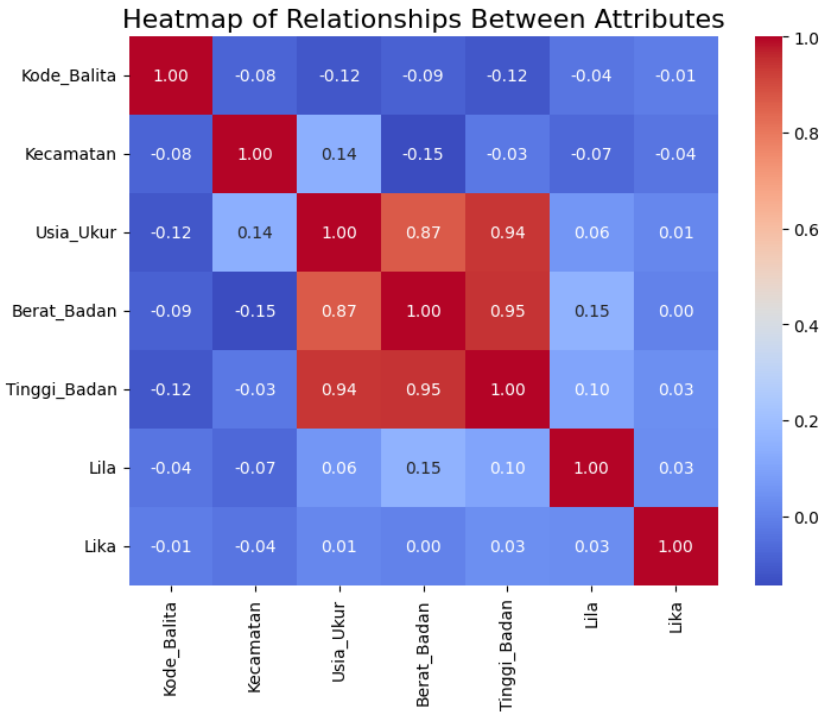


Figure 2. Heatmap of Relationship Between Attributes

In addition, the Kecamatan (Subdistrict) features are maintained despite having a low correlation because both of these features have an essential role in the context of the analysis. Kecamatan (Subdistrict) is essential for knowing the distribution of nutritional status in various regions, which can provide insight into geographic or demographic factors that affect the nutritional status of toddlers. This combination of features is expected to provide more accurate and structured grouping results to identify nutritional status in toddlers. The following is the data from the feature selection results, which can be seen in Table 2.

Table 2. Dataset After Feature Selection

Subdistrict	Age	Weight	Height
0	31,1	10,1	84,4
9	20,1	8,25	75,1
5	23,2	9,9	80,3
8	9,4	5,5	63
7	22,1	9,9	75,8
2	28,5	10,2	83
...	...	...	...
...	...	...	...
13	26,8	9,1	81,4

4.3 Data Transformation

After performing the data preprocessing stage, the next stage is to perform data transformation, at this stage, the researcher transforms the data into the encoder. The collected dataset contains essential attributes that are transformed and encoded to ensure compatibility with the classification algorithm. The encoding process uses LabelEncoder, which

can convert categorical attributes into numeric format without changing the data structure. This transformation makes it easier for the algorithm to read and process categorical data.

4.4 Data Mining

This research process uses the K-Means Clustering algorithm as the primary technique for grouping data based on specific characteristics. This study applies the Elbow Method, a graphical approach that identifies the "elbow point" on the inertia graph against the number of clusters to determine the optimal number of clusters. This point indicates the number of clusters where the decrease in inertia begins to slow down, resulting in a balance between efficiency and clustering accuracy. With this method, the optimal number of clusters can be selected systematically, ensuring that the clustering results can represent the data distribution pattern well. This approach provides a strong foundation in clustering analysis to understand the distribution of malnutrition status in the city. Figure 3 shows the elbow graph formed.

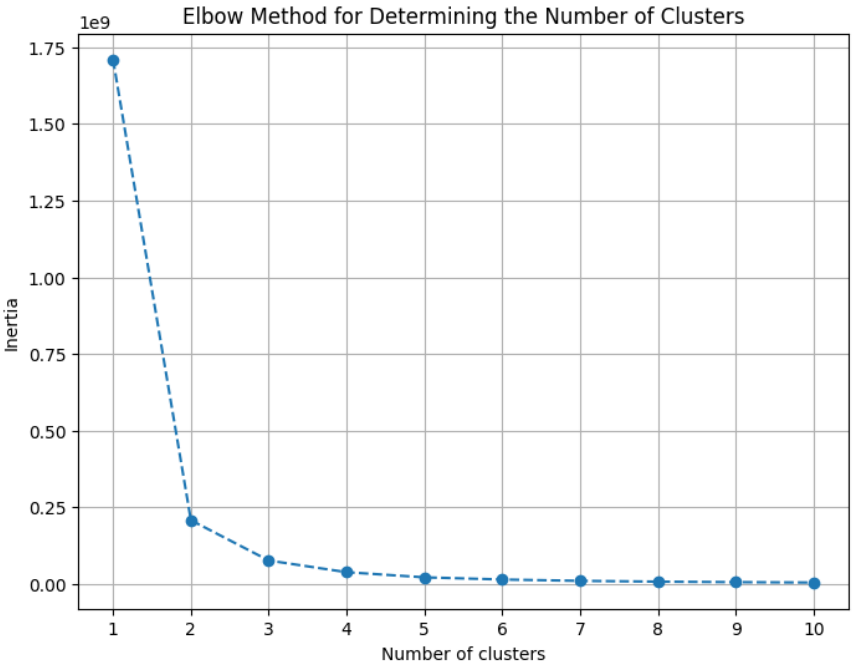


Figure 3. Elbow Method Graph

In the elbow graph, it can be seen that the elbow point is at the number of clusters 2, which indicates a significant decrease in inertia. However, to further ensure the optimization of the number of clusters, the researcher evaluated the number of clusters from 2 to 5. In this range, the decrease in inertia still occurs, indicating a better clustering. Furthermore, the DBI and Silhouette Score values were calculated for each cluster in the range, and the cluster with the lowest DBI value and the highest Silhouette Score was selected as the most optimal to describe the distribution of nutritional status in the area. The results of the cluster number test can be seen in Table 3.

Table 3. Cluster Test Result

C (Cluster)	DBI	Silhouette Score
2	0,361	0,799
3	0,499	0,743
4	0,568	0,569
5	0,583	0,573

After analyzing the DBI results for the number of clusters between 2 and 5, it was found that the lowest DBI value was in cluster 2. A low DBI value indicates that cluster 2 has better grouping between data distributions, where the distance between clusters is greater, and the similarity between data in one cluster is higher. A higher Silhouette value indicates that the clusters are more separated. This shows that cluster 2 is the optimal choice to describe the grouping of toddler data based on nutritional status because the results show an ideal balance between cluster accuracy and diversity. Thus, although the elbow point on the elbow graph occurs in cluster 2, further analysis using DBI ensures that cluster 2 is the most representative and provides the most stable and valid results in this study, especially in identifying the distribution of malnutrition in the studied area.

After determining the optimal number of clusters, which are 2 clusters labelled Cluster 0 and Cluster 1, the next step is to calculate the distance between the centroids of each cluster. Figure 4 shows a heatmap depicting the Euclidean distance between the centroids of the two clusters generated through K-means analysis. The horizontal and vertical axes represent the cluster labels (Cluster 0 and Cluster 1), while the values in the matrix indicate the magnitude of the distance between the centroids. The colours in the heatmap reflect the gradation of the distance, where blue indicates a smaller distance, while red indicates a more considerable distance. The analysis results show that the distance between Cluster 0 and 1 is 10.280. This information indicates the degree of separation between the clusters, where the distance between Cluster 0 and Cluster 1 reflects clusters that are more significantly separated in multidimensional space. The clustering results can be seen in Figure 5.

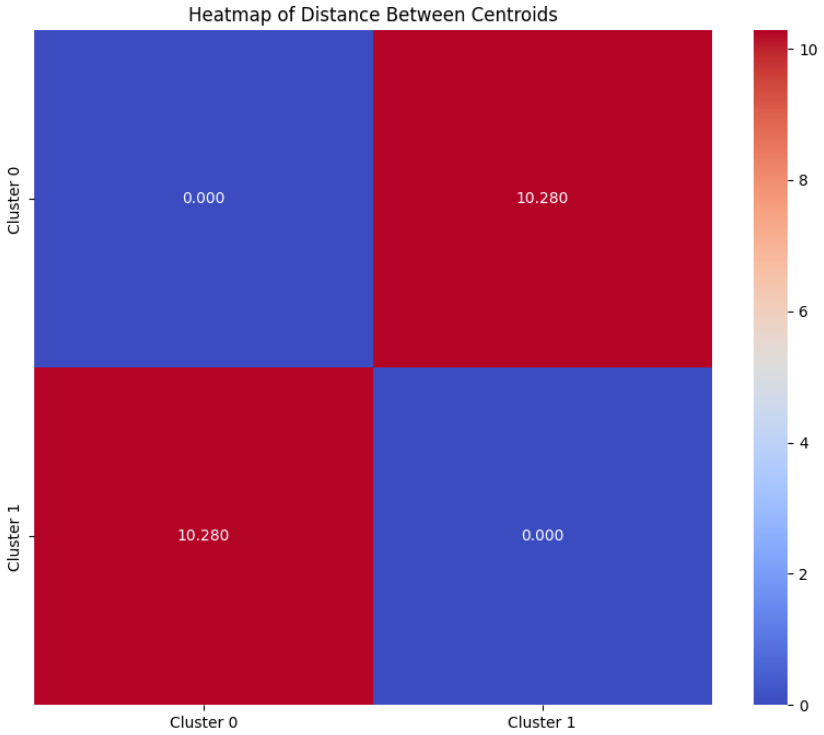


Figure 4. Heatmap of Distance Between Centroids

Figure 5 presents the visualization results of nutritional status data clustering using a clustering algorithm, which has been reduced in dimension into two main components with Principal Component Analysis (PCA). The first (Principal Component 1) and second (Principal Component 2) principal components are linear combinations of the original features, namely age, sub-district, height, and weight, representing the most significant variance in the data. The points on the graph represent individual data that are grouped into clusters based on the results of the clustering algorithm, with different colouring for each cluster. This visualization shows that the resulting clusters show quite good separation, with most of the data in each cluster distributed separately.

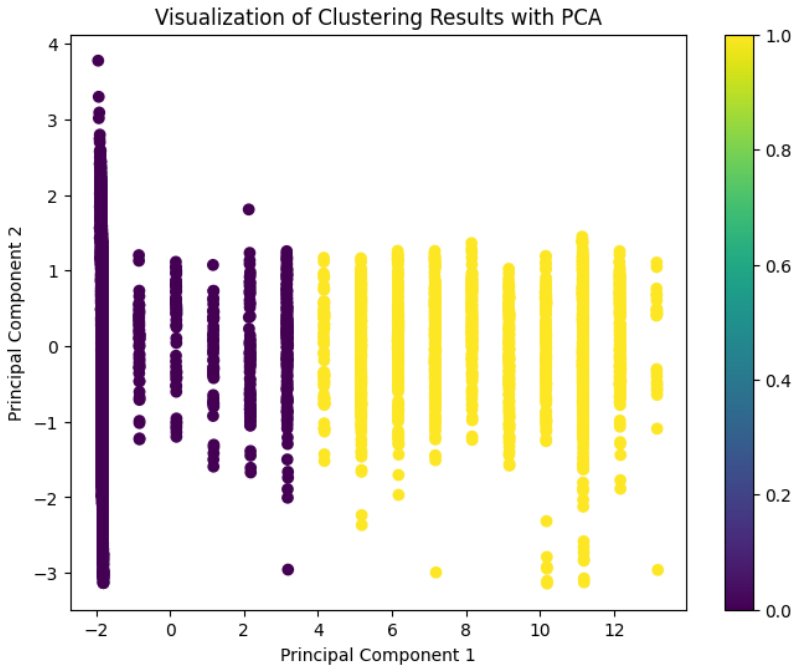


Figure 5. Visualization of Clustering Result with PCA

4.5 Evaluation

Based on Table 3, it can be seen that each cluster has a different DBI and Silhouette Score value, which are used to obtain the quality and validity of the clustering performed. DBI testing is carried out on the number of clusters C = 2 to C = 5, with variations in the applied clustering model. For your information, DBI measures the extent to which the formed clusters are separated. A smaller DBI value indicates better clustering results with higher cluster validity. Therefore, the closer to 0, the more optimal the clustering results achieved. Meanwhile, the Silhouette Score is used to assess the extent to which each data point matches its own cluster compared to other clusters. A Silhouette Score value close to 1 indicates that the formed clusters have good quality, with clear boundaries between clusters and well-degraded data. In this study, the smallest DBI value obtained in the Davies-Bouldin performance test was 0.361, indicating that the cluster formed had a reasonably good bandwidth. For the Silhouette Score, the value obtained was 0.799, indicating that the quality of the resulting cluster was quite good. These results were obtained in testing with the number of clusters 2. Anthropometric standards for determining the category of nutritional status of toddlers can be seen in Table 4.

Table 4. Cluster Test Result

No	Anthropometric Standards (Ideal Height Based on Age)
1	Baby Age 0 - 3 Months Height 40.4 - 60 cm
2	Baby Age 4 - 6 Months Height 60.5 - 66.0 cm
3	Baby Age 7 - 9 Months Height 67.5 - 70.5 cm
4	Baby Age 10 - 12 Months Height 72 - 74.5 cm
5	Toddlers aged 13 - 24 months Height 82 - 92 cm
6	Toddlers aged 25 - 60 months Height 93.1 – 118.9 cm

Based on anthropometric standards, the nutritional status of toddlers can be categorised as shown in Table 5. From the clustering results using the K-Means algorithm presented in Table 5, it can be concluded that Cluster 0 includes six sub-districts that show a relatively low prevalence of malnutrition from the number of toddlers in the sub-district. Meanwhile, cluster 1 consists of ten sub-districts that show a relatively high prevalence of malnutrition, where most of the sub-districts in this cluster face serious challenges related to malnutrition in toddlers. This condition requires immediate handling to reduce the number of malnutrition and improve the nutritional status of children in the area.

Table 5. Information on the Number of Toddlers Based on Nutritional Status, Sub-District and Cluster

Subdistrict	Good_Nutrition	Malnutrition	More_Nutrition	Obesity	Total_Toddler	Cluster
0	6.998	240	298	102	7.638	0
1	46	4	0	0	50	0
2	40	8	2	1	51	0
3	44	4	3	0	51	0
4	74	12	1	0	87	0
5	136	37	1	1	175	1
6	63	4	4	1	72	0
7	183	29	2	2	216	1
8	119	22	8	2	151	1
9	164	40	4	2	210	1
10	103	13	3	1	120	1
11	95	21	4	1	121	1
12	95	19	0	0	114	1
13	304	50	3	3	360	1
14	126	28	2	0	156	1
15	16	6	0	0	22	1

5. CONCLUSION

This study uses the K-Means algorithm with the elbow method to cluster the nutritional status of toddlers in the city and evaluate the distribution of malnutrition status in various sub-districts. The evaluation of the number of clusters was carried out using two main metrics, namely the DBI and the Silhouette Score, which produced an optimal configuration with several clusters of 2, where the DBI value was 0.361, and the Silhouette Score was 0.799. The clustering results identified the distribution pattern of malnutrition status of toddlers in each sub-district in the city. The results of the nutritional status clustering showed that cluster 0 (low) consisted of 6 sub-districts, and cluster 1 (high) consisted of 10 sub-districts. The formation of this cluster facilitates the identification of areas with a high prevalence of malnutrition status so that it can be a strong basis for designing more targeted and effective nutritional interventions to overcome malnutrition problems in the city.

ACKNOWLEDGEMENT

The authors would like to thank the two anonymous reviewers who have provided valuable suggestions to improve the article.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Femmi Widyawati: Data Preparation, Modeling, Validation, Writing – Original Draft Preparation;
Ahmad Yahya Dawod: Conceptualization, Methodology, Supervision, Review – Editing;
Heru Agus Santoso: Final approval of the version to be submitted.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] P. Arum, I. Nurmawati, N. Muna, I. Muflihatin, D. R. P. Mudiono, and A. P. Wicaksono, "Comparison of mother's and toddler's characteristics based on the nutritional status of the toddler," *International Journal of Health Information Systems*, vol. 1, no. 2, pp. 63–69, 2023, doi: 10.47134/ijhis.v1i2.4.

[2] Kementerian Kesehatan Indonesia, *Menuju Solusi Gizi Seimbang: Tantangan Dan Langkah-Langkah Konkrit Di Indonesia*. 2021.

[3] Y. O. Sari, A. Aminuddin, F. Hamid, P. Prihantono, B. Bahar, and V. Hadju, "Malnutrition in children associated with low growth hormone (Gh) Levels," *Gaceta Sanitaria*, vol. 35, pp. S327–S329, 2021, doi: 10.1016/j.gaceta.2021.10.046.

[4] Kementerian Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2020 Tentang Standar Antropometri Anak*. 2020.

[5] Kemenkes RI, "Status Gizi SSGI 2022," *BKPK Kemenkes RI*, pp. 1–156, 2022.

[6] J. H. A. M. Ikotun, A. E. Ezugwu, L. Abualigah, and B. Abuhaija, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences (New York)*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[7] M. Ula, A. Faridhatul Ulva, and I. Sahputra, "2021) 1910-1914 Accredited," *Jurnal Mantik*, vol. 5, no. 3, pp. 1910–1914, 2021.

[8] S. D. Raihannabil, "Penerapan metode hierarchical clustering untuk klasterisasi provinsi di Indonesia berdasarkan indikator status gizi anak baduta (bawah dua tahun) tahun 2023," *Emerging Statistics and Data Science Journal*, vol. 2, no. 3, pp. 424–436, 2024, doi: 10.20885/esds.vol2.iss.3.art32.

[9] W. Hadikurniawati, K. D. Hartomo, and I. Sembiring, "Spatial clustering of child malnutrition in central Java: A comparative analysis using K-means and DBSCAN," in *Proceedings of ICMERALDA 2023 - International Conference on Modeling E-Information Research Artificial Learning Digital Applications IEE Xplore*, 2023, pp. 242–247, doi: 10.1109/ICMERALDA60125.2023.10458202.

[10] O. Ardhiyanto, M. S. Asyidqi, A. Y. P. Yusuf, and T. A. Munandar, "Clustering of child nutrition status using hierarchical agglomerative clustering algorithm in bekasi city," *International Journal of Information Technology and Computer Science Applications*, vol. 1, no. 3, pp. 122–128, 2023, doi: 10.58776/ijitcsa.v1i3.42.

[11] G. Bazoukis, J. Hall, J. Loscalzo, and E. M. Antman, "The inclusion of augmented intelligence in medicine : A framework for successful implementation," pp. 1-8, 2022, doi: 10.1016/j.xcrm.2021.100485.

[12] K. D. A. N. K-medoids, "Klasterisasi kabupaten dan kota di jawa barat dalam kasus gizi buruk menggunakan algoritma k-means dan k-medoids," vol. 7, pp. 251–261, 2024, doi: 10.37600/tekinkom.v7i1.1387.

[13] M. D. Chandra, E. Irawan, I. S. Saragih, A. P. Windarto, A. K-means, and M. D. Chandra, "Penerapan algoritma K-means dalam mengelompokkan balita yang mengalami gizi buruk menurut provinsi," *Jurnal Teknologi Informasi dan Rekayasa Komputer*, vol. 2, no. 1, pp. 30–38, 2021, doi: 10.37148/bios.v2i1.191.

[14] M. R. Y. K. E. Lestari, A. Warmi, S. Winarni, S. Sylviani, Risnawita, and E. S. Nugraha, "Revealing the hidden pattern of under-five malnutrition prevalence distribution in West Java-Indonesia from canonical correspondence analysis and predictive the global burden of malnutrition has been an extremely critical and urgent issue. Malnutrition r," *Communications in Mathematical Biology and Neuroscience*, no. 2, pp. 1–30, 2024, doi: 10.28919/cmbn/8915.

[15] S. A. Siallagan and M. Safii, "Grouping of toddlers with malnutrition based on provinces in Indonesia using K-medoids algorithm," *Journal of Artificial Intelligence and Engineering Applications*, vol. 1, no. 1, 2021, doi: 10.59934/jaiea.v1i1.53.

[16] A. N. Sari and F. Harianto, "Regional mapping in bangkalan district based on potential indicators of total stunting using K-mode," *Media Gizi Indonesia (National Nutrition Journal)*, no. 1, pp. 76–82, 2022, doi: 10.20473/mgi.v17i1SP.76.

[17] S. K. C. White, "The future of augmented intelligence," *Bell Labs Technical Journal*, vol. 25, pp. 1–18, 2020, doi: 10.15325/BLTJ.2020.3015275.

[18] M. K. D. and I. Joe, "A deep-learned embedding technique for categorical features encoding.," *IEEE Access 9*, 2021, doi: 10.1109/ACCESS.2021.3104357.

[19] R. Shetty; G. M.; U. Dinesh Acharya and S. G.;, "Enhancing ovarian tumor dataset analysis through data mining preprocessing techniques," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3450520.

[20] P. Zhou, J. Chen, M. Fan, L. Du, Y.-D. Shen, and X. Li, "Unsupervised feature selection for balanced clustering," *Knowledge-Based Systems*, vol. 193, 2020, doi: 10.1016/j.knosys.2019.105417.

[21] P. G. *et. al*., "Fficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection technique," *IEEE Access 9*, vol. 10.1109/AC, 2021.

[22] A. Dewantoro and T. B. Sasongko, "Comparison of LSTM model performance with classical regression in predicting gaming laptop prices in Indonesia," *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 203–212, 2024, doi: 10.30871/jaic.v8i1.8137.

[23] N. Sakinah Sidik and U. Kebangsaan Malaysia AZLIN ALISA AHMAD, "Kelebihan dan Kekurangan Sukuk Blockchain: Satu Sorotan Literatur," 2021, doi: 10.26475/jcil.2021.6.2.13.\

[24] S. Palaniappan, R. Logeswaran, S. Khanam, and Z. Yujiao, "Machine Learning Model for Predicting Net Environmental Effects", *Journal of Informatics and Web Engineering*, vol. 4, no. 1, pp. 243–253, Feb. 2025.

[25] F. Sutomo *et al.*, "Optimization of the K-nearest Neighbors algorithm using the Elbow Method on stroke prediction," *Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 125–130, 2023, doi: 10.52436/1.jutif.2023.4.1.839.

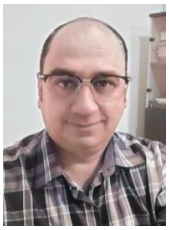
[26] R. Hariyanto and M. Z. Sarwani, "Optimizing K-measn algorithm using Particle Swarm Optimization to group student learning processes," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 6, no. 1, pp. 65–68, 2021, doi: 10.25139/inform.v6i1.3459.

[27] R. G. Prasasti Alam and Y. Everhard, "Optimasi K-means dengan Particle Swarm Optimization (PSO) dalam penentuan titik awal pusat klaster data telekomunikasi," *Techno.Com*, vol. 23, no. 1, pp. 96–111, 2024, doi: 10.62411/tc.v23i1.9743.

[28] S. Ramadhani, D. Azzahra, and T. Z, "Comparison of K-means and K-medoids algorithms in text mining based on Davies Bouldin Index testing for classification of student's thesis," *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 13, no. 1, pp. 24–33, 2022, doi: 10.31849/digitalzone.v13i1.9292.

[29] M. Sia, K.-W. Ng, S.-C. Haw, and J. Jayaram, "Chronic disease prediction chatbot using deep learning and machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 742–751, Feb. 2025, doi: 10.11591/eei.v14i1.8462.

BIOGRAPHIES OF AUTHORS

	**Femmi Widyawati** is an Undergraduate Computer Science student at Dian Nuswantoro University in Semarang, Indonesia. His primary interests are in data analytics, machine learning, artificial intelligence and website development, and holding a degree in Computer Science with specialization in Data Science. She can be contacted at email: femmiwdywti@gmail.com.
	Asst. prof. Dr. Ahmad Yahya Dawod is currently a lecturer at International College of Digital Innovation at Chiang Mai University, Thailand. He received his Ph.D. degree in Machine Learning and Artificial Intelligence from the National University of Malaysia in 2018 with the topic "Hand gesture recognition based on isolated and continuous sign language". He also graduated with his master's degree in computing and informatics from Multimedia University of Malaysia and had his bachelor's degree in computer science from The University of Mustansirya of Iraq. His research includes machine learning, pattern recognition, computer vision, robotics, and artificial intelligence. He has published 20 articles up to date with more than a hundred citations. He can be contacted at email: ahmadyahyadawod.a@cmu.ac.th.
	Heru Agus Santoso is currently an Associate Professor and the Head of the Centre for Medical Technology Innovation Group (CEMTI) at the Faculty of Engineering, Dian Nuswantoro University, Semarang, Indonesia. He has been a member of Indonesian Engineering Association (PII) and IEEE since 2021. He has received multiple research grants for his contributions to medical informatics and has actively participated in IEEE technical committees on engineering and informatics. His primary research interests include knowledge-based systems, ontologies, and information retrieval, particularly focusing on their applications in healthcare. His interdisciplinary research bridges engineering and public health, emphasizing both technical excellence and societal. He can be contacted at email: heru.agus.santoso@dsn.dinus.ac.id.