Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

Data-Driven K-Means Clustering Analysis for Stunting Risk Profiling of Pregnant Women

Desvita Dian Nazella[1], Heru Pramono Hadi[1,*], Farrikh Al Zami[1], Ayu Ashari[2], Yupie Kusumawati[1], Suharnawi[1], Rama Aria Megantara[3], Muhammad Naufal[3]

[1] Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Fakultas Kesehatan, Program Studi Rekam Medis dan Informasi Kesehatan, Universitas Dian Nuswantoro, Semarang, Indonesia
[3] Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]112202206874@mhs.dinus.ac.id, [2]heru.pramono.hadi@dsn.dinus.ac.id, [3]alzami@dsn.dinus.ac.id, [4]ayu.ashari@dsn.dinus.ac.id, [5]yupie@dsn.dinus.ac.id, [6]suharnawi@dsn.dinus.ac.id, [7]aria@dsn.dinus.ac.id, [8]m.naufal@dsn.dinus.ac.id
Correspondence Author Email: heru.pramono.hadi@dsn.dinus.ac.id
Submitted: **25/09/2025**; Accepted: **06/12/2025**; Published**: 08/12/2025**

Abstract−Stunting in children is influenced by maternal health conditions during pregnancy. This study aims to classify pregnant women to prevent stunting based on clinical, demographic, and environmental factors using the K-Means Clustering algorithm. A total of 229 data from the Primadona application (Disdalduk KB Kota Semarang) were analyzed using 14 normalized variables. The optimal number of clusters was determined using the Elbow Method and validated using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The Kruskal-Wallis test was performed to verify differences between clusters. This study resulted in seven clusters with different profiles, with a Silhouette Score of 0.134, Davies-Bouldin Index of 1.509, and Calinski-Harabasz Index of 29.54. These values indicate that the cluster structure is formed and reflects the variation in risk for pregnant women, although there is overlap due to differences in characteristics between individuals. The clustering successfully differentiated pregnant women with low to high risk, influenced by health and environmental factors. This study proves the effectiveness of K-Means in identifying stunting risk patterns in pregnant women and supports more targeted interventions, such as nutritional counseling, disease risk monitoring, education on cigarette smoke exposure, and referrals. Limitations of this study include the unbalanced distribution of data between and the use of cross-sectional data. Future research is recommended to improve pre-processing and compare other clustering methods such as K-Medoids or DBSCAN for more precise stunting risk analysis.

Keywords: Stunting; Pregnant Women; Risk Profilling; K-Means Clustering; Primadona App

1. INTRODUCTION

The risk profile of pregnant women is a crucial aspect in stunting prevention, but information on this is still limited, especially in Indonesia [1]. Despite various government efforts through maternal and child health programs, existing monitoring systems have not been able to identify specific risk factors based on clinical and environmental characteristics.

Maternal Mortality (MMR) remains a major global challenge.[2] reports approximately 260,000 maternal deaths due to complications of pregnancy or childbirth each year, or one death every two minutes. The total mortality rate in Indonesia is approximately 189 per 100,000 live births [3] . One of the main causes is Chronic Energy Deficiency (CED), with a prevalence of 8% [4]. Maternal nutritional status plays a significant role in increasing the risk of stunting [5]. Preeclampsia is also a worrying pregnancy complication, with a total of between 3-10% in hospitals [6]. Mid-Upper Arm Circumference (MUAC) can be a screening tool for detecting CED [4]. Hb, LiLA, and maternal weight during pregnancy are factors that contribute to stunting in children aged 1-3 years [7]. Digitizing health data, such as the regional health information system and the PRIMADONA (Provision of Data Information Homes for the Bangga Kencana Program) application developed by the Semarang City Population and Family Planning Office (Disdalduk KB) also supports the aforementioned policy to integrate and simplify data management.

The K-Means algorithm was chosen based on its ability to efficiently and stably detect pattern structures in data, with better computational performance than other methods such as K-Medoids or Hierarchical Clustering. This algorithm is also most suitable for clustering measurable numerical data, such as clinical and anthropometric parameters of pregnant women, in exploratory stunting risk analysis.

Several previous studies have used K-Means to analyze maternal health.[8] In the analysis, the nutritional status of pregnant women was grouped using K-Means to detect the risk of stunting early, determine the number of clusters using the Elbow method and evaluate their quality using the Silhouette Coefficient so that the cluster results were good. A study conducted by [9] categorized pregnant women into three homogeneous groups based on age, medical history, anemia, and pregnancy history using Sum of Square Error (SSE) analysis for quality measurement. This approach can be used by cadres to identify groups of pregnant women at similar risk so that more appropriate interventions can be provided.

K-Means, using unsupervised learning, is able to identify specific patterns and group them based on risk level [10]. The use of K-Means can reveal health patterns, especially when combined with Random Forest-based feature selection [11]. Clustering is a method for grouping data based on the level of similarity between individuals that can be applied to the K-Means method [12]. K-Means clustering is widely used to analyze stunting factors and to accurately group data based on stunting prevalence [13].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

However, several previous studies have not comprehensively addressed the impact of outliers on clustering results, even though clinical data from pregnant women often contain extreme values due to biological factors and data recording errors [14]. Furthermore, the use of Elbow or Silhouette is not sufficient to ensure model stability, necessitating a multi-metric validation approach to improve clustering effectiveness [15].

Previous research also demonstrated the use of the Elbow Method, Silhouette Index, and Davies-Bouldin Index [16] in determining the optimal number of clusters. However, these studies were limited by the small number of variables, so their effectiveness has not been tested on health data with a larger number of variables.

Besides K-Means, many other methods have been used for testing. [17] used machine learning (Random Forest) in pregnancy risk classification because they focused on unlabeled data. This study had limitations because it could not provide a clear group structure like clustering. Meanwhile, [18] introduced the Interpretable Clustering Ensemble algorithm, which produces more stable and easier-to-understand clusters. This method is rarely used and is relatively new in maternal health research.

Other methods, such as K-Medoids, are also widely used in analysis due to their resilience to outliers compared to K-Means. [19] used it in their research on COVID-19 in Indonesia, although this method has not been widely used in maternal health research. The DBSCAN method can detect irregularly shaped clusters [20], but it is sensitive to the epsilon parameter, which can affect clustering.

[21] in their study identified risk factors for stunting in children in Sambas Regency using the K-Means clustering method combined with t-SNE. This study resulted in a clustering method that can group children based on their risk, allowing for more precise prevention. However, its weakness is that it focuses only on early childhood, so analysis of mothers remains an area for further research.

2. RESEARCH METHODOLOGY

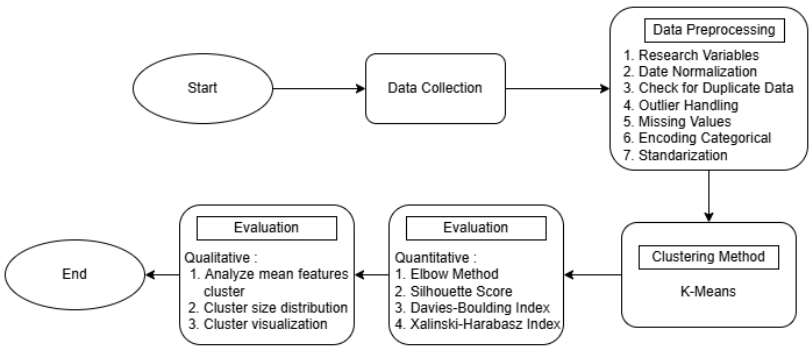

Figure 1. Stages of the Research Method

This study applied the K-Means algorithm to cluster data on pregnant women based on stunting risk factors. Figure 1 shows the research flow, which includes four main stages: data collection, pre-processing, application of the K-Means algorithm, and evaluation of the clustering results. Data collection, which explains the sources and characteristics of the dataset of pregnant women used. Data pre-processing, consisting of cleaning, normalization, and variable transformation, is then carried out. The K-Means algorithm is then applied to cluster the data of pregnant women based on the similarity of the analyzed indicators. The final stage is the quantitative and qualitative evaluation of the clustering results to assess the quality and interpretation of each cluster. A detailed explanation of each stage is presented in the following subsections.

2.1. Research Stages

This study applied the K-Means algorithm to cluster data on pregnant women based on stunting risk factors. Figure 1 shows the research flow, which includes four main stages: data collection, pre-processing, application of the K-Means algorithm, and evaluation of the clustering results.

2.2. Data Collection

The research data was obtained from the PRIMADONA application from the Semarang City Population and Family Planning Control Office (Disdalduk KB). This dataset contains reports on assistance to pregnant women in the stunting prevention program in the Central Semarang District from March to May 2025.

Overall, the March-May 2025 dataset consists of 229 entries with 23 features, including demographic data, clinical data, and environmental factors. Several pregnant women in the dataset were recorded at more than one visit.

2.3. Data Processing

Data processing was carried out in eleven stages, starting from variable identification to evaluation of the grouping results. Details of these stages are outlined in subsections 2.3.1 to 2.3.11.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

2.3.1 Research Variables

Of the 23 features in the dataset, 14 were used in the analysis, while the other nine were administrative. Identity variables such as Name, National Identification Number (NIK), and ID Data were excluded to maintain participant confidentiality. Administrative variables such as Serial Number and Status were also excluded from the analysis. The Status variable serves as a program cadre label. However, the Status variable was used in the evaluation phase as a benchmark to assess the suitability of the grouping results to the program's targets.

The features used included demographic data (maternal age, number of children, and gestational age), clinical data (weight, height, mid-upper arm circumference (MUAC), hemoglobin level, uterine fundal height (FHH), and medical history), and environmental factors (exposure to cigarette smoke, consumption of iron tablets, participation in the Family Hope Program (PKH), availability of clean water, and ownership of a toilet).

In addition, there are variables resulting from data processing, such as Body Mass Index (BMI), which is calculated based on weight and height, as well as additional risk variables, including nutritional risk (based on the Lifespan Index), anemia risk (based on hemoglobin levels), sanitation risk (based on sanitation facilities and toilet ownership), exposure to cigarette smoke, and consumption of iron supplements. The results of these variables are used to support the analysis of stunting risk factors.

2.3.2 Date Format Normalization

Date columns were not used directly in the analysis. To facilitate checking for missing values, all date columns were normalized to two-digit dd-mm-yy (day-month-year) format. In the final results, the date columns remain displayed to facilitate cadres' identification and matching of participant data.

2.3.3 Check For Duplicate Data

A data duplication check was performed to ensure there were no duplicate data that could affect the clustering results. Some participants had more than one visit during their pregnancy, and each visit represented a different maternal condition at each visit. Variables that changed between visits included weight, mid-upper arm circumference (MUAC), hemoglobin (Hb) levels, and uterine fundal height (FH). Because this analysis focused on the characteristics of the pregnant woman's condition at each measurement time, all visits were included in the clustering process.

2.3.4 Outlier Handling

The number of children column contains extreme values compared to the general distribution. These values were handled using winsorizing. Based on the data distribution, almost all participants had fewer than ten children, so the maximum value was limited to 10 to prevent extreme values from dominating the clustering results.

2.3.5 Missing Values

Check for missing values on each numeric value. Missing values are handled systematically.

a. Columns with more than 50% missing values are removed as they could potentially compromise model stability.
b. Columns with less than 50% missing values are filled using the median value to maintain a balanced data distribution.

2.3.6 Transforming Categorical Columns to Numerical Columns

Columns containing categorical data are converted into numeric form for processing by the algorithm. Variables such as health history, cigarette smoke exposure, iron supplementation tablets, Family Hope Program (PKH) participation, clean water availability, and toilet ownership are converted into binary variables (0 = no, 1 = yes).

2.3.7 Data Standardization

This normalization technique standardizes the data by setting the mean to 0 and the variance to 1. This is necessary because the K-Means algorithm is sensitive to large-scale features that can dominate the results. Standardization uses the z-score normalization method, with the formula:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

The data standardization process uses a formula as in equation (1), where z is the result of standardization. The symbol x is the raw or original data, while u shows the average value of each feature, and o is the standard deviation used as a measure of data distribution.

2.3.8 Principal Component Analysis (PCA)

The PCA method is used to reduce the dimensionality of data and display clustering results in two dimensions (2D). This technique helps visualize the separation patterns between clusters. PCA is used only for visualization purposes and does not affect the cluster formation process.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

2.3.9 Determining the Number of Clusters

The optimal number of clusters is determined using the Elbow Method, by calculating the within-cluster sum of squares (WCSS). The optimal point is determined when the WCSS curve begins to plateau (elbow point), i.e., when increasing the number of clusters no longer results in a significant decrease in the WCSS value.

2.3.10 K-Means Algorithm

K-Means is an unsupervised learning clustering algorithm capable of processing large data sets. Its weakness is the need to determine the number of clusters (k) upfront, so the Elbow Method is used as a reference. The K-Means objective function equation is shown in the following formula:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

In equation (2) of the K-Means algorithm, k represents the number of clusters, while C_i is the data set in the i-th cluster. The x value consists of a number of data that have similar characteristics, while the center point of the i-th cluster is called the cenroid (μ_i). This clustering process is carried out until the centroid position is stable and the distance between data in one cluster is minimal, so that the distribution of data in the cluster is optimal.

2.3.11 Evaluation

The evaluation was conducted using two approaches: quantitative and qualitative. The quantitative approach used four validation metrics: the Elbow Method, Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to assess cluster quality. The Elbow Method determines the optimal number of clusters based on the Within-Cluster Sum of Squares (WCSS), the Silhouette Score assesses internal coherence, and Davies-Bouldin and Calinski-Harabasz measure separation and dispersion between clusters. These four metrics can provide an evaluation of the results of the stunting risk grouping.

The qualitative approach analyzed feature averages, cluster size distributions, and visualization results to understand the characteristics of each group. Furthermore, the Kruskal-Wallis statistical test was used to verify differences between clusters on numerical variables due to its nonparametric nature and suitability for varied health data. This test was introduced by William Kruskal and W. Allen Wallis in 1952 as a non-parametric alternative. The formula is as follows:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{Rj^2}{nj} - 3(N+1) \quad (3)$$

In the Kruskal-Wallis test formula in equation (3), the H value is the statistical value resulting from the calculation of the difference in medians between groups. N represents the total number of sample data analyzed. The Rj value is the number of ranks obtained by the data in the jth group, while indicating the number of data contained in that group.

3. RESULT AND DISCUSSION

3.1. Correlation Heatmap between Features

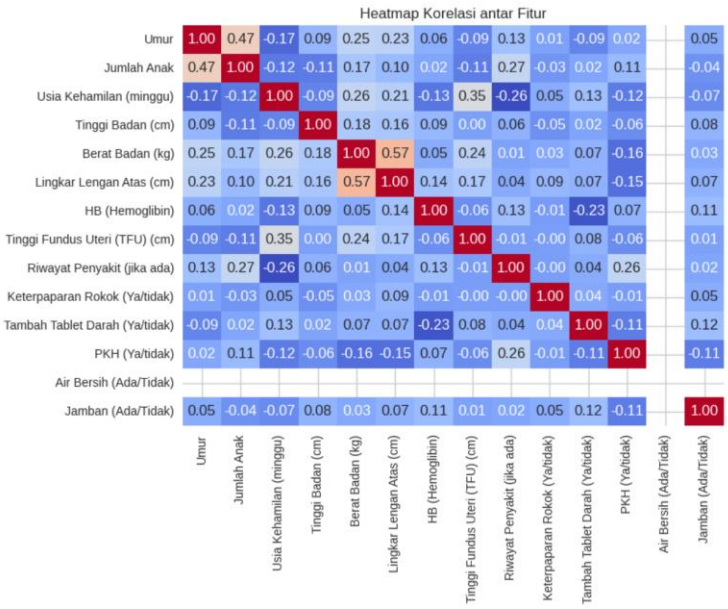

Figure 2. Correlation Heatmap between Features

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

The correlation heatmap in Figure 2 shows the relationships between variables. Red indicates a positive correlation, blue indicates a negative correlation; darker colors indicate a stronger correlation, and lower intensity colors indicate a weaker relationship.

The results show the highest correlation between Body Weight and Mid-Upper Arm Circumference (MUAC), with a value of approximately 0.57, indicating maternal nutritional status. Moderate correlations were also observed between Age and Number of Children (0.47), and between Gestational Age and Fundal Height (0.35), indicating a normal pregnancy physiology pattern. There was a negative correlation between Gestational Age and Medical History (-0.26), indicating the need for monitoring in pregnant women with a history of certain diseases. Furthermore, HB and Iron Supplement Tablets (-0.23) indicated a risk of anemia, necessitating regular consumption of iron supplements.

Furthermore, environmental factors such as the Family Hope Program (PKH), clean water, and toilet availability had very weak correlations, indicating that their influence on key health variables is indirect or relatively small.

3.2. Determining the Number of Clusters

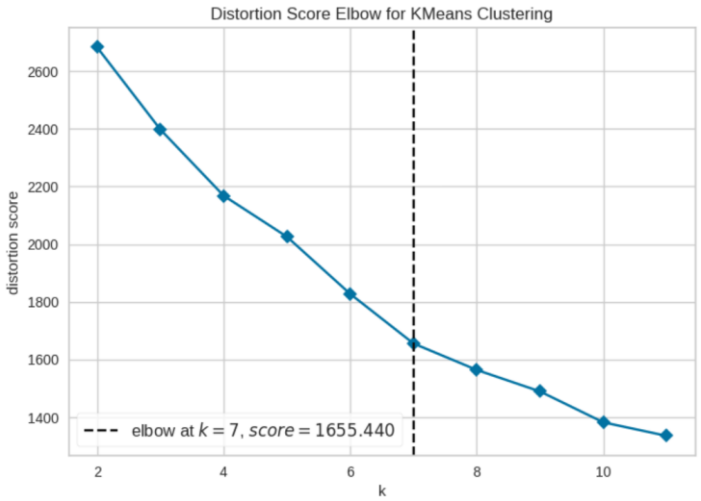

Figure 3. Determining the Number of Clusters

Figure 3 illustrates the determination of the optimal number of clusters using the Elbow Method in the k range between 2 and 12. The elbow point is seen at k = 7 with a score of 1655.440 which indicates a decrease in the Within-Cluster Sum of Squares (WCSS) before reaching a stable phase. This value is validated using the Sihouette Coefficient of 0.134 which indicates the presence of cluster differences although not all of them are clear, but still represent the characteristics of pregnant women. Thus, the selection of k = 7 is considered optimal because it produces a balance between model complexity and interpretability of results.

3.3. Cluster Distribution and Characteristics

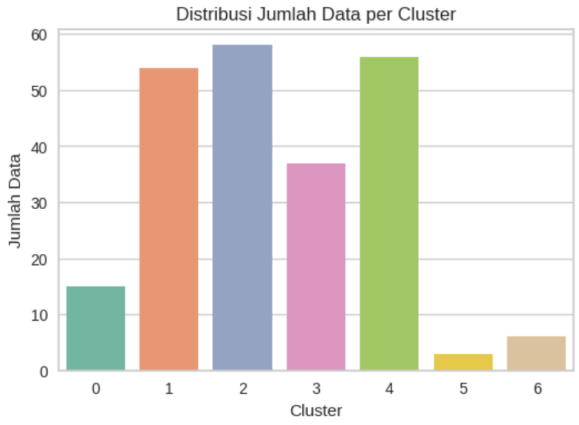

Figure 4. Number of Data per Cluster

The clustering process using the K-Means method yielded seven clusters with varying numbers of members, as depicted in Figure 4. Cluster 0 contains 15 data points, cluster 1 contains 54 data points, cluster 2 has the largest number of data points (58), cluster 3 contains 37 data points, and cluster 4 contains 56 data points. Cluster 5 has the smallest number of data points (3), and cluster 6 contains 6 data points.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

Based on Figure 5, the average of each cluster feature shows different results, which can be used as a basis for a more targeted and appropriate mentoring strategy. Each cluster has different characteristics and interpretations, as shown in Table 1.

Cluster	Umur	Jumlah Anak	Usia Kehamilan (minggu)	Tinggi Badan (cm)	Berat Badan (kg)	Lingkar Lengan Atas (cm)	HB (Hemoglibin)	Tinggi Fundus Uteri (TFU) (cm)	Riwayat Penyakit (jika ada)	Keterpaparan Rokok (Ya/tidak)	Tambah Tablet Darah (Ya/tidak)	PKH (Ya/tidak)	Air Bersih (Ada/Tidak)	Jamban (Ada/Tidak)
0	30.40	0.80	20.73	156.00	61.38	26.32	12.71	23.20	0.0	0.13	0.00	0.20	1.0	1.0
1	32.50	2.33	26.39	156.41	73.22	29.13	11.81	24.20	0.0	0.00	1.00	0.00	1.0	1.0
2	25.29	0.12	30.29	155.97	68.14	28.76	11.95	25.81	0.0	0.00	1.00	0.00	1.0	1.0
3	28.22	0.84	26.68	155.57	65.29	28.30	11.96	24.15	0.0	1.00	1.00	0.05	1.0	1.0
4	26.34	0.55	20.57	156.83	53.28	23.85	11.92	22.70	0.0	0.00	1.00	0.16	1.0	1.0
5	26.00	1.33	30.00	153.00	61.00	25.00	11.10	24.00	0.0	0.00	0.67	0.33	1.0	0.0
6	32.83	2.83	12.50	158.17	65.67	28.33	12.67	24.00	1.0	0.17	1.00	0.50	1.0	1.0

Figure 5. Average Profile of Each Cluster

Table 1. Profile and Interpretation of Characteristics of Each Cluster of Pregnant Women

Cluster	Main Characteristics	Interpretation
0	Age around (30.4 years), body weight (61.4 kg), hemoglobin level (12.7 g/dL), all members do not consume iron supplements.	Group of mothers with normal nutritional conditions and Hb, but need routine nutritional monitoring.
1	Age (32.5 years), weight (73.2 kg), large Upper Arm Circumference (LILA) (29 cm), slightly low hemoglobin level (11.8 g/dL).	Being overweight increases the risk of hypertension/diabetes. Nutritional monitoring is necessary for better weight control, blood sugar monitoring, and balanced nutrition.
2	Young age (25.3 years), gestational age (30.3 weeks), weight (68.1 kg) and LiLA (28.8 cm).	Young mothers in their final trimester are at risk of stunting due to lack of experience and require education on childbirth and newborn care.
3	Pregnant women exposed to cigarette smoke, body weight (65.3 kg), LiLA (28.3), Hb levels (11.96 g/dL).	The risk of fetal growth due to an unhealthy environment requires educational intervention on the dangers of cigarette smoke exposure in families, and fetal monitoring.
4	Lowest weight (53.3 kg), small LiLA (23.85 cm), second trimester gestational age (20.6 weeks), Hb level around (11.9 g/dL).	This cluster group includes those with malnutrition and can be at risk of stunting, which requires nutritional counseling, regular monitoring of body weight and LiLA, administration of iron-boosting tablets, and socialization of the necessary toddler nutrition programs.
5	Lowest hemoglobin level (11.1 g/dL), small LiLA (25 cm), no toilet.	Environmental factors and poor nutrition increase the risk of anemia and infectious diseases. Nutritional monitoring and environmental hygiene are necessary.
6	Age (32 years), more children, early trimester gestational age (12 weeks), Hb levels are quite good (12.7 g/dL), has a history.	Requires intensive supervision by health workers regarding the disease, monitoring of the first trimester of pregnancy, and if necessary, referral to a specialist.

3.4. Prevalence of Stunting Risk Factors per Cluster

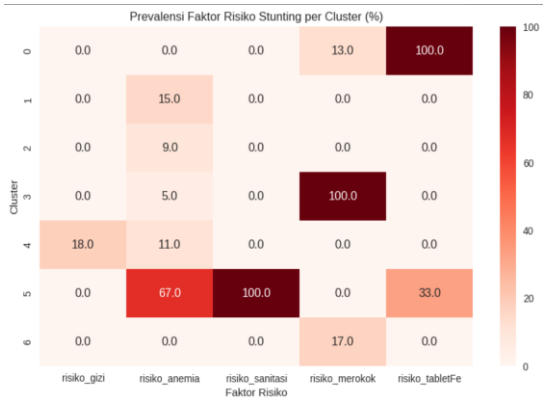

Figure 6. Prevalence of Stunting Risk Factors per Cluster

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

Each cluster has a different distribution of stunting risk factor prevalence as depicted in Figure 6. Stunting risk factors based on data variables are divided into 6, namely nutritional factors, anemia, sanitation (environment), exposure to cigarette smoke, and compliance with consumption of Fe tablets (blood supplements).

a. Cluster 0 has the highest risk of non-compliance with iron tablet consumption (100%), accompanied by 13% exposure to cigarette smoke. This puts pregnant women at risk of anemia and can affect fetal growth if not addressed promptly. Therefore, cadres need to provide education and monitor appropriate supplement consumption.
b. Cluster 1 has a 15% risk of anemia, followed by Cluster 2 at 9%. The risk of anemia in pregnant women is the risk of low birth weight, which is closely related to stunting. Regular provision of iron tablets, Hb levels, and education on balanced nutrition are necessary.
c. Cluster 3 has 100% exposure to cigarette smoke, accompanied by 5% anemia. This very high percentage of cigarette smoke exposure can trigger premature birth and increase the risk of stunting. Prevention can include educating parents and those around them to reduce or even stop smoking, as well as monitoring fetal growth and development.
d. Cluster 4 has a low limb (LILA) that leads to malnutrition (28%), accompanied by anemia (11%), which can put them at risk for chronic energy deficiency (CED), which can affect child growth and development. Solutions include appropriate dietary counseling for pregnant women with CED, weight monitoring, and nutrition programs.
e. The most complex cluster is Cluster 5, with a risk of poor sanitation (100%), 67% anemia, and 33% non-compliance with iron tablet consumption. These nutritional and environmental factors significantly increase the risk of stunting. Improvements to sanitation standards and healthy lifestyles, regular monitoring of iron tablet consumption, and nutritional counseling are needed.
f. Cluster 6 is at risk for stunted fetal growth because 17% of children are exposed to cigarette smoke, despite relatively better nutritional indicators. Preventive measures include family education and reducing exposure to cigarette smoke.

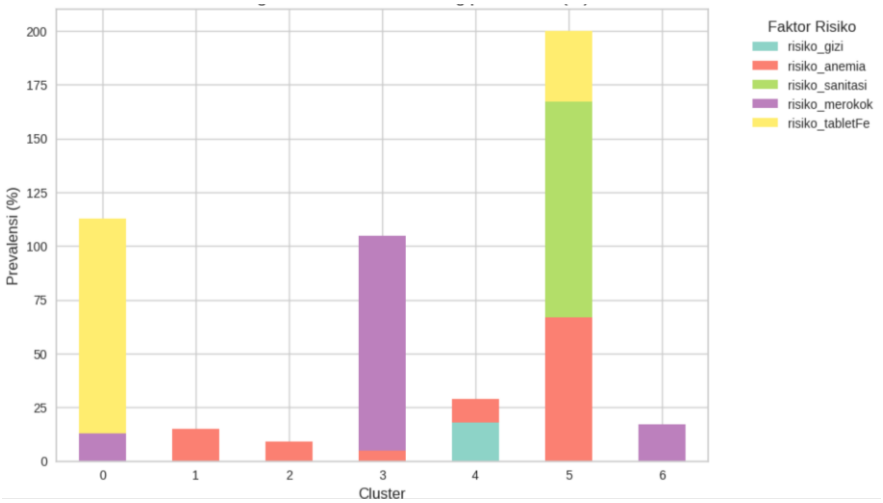

Figure 7. Comparison of Stunting Risk Factors per Cluster (%)

Figure 7 displays the interrelationships between risk factors in each cluster. Analysis of cluster 4 shows an overlap between low nutritional status (LILA) and mild anemia (11%), indicating a link between inadequate nutritional intake and decreased hemoglobin levels. Cluster 5 depicts the most complex combination of risk factors, with a high prevalence of anemia (67%), inadequate sanitation (100%), and non-compliance with iron (Fe) tablet consumption (33%), confirming that nutritional factors, the environment, and compliance with Fe tablet consumption can contribute to stunting risks. Environmental factors in cluster 3, dominated by cigarette smoke exposure (100%) without malnutrition, have an impact on fetal health even when the mother's nutritional status is good. All the patterns depicted in this stacked diagram strengthen the understanding of the combination of risk factors that can increase the risk of stunting.

3.5. Cluster Validation

Quantitatively, cluster validation yielded a Silhouette Score of 0.134, a Davies-Bouldin Index of 1.509, and a Calinski-Harabasz Index of 29.54. The Silhouette score was relatively low, indicating overlap between clusters. This is due to the multidimensional and non-mutually exclusive nature of social and clinical data for pregnant women, where individuals can have more than one risk factor simultaneously. The combination of clinical, non-compliance, and environmental variables resulted in unclear boundaries between clusters, thus affecting the level of separation between groups.

Nevertheless, the optimal number of clusters was determined based on the results of the Elbow Method, which demonstrated pattern stability at k=7. The Davies-Bouldin and Calinski-Harabasz values still indicated a meaningful

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

cluster structure. The characteristic patterns between cluster profiles were consistent, confirming that the clustering results remain exploratory and relevant as a basis for identifying stunting risk groups in pregnant women.

3.6. Cluster Difference Test

The Kruskal-Wallis test showed that most variables differed between clusters ($p<0.05$), except for height, which did not differ ($p=0.523$). This confirms that clustering can separate different characteristics of pregnant women, particularly demographic, clinical, and environmental factors associated with stunting.

3.7. Principal Component Analysis (PCA)

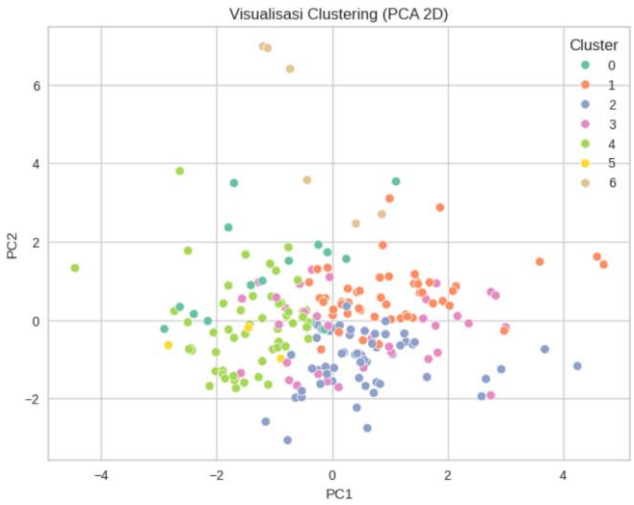

Figure 8. Principal Component Analysis

Figure 8 shows the PCA results visualized in a two-dimensional scatterplot (PC1 and PC2) with coloring based on the K-Means clustering results. PCA visualizes the split pattern between clusters, although some overlap remains. This overlap occurs because risk factors in pregnant women are multifactorial and intersecting, allowing individuals to be associated with more than one cluster.

3.8. Discussions

The application of the clustering method (K-Means) resulted in seven groups of pregnant women (clusters 0-6) with varying characteristics: cluster 0 (normal profile), cluster 1 (high weight), cluster 2 (young mothers in their final trimester), cluster 3 (exposure to cigarette smoke), cluster 4 (malnutrition), cluster 5 (anemia and environmental problems), and finally, cluster 6 (medical history). This grouping facilitates the identification of groups of pregnant women with appropriate profiles, enabling interventions to prevent stunting.

This aligns with research by [13] who used the K-Means method to identify health data on stunting risk, with clusters proven effective in grouping data based on appropriate risk. An imbalance in the number of cluster members can impact the results, particularly in clusters 5 and 6, which have small data sets, making them more susceptible to extreme values and input errors. Nevertheless, these two clusters represent the most complex risk groups and remain programmatically important and relevant as priority intervention targets.

The low Silhouette value and overlapping PCA visualizations between clusters indicate that the data for pregnant women is heterogeneous and multifactorial. Data preprocessing, such as normalization and winsorizing, contributed to reducing distortion due to scale differences and the presence of outliers.

This analysis has other limitations: (1) K-Means only identifies spherical and homogeneous clusters, potentially underdetecting data patterns. (2) There are clusters with very few members, and there may be data input errors that must be addressed through data verification and further data cleaning. (3) The data used was collected at a single point in time (cross-sectional), so clustering was based on similarity, not causality. (4) These findings have not been tested with other data sets, so their generalizability is still limited. Future research is recommended to optimize data cleaning, verify outliers, and compare them with other methods such as Hierarchical Clustering, K-Medoids, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to achieve more consistent results. It is recommended to use longitudinal datasets so that patterns of change in pregnant women can be analyzed better.

4. CONCLUSION

This study successfully applied the K-Means algorithm to group pregnant women in Central Semarang District into seven clusters based on clinical, demographic, and environmental factors. Validation results showed a Silhouette Score of 0.134, a Davies-Bouldin Index of 1.509, and a Calinski-Harabasz of 29.54, indicating that the cluster structure was formed exploratively and was able to represent risk variations despite natural overlap due to heterogeneity in social

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1628−1636
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8415

and clinical data. The K-Means method proved effective in uncovering risk patterns that cannot be directly observed through conventional approaches, thus making a significant contribution to the application of data mining in maternal and child health. The results can be used for further, more targeted actions by cadres in designing stunting prevention assistance programs tailored to the needs of each group of pregnant women, such as nutritional counseling, environmental education, intensive clinical monitoring, and medical referrals. However, this study still has several limitations, including sensitivity to outliers, imbalance in cluster size, and the cross-sectional nature of the data, which prevents direct causality. Further research is recommended to employ more appropriate data preprocessing, outlier verification, and comparisons with other methods such as K-Medoids or DBSCAN to obtain more stable and generalized results. Furthermore, the use of longitudinal data is necessary to analyze dynamic risk changes and improve the effectiveness of predictive models.

REFERENCES

[1] S. K. Pranindita, A. Yuniastuti, and S. R. Rahayu, "Hubungan Faktor Maternal Ibu dengan Kejadian Stunting pada Balita Usia 24-59 Bulan di Kabupaten Grobogan," *Indonesian Journal of Public Health and Nutrition*, vol. 5, no. 1, 2025, doi: https://doi.org/10.15294/ijphn.v5i1.28999.

[2] World Health Organization, *Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division*. Geneva: World Health Organization, 2023. Accessed: Aug. 09, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240068759

[3] Kementerian Kesehatan Republik Indonesia, *Laporan Kinerja Kementerian Kesehatan Republik Indonesia Tahun 2023*. Jakarta: Kementerian Kesehatan Republik Indonesia, 2024.

[4] L. Sulistianingrum, "Karakteristik dan tingkat pengetahuan ibu hamil dengan kejadian kurang energi kronis (KEK)," *Midwifery J. MJ*, vol. 3, no. 4, 2023, doi: 10.33024/mj.v3i4.13379.

[5] N. K. Pane, U. H. Almadany, and E. Sujoko, "Status Gizi Ibu Hamil sebagai Prediktor Kejadian Stunting pada Anak Usia 24–59 Bulan di Kecamatan Padangsidimpuan Selatan," *PubHealth J. Kesehat. Masy.*, vol. 4, no. 1, pp. 46–53, Jul. 2025, doi: 10.56211/pubhealth.v4i1.1026.

[6] W. Wulandari and W. D. Pangesti, "Prevalensi Preeklamsi dengan Komplikasi di Rumah Sakit Rujukan Kabupaten Banyumas Tahun 2017-2020," *J. Kebidanan Harapan Ibu Pekalongan*, vol. 9, no. 1, pp. 1–15, Feb. 2022, doi: 10.37402/jurbidhip.vol9.iss1.168.

[7] P. Hanum, S. Sumiaty, S. Sumiati, and S. Suryani, "Hubungan Kadar HB, Lila dan Berat Badan Ibu Saat Hamil Berisiko dengan Kejadian Stunting pada Anak Usia 1-3 Tahun," *MAHESA Malahayati Health Stud. J.*, vol. 4, no. 2, pp. 699–708, Feb. 2024, doi: 10.33024/mahesa.v4i2.13230.

[8] D. Sartika, F. Elfaladonna, and A. Octarina, "Kombinasi hybrid K-means untuk klasterisasi multivariat dalam analisis stunting," *Jurnal Jaringan Sistem Informasi Robotik (JSR)*, vol. 9, no. 1, pp. 64-72, 2025.

[9] J. Maulindar and E. P. Yudha, "Pengembangan Klastering Untuk Penanganan Ibu Hamil Menggunakan K-Means," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB) 2023*, Surakarta, Indonesia: Universitas Duta Bangsa Surakarta, Jul. 2023.

[10] B. P. Wongso, M. E. Johan, and M. I. Fianty, "Empowering Pregnancy Risk Assessment: A Web-Based Classification Framework with K-Means Clustering Enhanced Models," *J. Inf. Syst. Inform.*, vol. 5, no. 4, pp. 1221–1239, Nov. 2023, doi: 10.51519/journalisi.v5i4.568.

[11] R. Ishak, "Optimasi K-Means pada Clustering Penyakit Ibu Hamil Menggunakan Random Forest Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest," *Jambura J. Electr. Electron. Eng.*, vol. 7, no. 1, Jan. 2025.

[12] I. Indra, N. Nur, Muh. Iqram, and N. Inayah, "Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting," ISI, vol. 8, no. 2, p. 356, Nov. 2023, doi: 10.35314/isi.v8i2.3612.

[13] M. H. M. Rohman *et al.*, "Clustering Analysis of Stunting Risk Factors Using K-Means and Principal Component Analysis: A Case Study in Indonesian Regency," *Sinkron*, vol. 9, no. 1, pp. 65–77, 2025, doi: 10.33395/sinkron.v9i1.14311.

[14] N. Alharbe, M. A. Rakrouki, and A. Aljohani, "A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm," *Processes*, vol. 10, no. 6, p. 1199, Jun. 2022, doi: 10.3390/pr10061199.

[15] A. Aljohani, "Optimizing Patient Stratification in Healthcare: A Comparative Analysis of Clustering Algorithms for EHR Data," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 173, Jul. 2024, doi: 10.1007/s44196-024-00568-8.

[16] I. T. Utami, F. Suryaningrum, and D. Ispriyanti, "K-means cluster count optimization with silhouette index validation and Davies Bouldin index (case study: coverage of pregnant women, childbirth, and postpartum health services in Indonesia in 2020)," *BAREKENG J. Ilmu Mat. Dan Terap.*, vol. 17, no. 2, pp. 0707–0716, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0707-0716.

[17] R. D. Syaputra and A. Solichin, "Pregnancy Risk Level Classification Using The CRISP-DM Method," *J. Ris. Inform.*, vol. 5, no. 1, pp. 537–548, Dec. 2022, doi: 10.34288/jri.v5i1.487.

[18] H. Lv, L. Hu, M. Jiang, X. Liu, and Z. He, "Interpretable Clustering Ensemble," Jun. 06, 2025, *arXiv*: arXiv:2506.05877. doi: 10.48550/arXiv.2506.05877.

[19] S. Febriyanti and J. Nugraha, "Application of K-Medoids Clustering to Increase the 2020 Family Planning Program in Sleman Regency," *Enthusiastic Int. J. Appl. Stat. Data Sci.*, pp. 10–18, Apr. 2022, doi: 10.20885/enthusiastic.vol2.iss1.art2.

[20] D. T. Setiyawan, B. Berlilana, and A. S. Barkah, "Comparative Analysis of DBSCAN, OPTICS, and Agglomerative Clustering Methods for Identifying Disease Distribution Patterns in Banjarnegara Community Health Centers," J. Tek. Inform. (JUTIF), vol. 6, no. 3, pp. 1229–1240, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4577.

[21] H. P. Hadi *et al.*, "Mengungkap Heterogenitas Stunting pada Anak: Pendekatan Machine Learning untuk Intervensi yang Tepat Sasaran di Sambas, Indonesia," *Jatekom J. Apl. Teknol. Dan Komputasi*, vol. 1, no. 2, pp. 94–108, Jun. 2025.