Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2434~2442
e-ISSN: 2548-6861

Comparison of Linkage Methods in Hierarchical Clustering for Grouping Districts/Cities in East Java Based on Stunting Determinants

Dinda Rima Rachcita Putri[1], Nurissaidah Ulinnuha[2*], Putroue Kumala Intan[3]
[1,2,3]Mathematics, Science and Technology, UIN Sunan Ampel Surabaya, Indonesia
09040222054@student.uinsby.ac.id [1], nuris.ulinnuha@uinsa.ac.id [2*], puput.in@uinsa.ac.id [3]

Article Info

Article history:

Received 2025-08-26
Revised 2025-09-11
Accepted 2025-09-20

Keyword:

Agglomerative Hierarchical Clustering,
Stunting,
Centroid Linkage.

ABSTRACT

Stunting is a long-term nutritional problem that generally occurs in children under five years old and is characterized by a shorter body than other children of the same age due to continuous dietary deficiencies. As a result of the Indonesian Health Survey (SKI) conducted in 2023, the stunting rate in East Java decreased to 17.7%. In 2024, the target is to reduce it to 14%. This study aims to group regencies and cities in East Java based on indicators of child nutritional status by using five linkage approaches in the hierarchical clustering method. This study found areas with similar causes of stunting so that intervention programs can be more targeted. The analysis showed that the centroid linkage methods formed two clusters with the highest cophenetic correlation coefficient of 0.8619. The first cluster consists of 37 regencies/cities with a low stunting category, and the second cluster consists of one regency/city with a high stunting category. The model in this clustering has a silhouette value of 0.6155, which indicates that the model is in the good category.

I. INTRODUCTION

Stunting is a chronic nutritional problem caused by long-term malnutrition. It generally occurs in children (under 5 years old) and is characterized by a short body type compared to other children of the same age [1]. Childhood, which ranges from conception to the age of 2 years, is a critical stage for children's intellectual development and future health [2]. Optimal nutrition in the first two years of a child's life is essential for growth, health, and development. Malnutrition during this period is common and can have long-term consequences on a child's physical and cognitive development [3]. This topic concerns governments in various countries worldwide, including Indonesia, to further review nutritional issues for all groups, from children to adults, by providing education on the nutritional content of food consumed daily [4].

It can be seen that the prevalence of stunting is influenced by various factors, including family living conditions, employment opportunities, and clean water and sanitation facilities, which have a significant impact on child growth and development [5]. Indonesia is a country with a vast topography and varying socioeconomic conditions. Socioeconomic conditions are a determining factor in nutritional status. Conditions such as high poverty, low levels of education, and the need for access to adequate health care can lead to poor health [6]. The vast topographic area certainly causes differences in nutritional status between regions in Indonesia, which causes differences in economic, educational, and health conditions in each region [7].

The Government of the Republic of Indonesia targets to reduce stunting cases in East Java to 14% by 2024. According to the Indonesian Nutritional Status Study (SSGI) from the Ministry of Health, the dominance of stunting in Indonesia is targeted to reach 21.6% by 2022 [8]. In East Java, the spread of stunting has decreased from year to year. In 2022, the prevalence of stunting reached 19.2%, with a significant decrease of 23.5% in 2021. Despite the decline, East Java still stipulates that the prevalence of this epidemic must decrease to 18.4% in 2022 [6]. In addition, based on the Indonesian Health Survey (SKI) in 2023, East Java experienced a decrease in stunting prevalence of 17.7%, which is expected to reach 14% by the end of 2024. In particular, attention should be paid to the suitability of drinking water sources, the availability of adequate toilets, the condition of community housing, and the nutritional status of children, so that it can be

known which areas require special attention regarding dietary problems in children in balancing the consistency of reducing stunting prevalence [9].

Therefore, regional grouping in East Java was carried out using a clustering method based on the characteristics of the causes of stunting in children, so that it could help identify areas with different nutritional problems. Clustering is one of the clustering methods in data mining. Clustering is the collection of objects into a group so that one cluster contains objects that are similar and different from other objects in other clusters [10]. A fairly popular clustering method is hierarchical clustering [11]. In hierarchical clustering, there is an advantage in data combination, and the creation of a hierarchy is that similar data will be placed in a dense hierarchy. In contrast, data that is not similar will be placed in layers far apart [12].

Based on a previous study by Azzahrah and Wijayanto (2022) that grouped provinces based on maternal health services, a comparison was conducted between Agglomerative Hierarchical and K-Means. The best clustering method for 34 provinces in Indonesia was found to be the hierarchical clustering method with a similarity level approach in the form of average linkage and resulting in a total of five clusters, based on internal validation, including a connectivity index of 13.9, a dunn index of 0.18, and a silhouette index of 0.51 [13]. The agglomerative hierarchy also outperformed the K-Means method in the study of Sebriana and Hasanah (2025), who used a clustering approach to observe employment indicators in East Java in 2023. The findings showed that the hierarchical clustering method (Complete Linkage) was the most effective technique. This was obtained by comparing the validity values of the best silhouette approach (0.4317) and the largest Dunn (0.1910), as well as the smallest connectivity approach (7.8861) [14].

The research by Yahya et al (2022) in the selection of the hierarchical cluster method in Southeast Sulawesi province based on the type of disease produced a cophenetic correlation value of 0.990, with the results of the study showing that the average linkage approach is the most effective clustering technique [15]. The research of Irwan et al (2024) comparing the ward linkage and complete linkage methods in analyzing the HDI in Indonesia showed that the Ward technique with five clusters produced optimum clustering based on the smallest standard deviation ratio, which was 0.282 [16]. Another study conducted by Alfirdausy et al, which compared three hierarchical clustering algorithms (Single linkage, complete linkage, and average linkage), obtained the most considerable cophenetic coefficient value with a value of 0.8105891 by the average linkage method [17]. The research by Meilani et al. also showed that the causes of stunting in Indonesia can be grouped using the centroid linkage method as the best hierarchical clustering method with the highest cophenetic result of 0.7797 [9].

Previous research has shown that trials of various linkage methods in hierarchical clustering have demonstrated that no single method consistently produces superior cluster quality. This study used a hierarchical clustering analysis to compare several linkage techniques — single, average, complete, ward, and centroid linkage to determine the most effective method based on the Cophenetic Correlation Coefficient and Silhouette Score. It is known that the five linkage techniques have not been directly applied in previous studies on stunting data in the East Java region. To improve the precision and interpretability of clustering results in a relevant context, this study helps determine a more appropriate linkage approach based on data features. This method allows grouping districts/cities in East Java Province based on similar characteristics that cause stunting, resulting in clusters of regions with relatively identical problems. Therefore, the researchers aim to cluster districts/cities in East Java, ensure that decision-makers plan strategic programs in areas requiring special attention regarding nutrition issues, and ensure that these programs are more targeted.

II. METHOD

The data comes from the 2023 Indonesian Health Survey (SKI) conducted by the East Java BKKBN, accessed on the official website https://siga.bkkbn.go.id/. The data is summarized by district or city (aggregated data), not individual data. Therefore, each row in the data represents a district or city with indicators related to stunting. The data were used in this quantitative research. The variables included in the data analysis were the prevalence of underweight toddlers (%) and the percentage of toddlers whose weight is below the standard for their age. This condition indicates a nutritional problem that can interfere with a child's growth and development. Families at risk of stunting refer to families with several factors that can increase the likelihood of stunting in children. These factors include economic capacity, maternal and child health, parental upbringing, and living conditions. Families without a primary source of adequate drinking water are the number of families who cannot access safe and standard drinking water, both in terms of quality and availability. Families without sufficient toilets are the number of families that do not have basic sanitation facilities such as clean and usable toilets. Lastly, families with uninhabitable homes show the number of families living in houses with physical conditions or facilities that do not meet basic living needs.

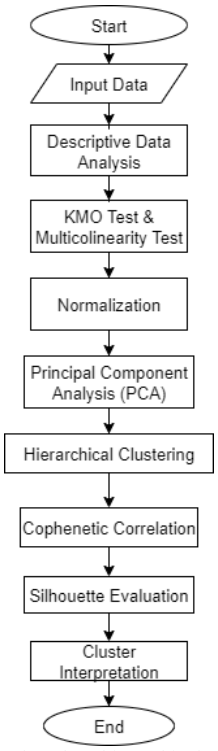


Figure 1. Flowchart hierarchical clustering

The research subjects were 38 regencies/cities in East Java Province. The data processing method used was Google Colab. Using the Hierarchical Clustering method, the cluster data was processed using the following processes and analyses:

1. Collecting input information
2. Descriptive data analysis.

   Descriptive statistical analysis of data is carried out to obtain a general overview of the data used in descriptive statistical analysis, which includes minimum, average, maximum, standard deviation, and the quantity of data [17].
3. Kaiser Meyer Olkin (KMO) test

   In cluster analysis, the KMO assumption is needed to ensure that the sample is appropriate and accurately represents the population by evaluating the suitability of each indicator and the sample as a whole. The requirement is met if the KMO value is > 0.5. With the following formula, the KMO test can determine sample adequacy [9]:

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1\,k\neq j}^{p} r_{x_j x_k}^2}{\sum_{j=1}^{p}\sum_{k=1\,k\neq j}^{p} r_{x_j x_k}^2 + \sum_{j=1}^{p}\sum_{k=1\,k\neq j}^{p} p_{x_j x_k x_l}^2} \quad (1)$$

   description:
   p = number of variables
   $r_{x_j x_k}^2$= simple correlation between variable x_j and x_k
   $p_{x_j x_k}^2$= partial correlation between variable x_j and x_k

4. Multicollinearity Test.

   Multicollinearity is vital in cluster analysis because it can affect the interpretation and results of clustering, given that this condition causes excessive information. Cluster construction becomes uninformative or fails to distinguish various cluster members effectively if highly correlated variables exist. Variance Inflation Factor (VIF) determines the presence of multicollinearity. Multicollinearity problems are indicated by VIF numbers greater than 10. Principal Component Analysis (PCA) can be used to solve multicollinearity problems [18]. One way to determine the VIF value is as follows [19]:

$$VIF = \frac{1}{1 - R_j^2} \quad (2)$$

   description:
   R_j^2 = coefficient of determination from regressing x_j on other independent variables
5. Normalization

   The data in this study used the z-score for normalization. Normalization using the z-score method is a method that uses the mean and standard deviation of the data to produce normalized values. This method remains stable even when some extreme values or data are greater than the maximum value or less than the minimum value [20].
6. Principal Component Analysis (PCA).

   Principal Component Analysis (PCA) is an analysis technique that uses linear transformation to simplify data. To facilitate data interpretation, its primary goal is to reduce the dimensionality of highly correlated variables [21]. Furthermore, the results of principal component analysis can be used for cluster analysis and other analytical calculations.
7. Hierarchical Clustering Method.
   a. Single Linkage

      The distance between each observation from different clusters is initially measured to determine the (dis)similarity or distance between clusters. The (dis)similarity measure between clusters will be the shortest or minimum distance. The cluster with the least (dis)similarity will form the dendrogram. This results in a closer dendrogram between clusters, meaning the clusters are merged at smaller distances.

      This is the formula for cluster distance with single linkage:

$$d(C_i, C_j) = \min_{x \epsilon C_i, y \epsilon C_j} d(x, y) \quad (3)$$

      description:
      C_i, C_j = cluster i dan j
      x, y = data points from different clusters
      $d(x, y)$ = distance (e.g., Euclidean) between data points
   b. Average Linkage

The first step is to measure the pairwise distances between each observation from the various clusters to determine the (dis)similarity or distance between them. The (dis)similarity measure between clusters is then determined by calculating the average pairwise distance. The cluster with the least (dis)similarity will form the dendrogram. Clusters created using this technique are often neither too "loose" nor too "dense."

The following formula uses average linkage to calculate the distance between clusters:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|}\sum_{x\epsilon C_i}\sum_{y\epsilon C_j} d(x,y) \quad (4)$$

description:

$|C_i||C_j|$ = number of data points in each cluster

c. Complete Linkage

First, the distance between each observation from different clusters is measured to determine the (dis)similarity or distance between clusters. The (dis)similarity measure between clusters will be the most significant or the maximum distance. The clusters with the least (dis)similarity will then be combined to create a dendrogram. This results in "dense" clusters, further separating the dendrogram clusters.

The following formula uses complete linkage to calculate the distance between clusters:

$$d(C_i, C_j) = \max_{x\epsilon C_i, y\epsilon C_j} d(x,y) \quad (5)$$

d. Centroid Linkage

The distance between the centroids of two clusters is measured to calculate the (dis)similarity or distance between them. Here, the mean of the x variable is used to calculate the centroid. The cluster with the smallest distance between the centroids will form the basis for the dendrogram.

The following formula uses centroid linkage to calculate the distance between clusters and the new centroid [22]:

$$d(C_i, C_j) = \|\bar{x}_i - \bar{x}_j\| \quad (6)$$

new centroid formula:

$$\bar{x}_{new} = \frac{N_i\,\bar{x}_i + N_j\bar{x}_j}{N_i + N_j} \quad (7)$$

description:

$\bar{x}_i - \bar{x}_j$ = centroid vectors of cluster C_i and C_j

$N_i + N_j$ = number of data points in cluster C_i and C_j

$\|\bar{x}_i - \bar{x}_j\|$ = euclidean distance between two centroids

e. Ward Linkage

Clusters will be created using this procedure at each iteration, and the sum of squares within each cluster will then be determined. The sum of squares can be understood as the total distance of each observation from the cluster's mean. A complete dendrogram will be created by combining the clusters that yield the smallest sum of squares.

$$SSE = \sum_{j=1}^{p}\left(\sum_{i=1}^{n} x_{ij}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right)^2\right) \quad (8)$$

Description:

n = number of data points in the cluster

p = number of variables

x_{ij} = value of variable *j* for observation *i*

ward's linkage merges clusters that inimize the increase in *SSE*

8. Cophenetic Correlation

The cophenetic correlation coefficient in a cluster tree is the linear correlation coefficient between the cophenetic distance (obtained from the dendrogram) and the original distance (dissimilarity) used to form the tree. This value measures how well the dendrogram represents the variation among observations. Technically, the cophenetic correlation coefficient is calculated as the correlation between the elements in the cophenetic matrix (formed based on the distance measure and linkage method used) and the elements in the original distance matrix (e.g., the Mahalanobis distance matrix) [23]. The equation can be shown as follows :

$$r_{coph} = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{Cik} - \bar{d}_C)}{\sqrt{\left[\sum_{i<k}\left[\left(d_{ik} - \bar{d}\right)^2\right]\left[\sum_{i<k}\left[\left(d_{Cik} - \bar{d}\right)^2\right]\right]\right]}} \quad (9)$$

Description:

d_{ik} = original distance between object *i* and *k*

d_{Cik} = cophenetic distance from the dendogram

$\bar{d}, \bar{d}_C$ = mean of original and cophenetic distances

9. Silhouette Evaluation

The number of clusters and the assessment of clustering quality can be determined using the Silhouette Coefficient. The number of clusters that has the highest average Silhouette Coefficient value or is closest to 1 is the ideal number of clusters [24].

$$s(i) = \left(\frac{b(i) - a(i)}{max(a(i), b(i))}\right) \quad (10)$$

description:

$a(i)$ = average distance of object *i* to all points in the same cluster

$b(i)$ = minimum average distance of object *i* to points in another cluster

$s(i)\ \epsilon\ [-1,1]$, closer to 1 indicates better clustering quality

10. Interpretation of the best cluster mapping results.

Determine the type of cluster based on the highest and lowest centroid values and display the ideal cluster results in a table and a map of East Java.

III. Results and Discussion

The analytical method used to group districts/cities based on stunting indicators is hierarchical clustering. This approach facilitates identifying groups of regions with essential characteristics related to the causes of stunting. In this study, the single linkage, average linkage, complete linkage, Ward linkage, and centroid linkage methods were analysed to determine the most appropriate approach.

At this stage, we need to describe the properties of the variables used in the research, including the minimum, maximum, average, and distribution values of each variable.

Table I
Stunting Indicator Data Description

Variables	N	Min	Max	Mean	Standard Deviation
Prevalence of Underweight Toddlers (%)	38	2,8	25,7	12,88	4,73
Families at Risk of Stunting	38	744	194523	72582,23	48999,36
Unsuitable Drinking Water Source	38	68	62920	11662,53	13632,93
Unsuitable Toilets	38	1126	217195	47594,42	44694,76
Uninhabitable House	38	3610	170519	55687,08	39362,92

Based on Table I, Probolinggo Regency has the highest prevalence of underweight toddlers, while Situbondo Regency has the lowest. For the indicator of families at risk of stunting, the lowest number is found in Blitar City, while the highest is in Malang Regency. Meanwhile, Madiun City consistently ranks lowest on the indicators of families with unsuitable drinking water sources, unsuitable toilets, and uninhabitable housing. Conversely, Jember Regency has the highest number of families across all three indicators. After conducting descriptive data analysis, the Kaiser-Meyer-Olkin (KMO) test was used to determine whether the sample size was adequate.

Table II
KMO Analysis

KMO Overall	0.7918348752851379

Table II shows the KMO value of the data at 0.7918, which is higher than 0.5. This indicates that the sample is representative, or the assumption of sample adequacy is met. Next, the VIF value was calculated to detect multicollinearity. The results are shown in Table III, with a VIF value greater than 10.

Table III
VIF Value on Stunting Indicator

X1	X2	X3	X4	X5
3,093	14,311	4,502	8,424	15,106

Two variables, X2 (Families at Risk of Stunting) and X5 (Families Without Adequate Toilets), have VIF values greater than 10. This indicates that the multicollinearity assumption is not met. Before conducting PCA and clustering analysis, all variables were transformed using the z-score standardization method. This standardization ensures that each variable has the same scale, with a mean of zero and a standard deviation of one. This ensures that no variable is too large or too small due to differences in units or value ranges, which could affect the analysis results. Principal component analysis (PCA) can be used afterwards to address the multicollinearity problem.

Table IV
Main Component Description

Main Components	Root Value of Characteristics	Proportion (%)	Cumulative Proportion (%)
PC1	3,475	67,68	67,68
PC2	0,949	18,49	86,17
PC3	0,380	7,41	93,58
PC4	0,211	4,11	97,70
PC5	0,117	2,29	100

The percentage of variance accounted for by each principal component is shown in Table IV. The scree plot in Figure 2 aims to calculate the principal components used. When the line reaches the second principal component, it forms the smallest angle (resembling a right angle). Consequently, the cluster analysis will be conducted using two principal components, PC_1 and PC_2, with a total explained variance of 86.17%.

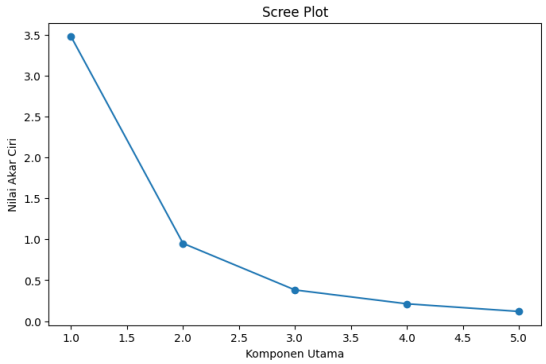


Figure 2. Scree plot of the number of principal components used

Table V
PCA Loading Value

	X1	X2	X3	X4	X5
PC1	0.205960	0.494380	0.466089	0.501962	0.493839
PC2	0.954329	-0.233718	0.038753	-0.017237	-0.181197

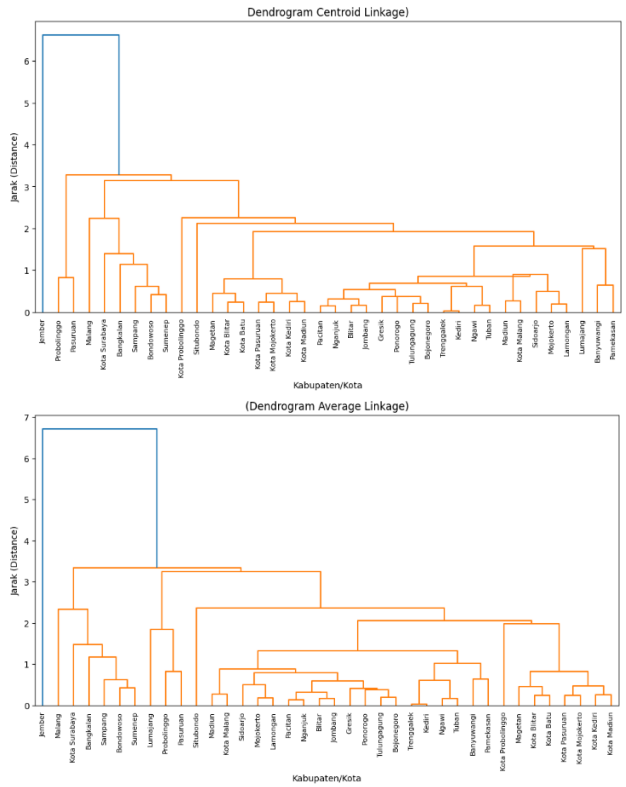
Table V displays the dominant variables in each principal component, with PC1 having four variables that yield significantly higher positive values. This indicates that variables X2 to X5 are reduced and interpreted through PC1, which covers environmental conditions and the risk of stunting. PC2 only has one dominant variable, X1 (Prevalence of Underweight Toddlers). The data transformation results using two principal components are then displayed in Table VI.

TABLE VI
PCA COEFFICIENT VALUE

No	PC1	PC2
1	-0.191087	-0.125596
2	-0.392117	0.168603
3	0.311033	0.100036
...	...	...
37	1.462094	-2.524904
38	-1.992085	0.542469

Each district/city is represented by two principal components (PC1 and PC2). These two elements can explain 86.17% of the data variation. After reducing the dimensionality and eliminating correlations between variables, the data transformation results were processed for clustering. Several linkage methods, including single linkage, complete linkage, average linkage, ward linkage, and centroid linkage, were applied in this hierarchical clustering process.

Figure 3 shows that the arrangement of each dendrogram has different distances. However, further cluster analysis should be conducted by comparing the cophenetic correlation values of the five hierarchical clustering approaches. The optimal approach in categorizing stunting levels in districts/cities is the approach with the highest cophenetic coefficient value.



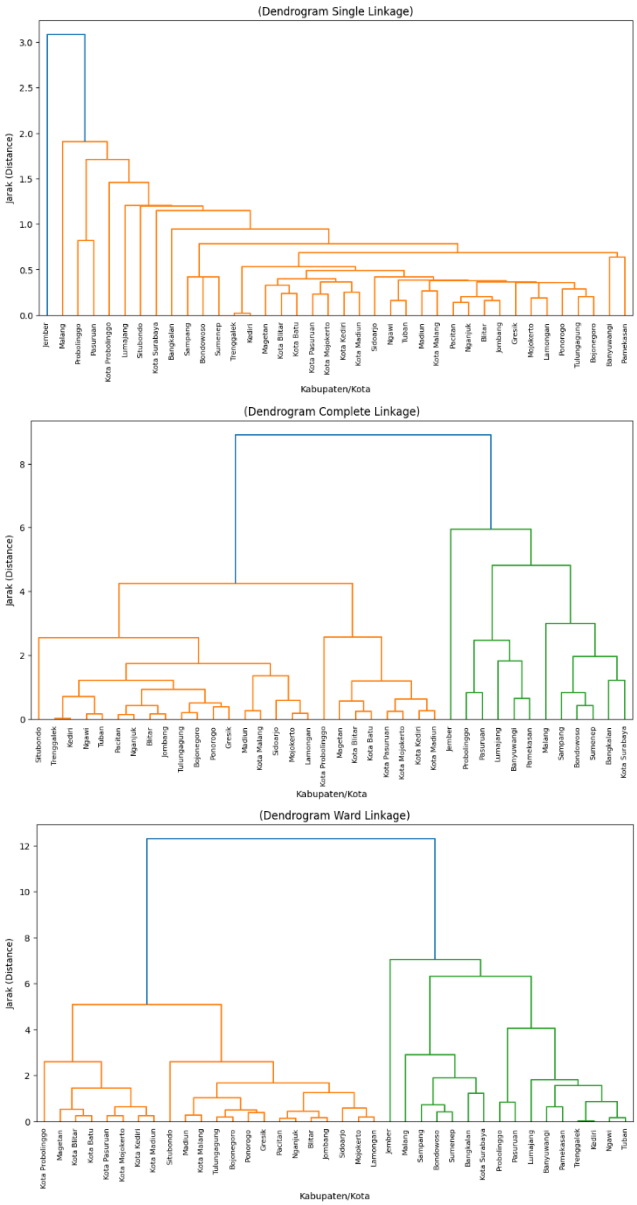


Figure 3. Dendrogram of each Hierarchical Clustering method

TABLE VII
COPHENETIC COEFFICIENT VALUE

HC Approach	Cophenetic Coefficient
Single Linkage	0,8227
Average Linkage	0,8505
Complete Linkage	0,6764
Centroid Linkage	**0,8619**
Ward Linkage	0,5435

Table VII shows that the Centroid Linkage method, with a value of 0.8619, has the highest cophenetic coefficient among the five methods tested. Therefore, this method is the optimal clustering technique for grouping regions based on the causes of stunting. Furthermore, the silhouette coefficient value will be used to calculate the optimal number of clusters for the best method.

TABLE VIII
SILHOUETTE COEFFICIENT VALUE

Number of Clusters	Silhouette Score
2	**0,6155**
3	0,3962
4	0,4776
5	0,3357
6	0,3087
7	0,2571
8	0,3854
9	0,3904
10	0,3637

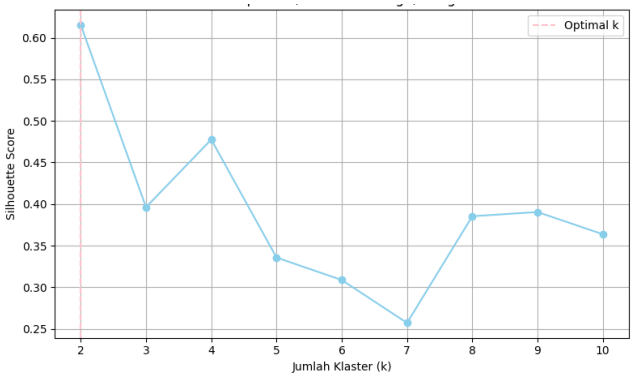


Figure 4. Silhouette Score Optimal Number of Clusters

Table VII and Figure 4 show that the optimal number of clusters for the centroid linkage method is 2, with the highest silhouette value of 0.6155. This value indicates that the clustering results have a pretty good level of separation. The silhouette value theoretically ranges from -1 to 1, where values closer to 1 indicate highly compact and well-separated clusters. Thus, this clustering result can be considered sufficiently good in dividing the data according to the characteristics of the causes of stunting in East Java.

TABLE IX
AVERAGE VARIABLES IN EACH CLUSTER

Variable	Cluster		National
	1	2	
Environmental Conditions and the Risk of Stunting	-0,172025	6,364939	1,168656
Child Nutritional Status	-0,027920	1,033055	8,764919

Based on Table VI, PC1 can be called 'Environmental Conditions and Stunting Risk', and PC2 is called 'Child Nutritional Status'. The national average is used as a reference to classify each variable within a cluster as having a good or bad value. Further examination of Table VIII reveals that several variables within cluster 2 are higher than those outside cluster 1. This indicates that districts/cities within cluster 2 have higher levels of stunting-causing characteristics, while districts/cities within cluster 1 have lower levels of stunting-causing characteristics.

TABLE IX
RESULTS OF DISTRICT/CITY CLUSTER MAPPING

Cluster	Regency/City
1	Malang, Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Kediri, Lumajang, Banyuwangi, Bondowoso, Situbondo, Probolinggo, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Bojonegoro, Tuban, Lamongan, Gresik, Bangkalan, Sampang, Pamekasan, Sumenep, Kota Kediri, Kota Blitar, Kota Malang, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Surabaya, dan Kota Batu
2	Jember

Table X shows that there are two clusters with the number of 37 regencies/cities in cluster 1 (low stunting-risk cluster), which include Malang, Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Kediri, Lumajang, Banyuwangi, Bondowoso, Situbondo, Probolinggo, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Bojonegoro, Tuban, Lamongan, Gresik, Bangkalan, Sampang, Pamekasan, Sumenep, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City, and Batu City. Jember becomes the only regency/city in cluster 2, or a high stunting-risk cluster.

The regencies/cities in cluster 2, namely Jember, have a relatively high stunting prevalence rate of 29.7% in 2023. This pushes this region into the high cluster group. Further attention is needed to address stunting in relatively high-risk areas to ensure more equitable stunting prevalence. Jember Regency is included in the high stunting category because indicators of poverty and poor access to basic infrastructure, such as families at risk of stunting, uninhabitable housing, and inadequate sanitation, are far from national standards. This phenomenon is quite natural in the geographic context of East Java, as Jember is a regency with a vast area, high population density, and uneven infrastructure distribution, which makes equitable distribution of basic infrastructure and nutrition services a significant challenge.

TABLE XI
PREVIOUS RESEARCH

Location	n Cluster	Method	Best Silhouette Score
West Java	4	K-Means [25]	0.33
A community health center	2	PCA & K-Means [26]	0.58
Indonesian's Provincies	2	PCA & K-means [27]	0.55
East Java	2	Hierarchical Clustering (our research)	0,6155

Previous research at Table XI has shown that the quality of clustering results depends on the method used. This comparison demonstrates that method selection is crucial in forming more transparent and more representative clusters. Therefore, the hierarchical clustering method was chosen in this study because it offers a more understandable and interpretable cluster structure, thus supporting stunting analysis based on multi-variable indicators.

A map of East Java based on clusters is also shown in Figure 5. This map illustrates the distribution of districts/cities based on the clustering results. Cluster 1 (low cluster), shown in green, comprises 37 districts/cities across East Java. Meanwhile, Cluster 2 (high cluster), marked in red, consists solely of Jember District, indicating that this area has distinct characteristics compared to other regions, particularly regarding stunting risk factors.

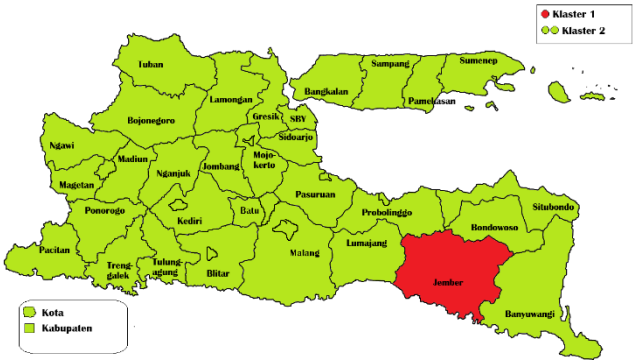


Figure 5. Map of East Java Regency/City based on clusters

IV. Conclusion

Based on the characteristics of stunting causes, regencies/cities in East Java were divided into two clusters. The clusters were created using the centroid linkage with the highest cophenetic correlation coefficient value. Regencies/cities with low stunting causes formed the first cluster, which consists of 37 regencies/cities, while those with high stunting causes formed the second cluster, represented only by Jember Regency. The prevalence of stunting was significantly influenced by the number of underweight toddlers and families at risk due to inadequate drinking water sources, inadequate toilets, and uninhabitable houses. This indicates that household environmental hygiene has a significant influence. The results of this categorization have important policy implications. Regions in the high-risk group, such as Jember, require greater focus on nutrition management programs, improving sanitation, and enhancing basic facilities. However, regions in the low-risk group also require prevention and monitoring programs to prevent future increases in stunting cases. Therefore, the results of this categorization can help local governments develop more effective, flexible, and locally tailored stunting management strategies.

Acknowledgment

Thanks to the Dalduk Division of the Ministry of Population and Family Development/BKKBN Representative Office of East Java Province for their cooperation and guidance in completing this research.

References

[1] M. R. Anggraeni, U. Yudatama, dan M. Maimunah, "Clustering Prevalensi Stunting Balita Menggunakan Agglomerative Hierarchical Clustering," *mib*, vol. 7, no. 1, hlm. 351, Jan 2023, doi: 10.30865/mib.v7i1.5501.

[2] C. G. Victora *dkk.*, "Maternal and child undernutrition: consequences for adult health and human capital," *The Lancet*, vol. 371, no. 9609, hlm. 340–357, Jan 2008, doi: 10.1016/S0140-6736(07)61692-4.

[3] World Health Organization, "Infant and young child feeding : model chapter for textbooks for medical students and allied health professionals," hlm. 99, 2009.

[4] S. D. Raihannabil, "Penerapan Metode Hierarchical Clustering untuk Klasterisasi Provinsi di Indonesia berdasarkan Indikator Status Gizi Anak Baduta (Bawah Dua Tahun) Tahun 2023: Penerapan Metode Hierarchical Clustering untuk Klasterisasi Provinsi di Indonesia berdasarkan Indikator Status Gizi Anak Baduta (Bawah Dua Tahun) Tahun 2023," *ESDS*, vol. 2, no. 3, hlm. 424–436, Okt 2024, doi: 10.20885/esds.vol2.iss.3.art32.

[5] K. Komalasari, E. Supriati, R. Sanjaya, dan H. Ifayanti, "Faktor-Faktor Penyebab Kejadian Stunting Pada Balita," *maj. kesehat. indones.*, vol. 1, no. 2, hlm. 51–56, Okt 2020, doi: 10.47679/makein.202010.

[6] A. N. A. M. Pertiwi dan L. Y. Hendrati, "Literature Review: Analisis Penyebab Kejadian Stunting Pada Balita Di Provinsi Jawa Timur," *Amnt*, vol. 7, no. 2SP, hlm. 320–327, Des 2023, doi: 10.20473/amnt.v7i2SP.2023.320-327.

[7] S. Wulandari, "Clustering Indonesian Provinces on Prevalence of Stunting Toddlers Using Agglomerative Hierarchical Clustering," *FaktorExacta*, vol. 16, no. 2, Jul 2023, doi: 10.30998/faktorexacta.v16i2.17186.

[8] Syahrul Aziz Pamungkas, "Implementasi Metode Agglomerative Hierarchical Dalam Penelitian Klasterisasi Indeks Khusus Penanganan Stunting", 2024.

[9] D. Meilani, M. Masjkur, dan F. M. Afendi, "Grouping Provinces in Indonesia Based on the Causes of Stunting Variables using Hierarchical Clustering Analysis: Pengelompokan Provinsi di Indonesia Berdasarkan Peubah Penyebab Stunting Menggunakan Analisis Cluster Hierarki," *IJSA*, vol. 7, no. 1, hlm. 32–43, Okt 2023, doi: 10.29244/ijsa.v7i1p32-43.

[10] S. Wulandari, "Clustering Kecamatan Di Kota Bandung Berdasarkan Indikator Jumlah Penduduk Dengan Menggunakan Algoritma K-Means," 2020.

[11] G. R. Suraya dan A. W. Wijayanto, "Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting: Perbandingan Metode Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia Menurut Indeks Khusus Penanganan Stunting," *IJSA*, vol. 6, no. 2, hlm. 180–201, Agu 2022, doi: 10.29244/ijsa.v6i2p180-201.

[12] Ersa Riga Puspita dan Mujiati Dwi Kartikasari, "Identifying The Cluster Of Families At Risk Of Stunting In Yogyakarta Using Hierarchical And Non-Hierarchical Approach," *kursor*, vol. 12, no. 4, hlm. 159–166, Des 2024, doi: 10.21107/kursor.v12i4.358.

[13] A. Azzahra dan A. W. Wijayanto, "Comparison of Agglomerative Hierarchical and K-Means in Grouping Provinces Based on Maternal Health Services," *SISTEMASI*, vol. 11, no. 2, hlm. 481, Mei 2022, doi: 10.32520/stmsi.v11i2.1829.

[14] E. I. Sebriana dan S. H. Hasanah, "Analisis Indikator Ketenagakerjaan Di Jawa Timur Tahun 2023 Dengan Pendekatan Clustering," vol. 2, no. 1, 2025.

[15] I. Yahya, G. N. A. Wibawa, dan L. Laome, "Penggunaan Korelasi Cophenetic Untuk Pemilihan Metode Cluster Berhierarki Pada Mengelompokkan Kabupaten/Kota Berdasarkan Jenis Penyakit Di Provinsi Sulawesi Tenggara Tahun 2020," 2022.

[16] I. Irwan, W. Sanusi, dan A. Hasanah, "Perbandingan Analisis Cluster Metode Complete Linkage dan Metode Ward dalam Pengelompokkan Indeks Pembangunan Manusia di Sulawesi Selatan," *JMATHCOS*, vol. 7, no. 1, hlm. 75–86, Apr 2024, doi: 10.35580/jmathcos.v7i1.2089.

[17] R. J. Alfirdausy, N. Ulinnuha, dan Moh. Hafiyusholeh, "Analysis of Regency/City Human Development Index Data in East Java Through Grouping Using Hierarchical Agglomerative Clustering Method," *SISTEMASI*, vol. 12, no. 3, hlm. 811, Sep 2023, doi: 10.32520/stmsi.v12i3.2959.

[18] Jurusan Matematika, Fakultas MIPA, Universitas Lampung *dkk.*, "Simulasi Pemilihan Metode Analisis Cluster Hirarki Agglomerative Terbaik Antara Average Linkage Dan Ward Pada Data Yang Mengandung Masalah Multikolinearitas," *JSM*, vol. 1, no. 2, Sep 2020, doi: 10.23960/jsm.v1i2.2497.

[19] N. Sari, H. Yasin, dan A. Prahutama, "Geographically Weighted Regression Principal Component Analysis (Gwrpca) Pada Pemodelan Pendapatan Asli Daerah Di Jawa Tengah".

[20] R. G. Whendasmoro dan J. Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," *Jur. Ris. Kom.*, vol. 9, no. 4, hlm. 872, Agu 2022, doi: 10.30865/jurikom.v9i4.4526.

[21] S. Mishra *dkk.*, "Principal Component Analysis," *Int. J. Livest. Res.*, hlm. 1, 2017, doi: 10.5455/ijlr.20170415115235.

[22] R. Silvi, "Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokkan Indikator HIV/AIDS di Indonesia," *JMM*, vol. 4, no. 1, hlm. 22–31, Mei 2018, doi: 10.15642/mantik.2018.4.1.22-31.

[23] I. Yahya, G. N. A. Wibawa, dan L. Laome, "Penggunaan Korelasi Cophenetic Untuk Pemilihan Metode Cluster Berhierarki Pada Mengelompokkan Kabupaten/Kota Berdasarkan Jenis Penyakit Di Provinsi Sulawesi Tenggara Tahun 2020," 2022.

[24] Krisman Pratama Simanjuntak dan Ulfa Khaira, "Pengelompokkan Titik Api Di Provinsi Jambi Dengan Algoritma Agglomerative Hierarchical Clustering: Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm," *Indonesian Journal of Machine Learning and Computer Science*, vol. 1, no. 1, hlm. 7–16, Apr 2021.

[25] A. Tiara, M. R. Mujahid, dan N. P. Salsabila, "Analisis Klasterisasi Mengenai Tingkat Prevelensi Stunting di Jawa Barat Tahun 2023," vol. 1, no. 1, 2025.

[26] D. Sartika, F. Elfaladonna, dan A. Octarina, "Kombinasi Hybrid K-Means Untuk Klasterisasi Multivariat Dalam Analisis Stunting," *Vol* ., no. 1, 2025.

[27] F. S. P. Wiyono, L. Kaffi, M. H. Maulana, dan T. Damaliana, "Segmentasi Wilayah Provinsi di Indonesia Berdasarkan Indeks Penanganan Stunting Menggunakan PCA dan Partition Clustering," 2025.