

Analysis of Family Status at Risk of Stunting in Babat District using the K-Means Clustering Method

Susmita Amelia^{1*}, Retno Wardhani², Musta'in³

*^{1,2,3} Department of Informatics Engineering, Lamongan Islamic University, Indonesia
mymellia1@gmail.com ^{1*}, retzno.teknik@unisla.ac.id ²*

Abstract

Stunting is a national public health priority because it affects children's physical growth, cognitive development, and the quality of future human resources. To support stunting prevention efforts, it is important to identify areas with similar risk characteristics so that intervention programs can be implemented more effectively. This study aims to analyze the clustering of villages in Babat Regency based on stunting risk using the K-Means clustering algorithm. The variables used include the number of infants under two years of age, the number of children under five years of age, the number of pregnant women, and environmental factors represented by access to proper sanitation and clean water sources. A total of 21 villages were analyzed and grouped into three clusters representing high, medium, and low stunting risk. The clustering results showed that five villages were classified as high risk, four villages as medium risk, and 12 villages as low risk. These results were then compared with the original village data and showed that the clustering accurately represented the actual conditions, indicating that the K-Means algorithm effectively identified villages with similar stunting risk characteristics. These findings indicate that K-Means clustering can provide valuable information for local governments in determining intervention priorities, allocating resources efficiently, and improving the effectiveness of stunting prevention and management programs in Babat Regency..

Keywords: *Stunting, Analysis, Clustering, K-Means Algorithm, data mining*

1. Introduction

Stunting is a public health issue that remains a national priority because it impacts physical growth, cognitive development, and the quality of future human resources. Stunting is a condition of failure to thrive in children due to chronic malnutrition, particularly during the first 1,000 days of life [1]. According to the World Health Organization (WHO), stunting status is determined based on the height-for-age (H/U) index using the Z-score value, where children with a value of less than -2 standard deviations (SD) are categorized as being at risk of stunting [2]. Indonesia remains one of the countries with the highest stunting rates in the world. According to data from the Indonesian Ministry of Health, the stunting rate in Indonesia reached 21.6% in 2022.

Although the prevalence of stunting has decreased, reaching 24.4% in 2021, it still falls short of the WHO standard, which stipulates that the stunting prevalence should be less than 20% [3]. In developing countries like Indonesia, many parents are unable and unaware of how to provide nutritious food for their toddlers. This can result in malnutrition. A toddler's nutritional status is closely related to the parents, as it determines the quality of the family's offspring [4].

Clustering is an unsupervised machine learning method useful for grouping data into clusters or groups based on similarities or differences between the data. One machine learning method used to uncover hidden patterns or structures in data is k-means clustering [5]. K-means clustering is a method that can be used to group children based on similar characteristics. This method is a non-hierarchical data mining algorithm that divides data into groups based on the proximity between objects so that each cluster has relatively homogeneous characteristics [6].

The k-Means clustering algorithm is used to group data points into clusters by minimizing the distance from the nearest cluster center. The centroid value determines the optimal grouping of data, allowing them to be categorized into different clusters [7]. Many studies have raised cases at risk of stunting using various methods and there are also several studies that use the K-means clustering method which aims to conduct cluster analysis of sub-districts or villages at risk of stunting [8]. The K-Means method is able to optimally group the stunting status of toddlers based on anthropometric characteristics [9], K-Means can be used to support stunting management through grouping based on age [10], Factors that influence stunting are height, weight and environmental factors [11].

The K-means algorithm is relevant for grouping toddler nutritional status data based on weight and height, thus facilitating the identification of toddler nutritional status [12]. The K-means clustering method is expected to provide a deeper understanding of population groups at higher risk of stunting, allowing for targeted interventions [13]. With this method, analyzed data can be systematically processed, producing

more accurate information that can be used in decision-making [14]. It is hoped that the results of this study will contribute to the efforts of the government and other policymakers to address stunting issues more effectively [15].

2. Method

2.1. Research Stages

This quantitative study focused on processing, measuring, and analyzing numerical data using a data mining approach, specifically the K-Means Clustering method. The study began with advanced problem analysis by collecting data from each village in Babat sub-district. The collected data was then processed using the K-Means Clustering method to categorize the villages into three groups: high, medium, and low risk of stunting.

This stage begins with data collection and then continues with the selection of attributes deemed most relevant. The next step is determining the number of clusters to be used, followed by iterations of the K-Means algorithm until convergence or stable results are achieved. The clustering results are then interpreted to identify patterns, characteristics, and important information. To facilitate understanding of the entire process, a flowchart system is created, as shown in the figure 1 below.

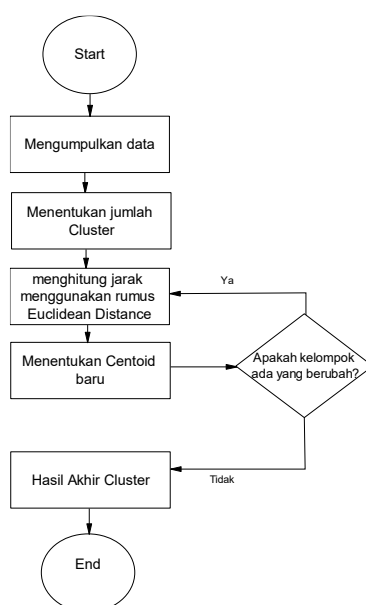


Fig. 1: Research Flow

2.2. Data Collection

Data were collected from the Babat sub-district family planning office through direct observation and documentation. The research dataset used in this cluster analysis used 21 data sets. These data have five attributes in their processing. The attributes for clustering sub-districts or villages at risk of stunting are: the number of families with infants under two years old (A), the number of families with infants under five years old (B), the number of families with pregnant women (C), the number of families with inadequate sanitation (D), and the number of families with inadequate water sources (E). These attributes serve as the basis for grouping the data and calculating distances between data sets. Based on the above attributes, the data can be presented in Table 1.

Table 1: Village Table

No	village name	A	B	C	D	E
1	Tritunggal	18	120	15	25	5
2	Puncak wangi	9	25	5	0	0
3	Patihan	10	43	9	0	0
4	Bedahan	12	32	3	0	0
5	Sogo	15	23	5	0	0
6	Sambangan	9	22	9	1	0
7	Truni	10	36	6	0	0
8	Sumur genuk	6	20	11	2	2
9	Kebalanpelang	8	19	9	0	1

Table 2: Village Table

No	village name	A	B	C	D	E
10	Keyongan	12	28	7	1	0
...	...	...	...	...	...	...
16	Kebalandono	18	103	10	6	0
17	Gendogn kulon	11	27	14	2	0
18	Bulumargi	7	22	4	4	0
19	Karang kembang	12	28	9	0	0
20	Terpan	8	19	5	0	0
21	Plaosan	10	48	14	0	2

The initial centroid values were randomly determined from the data used in the study. The initial centroids were the cluster centers. The centroids were taken from data points 3, 18, and 20. The following is the initial centroid format, presented in Table 2.

Table 3: Initial Centroid

No	village name	A	B	C	D	E
C1	Patihan	10	43	9	0	0
C2	Bulumargi	7	22	4	4	0
C3	Terpan	8	19	5	0	0

2.3. K-Means Clustering Method

The K-Means Clustering algorithm is a distance-based method that divides data into clusters. This algorithm is included in partitioning clustering, separating data into K separate regions. K-Means is a non-hierarchical data clustering method that attempts to partition existing data into one or more clusters.

1. Allocate data points to the nearest cluster

$$d(x, m) = \sum_{i=1}^n (x_i - m_i)^2. \quad (1)$$

Where :

$d(x, m)$ = is the distance between data point x and centroid m

(X_i) = is the data value for the i-th attribute.

(M_i) = is the centroid value for the i-th attribute.

(N) = is the number of attributes (variables) for which the distance is calculated.

2. Recalculate the centroid

$$c_k = \frac{1}{N_k} \sum_{j=1}^{N_k} x_i \quad (2)$$

Where :

c_k = is the new cluster centroid of the kth cluster

(N_k) = is the number of data points in the kth cluster

$x_k(j)$ = is the jth data point in the kth cluster

$\sum_{j=1}^{N_k} x_i$ = is the total number of all data points in the kth cluster

2.4. K-Means Algorithm Analysis

At this stage, the required data is sub-district data. This sub-district data consists of five attributes and totals 21 items. This data will be processed and produced three clusters that are expected to assist the government in reducing stunting.

2.5. Result and Discussion

The initial centroid values are randomly determined, with a value corresponding to the number of clusters. Next, calculations are performed using the Euclidean Distance formula. Based on the steps taken in K-Means, the initial step is to calculate the distances of the 1st, 2nd, and 3rd data points to the initial centroid.

1. Centroid Search (C1)
 - a. $C1(A) = (18+26+42+50+18)/5 = 30,8$
 - b. $C1(B) = (120+102+160+121+103)/5 = 121,2$

- c. $C1(C) = (15+7+16+14+10)/5 = 12,4$
 - d. $C1(D) = (25+20+4+4+6)/5 = 11,8$
 - e. $C1(E) = (5+5+6+5+0)/5 = 4,2$
2. Centroid Search (C2)
- a. $C2(A) = (10+12+7+10)/4 = 9,75$
 - b. $C2(B) = (43+28+22+48)/4 = 35,25$
 - c. $C2(C) = (9+7+4+14)/4 = 8,5$
 - d. $C2(D) = (0+2+4+0)/4 = 1,5$
 - e. $C2(E) = (0+0+0+2)/4 = 0,5$
3. Centroid Search (C3)
- a. $C3(A) = (9+12+5+9+10+6+8+11+12+14+10+8)/12 = 12$
 - b. $C3(B) = (25+32+23+22+36+20+19+45+25+27+28+19)/12 = 26,75$
 - c. $C3(C) = (5+3+5+9+6+11+9+9+5+14+9+5)/12 = 7,51$
 - d. $C3(D) = (0+0+0+1+0+2+0+0+20+14+9+5)/12 = 4,25$
 - e. $C3(E) = (0+0+0+0+0+2+1+5+6+0+0+0)/12 = 1,166$

Do the calculation continuously until the cluster remains in place or you could say the final cluster result is the same as the previous cluster result. The calculation results of the overall data with each cluster can be seen in Table 4.

Table 4: The most stable iteration results

No	C1	C2	C3	Closest Distance	Cluster
1	37.0049	129.5555	187.9143	37.0049	1
2	165.3625	72.2989	14.1468	14.1468	3
3	107.0278	14.0401	44.5268	14.0401	2
4	179.6117	86.6264	28.4306	28.4306	3
5	178.147	85.0507	26.6889	26.6889	3
6	135.7305	42.8675	16.8957	16.8957	3
7	151.3432	58.2891	3.6007	3.6007	3
8	135.8373	42.8675	16.3084	16.3084	3
9	139.181	46.2615	12.9151	12.9151	3
10	106.8483	14.1819	44.8679	14.1819	2
...	...	...	...	...	...
16	38.5014	57.1937	115.1411	38.5014	1
17	138.023	45.2341	14.5189	14.5189	3
18	96.7459	6.3344	55.0935	6.3344	2
19	136.3912	43.4468	15.7204	15.7204	3
20	147.9762	55.2008	6.8895	6.8895	3
21	62.7484	32.2355	90.0081	32.2355	2

3. Result and Discussion

At This stage aims to implement and compare the calculation results with the website-based system application, so that the results obtained from the application can be verified. The final results were 12 sub-districts/villages with a low risk of stunting, 4 sub-districts/villages with a moderate risk of stunting, and 5 sub-districts/villages with a high risk of stunting. The sub-districts/villages can be seen in Table 5.

3.1 Results

At This stage aims to implement and compare the calculation results with the website-based system application, so that the results obtained from the application can be verified. The final results were 12 sub-districts/villages with a low risk of stunting, 4 sub-districts/villages with a moderate risk of stunting, and 5 sub-districts/villages with a high risk of stunting. The sub-districts/villages can be seen in Table 4.

Table 5: The Most Table Iteration Results

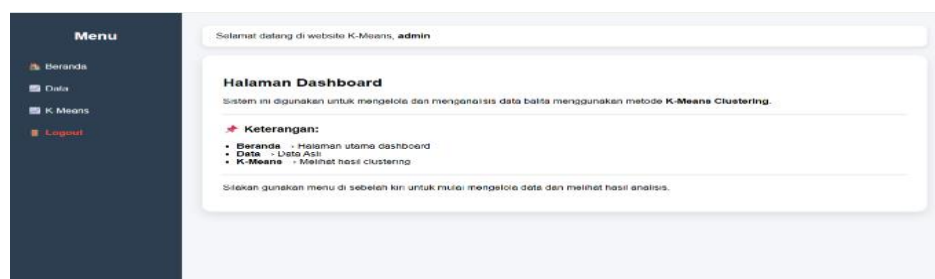
No	village name	Cluster	information
1	Tritunggal	C1	Tinggi
2	Puncak wangi	C3	Rendah
3	Patihan	C2	Sedang
4	Bedahan	C3	Rendah
5	Sogo	C3	Rendah
6	Sambangan	C3	Rendah
7	Truni	C3	Rendah
8	Sumur genuk	C3	Rendah
9	Kebalanpelang	C3	Rendah
10	Keyongan	C2	Sedang
...	...	...	...
16	Kebalandono	C1	Tinggi
17	Gendogn kulon	C3	Rendah
18	Bulumargi	C2	Sedang
19	Karang kembang	C3	Rendah
20	Terpan	C3	Rendah
21	Plaosan	C2	Sedang

3.2 Sistem Implementasion

The login page is the initial display that appears when a user first accesses a website. This page serves as the main gateway to the system, allowing only authorized users to proceed to the next page. Here, users are required to enter their previously registered username and password. This process serves to verify user identity, maintain data security, and prevent unauthorized access. If the entered data is correct, the user will be redirected to the system's main page, such as the dashboard. Conversely, if the username or password entered is incorrect, the system will deny access and require the user to re-login. Therefore, the login page plays a crucial role in maintaining security and controlling access to all website features. The login page can be seen in Figure 1 below:

**Fig.1:** Login Page

The dashboard page is the main page displayed after a user successfully logs into the system. This page serves as a navigation center for accessing various available features, such as data input and data processing results. The dashboard is designed to be simple, informative, and easy to use so that users can quickly understand the available functions and operate the system more efficiently. The dashboard page display can be seen in Figure 2.

**Figure 2:** Dashboard

On this page, the user enters the initial centroid value as the starting point for the clustering process. Next, the user determines the number of clusters, enters the amount of data to be processed, and fills in the centroid values based on the research variables. Once all data has been entered correctly, the user can click the Process button to begin the calculation using the K-Means method. The K-Means Process page display can be seen in Figure 3.

— Kembali Ke Beranda

INPUT K-MEANS

K Cluster:
3

Data dari:
1

Data sampai:
21

Centroid awal:
3, 10, 20

PROSES K-MEANS

Fig. 3: K-Means Process

After the initial centroid values are entered, the system processes the data using the K-Means Clustering method through an iterative process. At each iteration, the system calculates the distances between data points, groups them into clusters, and updates the centroid values until optimal clustering results are achieved. The K-Means Process page display can be seen in Figure 4.

C1	C2	C3	Jarak Terdekat	Cluster
37.0049	136.877	189.8937	37.0049	1
165.3625	65.0102	12.1754	12.1754	3
107.0278	6.9656	46.5048	6.9656	2
179.6117	79.3544	26.4277	26.4277	3
178.147	77.7375	24.7141	24.7141	3
135.7305	35.5741	18.8309	18.8309	3
151.3432	51.0129	4.1191	4.1191	3
135.8373	35.4982	18.3244	18.3244	3
139.181	38.9579	14.8497	14.8497	3
106.8483	7.571	46.834	7.571	2
181.416	81.0415	28.2354	28.2354	3
129.8536	29.5147	24.2948	24.2948	3

Fig. 4: Iterasi

3.3 K-Means Clustering

The grouping results are the final output from the application of the K-Means Clustering algorithm which has gone through several stages of iteration until it reaches a convergent condition. Thus, the results obtained can be said to be stable and able to represent the data conditions more accurately. The K-Means Process page display can be seen in Figure 5.

HASIL AKHIR CLUSTER

No	Nama Desa	Cluster	Keterangan
1	Tribunggal	C1	Tinggi
2	Puncak Wengi	C3	Rendah
3	Palihan	C2	Sedang
4	Bedahan	C3	Rendah
5	Sogo	C3	Rendah
6	Sambangan	C3	Rendah
7	Truni	C3	Rendah
8	Sumur Genuk	C3	Rendah
9	Kebalanpelang	C3	Rendah
10	Keyongan	C2	Sedang
11	Dati Nawong	C3	Rendah
12	Kuripan	C3	Rendah
13	Moropelang	C1	Tinggi

FIG. 5. Recapitulation of results

Based on the clustering results graph, cluster C1 (high stunting risk) consists of 5 villages, cluster C2 (moderate stunting risk) consists of 4 villages, and cluster C3 (low stunting risk) consists of 12 villages. These results are consistent with data obtained from the Babat District Family Planning Office. The graphic visualization makes it easier for users to understand data distribution and interpret the clustering results. The graphic display can be seen in Figure 6.

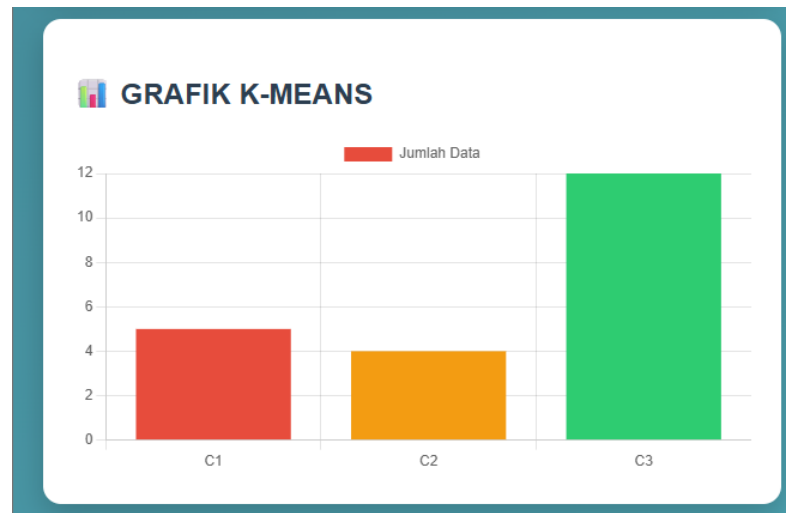


Fig. 6. Results graph

3.4 Discussion - Example of Source Code Presentation

The implementation results show that the K-Means Clustering method is able to effectively group village data based on stunting risk levels in Babat District. The developed web-based system allows users to perform the clustering process by specifying the number of clusters and the initial centroid value, resulting in the grouping of villages into three categories: high, medium, and low stunting risk. The resulting clustering results align with the original data obtained from the Babat District Family Planning Office, thus supporting decision-making in prioritizing stunting management and prevention.

To implement this clustering process, the K-Means algorithm was integrated into a web-based application. The clustering process begins with determining the initial centroid value, then the system calculates the distance of each data item to each centroid using the Euclidean distance method. Next, each data item is assigned to the cluster with the closest distance, and the centroid value is updated based on the average of each cluster's members. This process is repeated until the centroid value remains unchanged or converges. The implementation of the K-Means Clustering calculation process in the application is presented in the following code snippet.

Source Code 1: K-Means Clustering Analysis Process

```

1  for ($iter=0; $iter<20; $iter++) {
2    foreach ($subset as $d)
3      $distances = []
4      for ($i=0; $i<$k; $i++)
5        $x = $d['norm'];
6        $m = $centroids[$i]; // centroid hasil update dari iterasi
7        $sum = 0;
8        for ($j=0; $j<8; $j++) {
9          $distances[$i] = sqrt($sum);
10         $minCluster = array_keys($distances, min($distances))[0];
11         $clusters[$minCluster][] = $d['norm'];
12         $new_centroids[$i][$j] /= count($clusters[$i]);

```

Compared with previous research, the proposed model's performance is consistent with previous studies that reported good K-Means clustering performance. The K-Means method is able to optimally group the stunting status of toddlers based on anthropometric characteristics [9], K-Means can be used to support stunting management through grouping based on age [10]. The K-means algorithm is relevant for grouping toddler nutritional status data based on weight and height, thus facilitating the identification of toddler nutritional status [12].

The K-means clustering method is expected to provide a deeper understanding of population groups at higher risk of stunting, allowing for targeted interventions [13]. With this method, analyzed data can be systematically processed, producing more accurate information that can be used in decision-making [14]. This finding confirms that the K-Means Clustering algorithm remains an effective method for grouping data based on their similar characteristics.

4. Conclusion

The K-Means Clustering method was successfully applied to group the stunting risk level in 21 villages in Babat District using five research variables, namely the number of toddlers under 2 years old, the number of toddlers under 5 years old, the number of pregnant women, the number of families with inadequate sanitation, and the number of families with inadequate water sources. The clustering process was

carried out by determining the number of clusters, determining the initial centroid, calculating the distance using Euclidean Distance, and iterating until stable results were obtained.

Based on the results of the clustering process using the K-Means method, the system successfully grouped the data according to the groupings in the original data. The clustering results show that out of 21 villages in Babat sub-district, there are 5 villages with a high risk of stunting, 4 villages with a moderate risk of stunting, and 12 villages with a low risk of stunting, this has been correctly identified, so the clustering results are in accordance with the actual data conditions or original data.

References

- [1] D. Yusuf, F. Razi, S. A. Arman, V. Terisia, and R. Nurjayanti, "Prediksi Risiko Stunting pada Balita menggunakan Algoritma Logistic Regression dan Decision Tree berbasis Data Terbuka," *Pros. Semnastek*, pp. 1–7, 2025.
- [2] A. S. Hibatulloh and U. Zaky, "APLIKASI DETEKSI DINI STUNTING PADA BATITA MENGGUNAKAN KLASIFIKASI TINGGI BADAN TERHADAP UMUR BERDASARKAN Z-SCORE WORLD HEALTH ORGANIZATION," (*Jurnal Mhs. Tek. Inform.*, vol. 9, no. 6, pp. 9441–9448, 2025).
- [3] N. Nur, M. Iqam, and N. Inayah, "Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting," pp. 356–367, 2023.
- [4] A. Subayu, "PENERAPAN METODE K-MEANS UNTUK ANALISIS STUNTING GIZI PADA BALITA :," vol. 2, no. 1, 2022, doi: 10.20885/snati.v2i1.18.
- [5] M. F. Amalia and D. B. Arianto, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING DALAM KLASTERISASI KABUPATEN / KOTA PROVINSI JAWA BARAT BERDASARKAN FAKTOR PEMICU STUNTING PADA BALITA," vol. 9, no. 1, pp. 36–46, 2024.
- [6] I. S. Tinendung and I. Zufria, "Pengelompokan Status Stunting Pada Anak Menggunakan Metode K-Means Clustering," vol. 7, pp. 2014–2023, 2023, doi: 10.30865/mib.v7i4.6908.
- [7] N. Nafiyah, R. Wardhani, M. N. Huda, and A. P. Setyan, "K -means clustering for data-driven traffic accident in east java," vol. 10, no. 4, pp. 3679–3688, 2025.
- [8] D. M. Cytry, S. Defit, and G. W. Nurcahyo, "Jurnal KomtekInfo Penerapan Metode K-Means dalam Klasterisasi Status Desa terhadap Keluarga Beresiko Stunting," vol. 10, no. 3, pp. 122–127, 2023, doi: 10.35134/komtekinfo.v10i3.423.
- [9] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," 2023.
- [10] N. S. Hanapiah *et al.*, "MENINGKATKAN PENANGANAN STUNTING PADA ANAK MELALUI KLASIFIKASI PEMBERIAN MAKANAN TAMBAHAN BERDASARKAN USIA DENGAN METODE K-MEANS DI DESA CINTARASA," vol. 7, no. 6, pp. 3484–3493, 2023.
- [11] M. R. Nugroho, R. N. Sasongko, and M. Kristiawan, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Usia Dini di Indonesia," *J. Obs. J. Pendidik. Anak Usia Dini*, vol. 5, no. 2, pp. 2269–2276, 2021, doi: 10.31004/obsesi.v5i2.1169.
- [12] S. Nurul Rizki Octaviyani¹, Rini Mayasari² and Received:, "Implementasi Algoritma K-Means Clustering Status Gizi Balita Nurul Rizki Octaviyani 1 , Rini Mayasari 2 , Susilawati 3 Universitas Singaperbangsa Karawang," vol. 8, no. 13, pp. 370–381, 2022.
- [13] 4Baenil Huda 1Fuad Anwar Yuwono, 2 April Lia Hananto, 3Fitri Nurapriani, "Implementasi Algoritma K-means untuk Klasterisasi Data Stunting di Kabupaten Bekasi," vol. 9, no. 2, pp. 626–633, 2025.
- [14] T. Hardiani and R. N. Putri, "Implementasi Metode Naïve Bayes Classifier Untuk Klasifikasi Stunting Pada Balita," *Digit. Transform. Technol.*, vol. 4, no. 1, pp. 621–627, 2024, doi: 10.47709/digitech.v4i1.4481.
- [15] A. Fikri, B. F. Hutabarat, and U. Khaira, "Komparasi K-Means Clustering Dan Complete Linkage Dalam Pengelompokan Penyaluran Pinjaman Financial Technology," vol. 17, no. 2, pp. 228–239, 2023.