International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Identifying High-Risk Stunting Districts through Data Mining Clustering for Targeted Policy Making

Nazwa Akilla Zahra[1], Yazid Hilmi Allamsyah[2], Abiyyu Rasyiq Muhadzzib[3], Mario Mora Siregar[4], Detty Purnamasari[5], Milda Safrila Oktiana[6]
[1,2,4]Informatics, Universitas Gunadarma, Depok, West Java, Indonesia-16424
[3,6]Information System, Universitas Gunadarma, Depok, West Java, Indonesia-16424
[5]Information Technology, Universitas Gunadarma, Depok, West Java, Indonesia-16424

Abstract*—Stunting is a major public health issue that impairs children's physical growth and cognitive development, ultimately affecting human capital quality and national productivity. According to the 2023 Indonesia Health Survey, the prevalence of stunting in Indonesia reached 21.5%. Contributing factors include poor sanitation, limited access to clean water, inadequate knowledge of child nutrition, and weak cross-sectoral coordination. These challenges highlight the need for data-driven intervention strategies to objectively identify high-risk areas. This study aims to develop a clustering model to classify 430 districts and municipalities in Indonesia into three risk categories—high, medium, and low—as a basis for targeted policy interventions. The research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The dataset was obtained from the Indonesian Ministry of Villages, Development of Disadvantaged Regions, and Transmigration (2023), comprising 75,265 village-level records and 67 variables related to demographics, health, sanitation, and governance. Three clustering algorithms were evaluated: K-Means, DBSCAN, and Fuzzy C-Means. Model performance was assessed using the Silhouette Score and Davies–Bouldin Index. The best-performing model achieved a Silhouette Score of 0.52 and a Davies–Bouldin Index of 1.11, indicating a satisfactory clustering structure with distinct intervention profiles across clusters. The model classified the 430 districts and municipalities into three clusters: (i) High Stunting Risk (9 districts/municipalities; 2.09%), (ii) Medium Stunting Risk (45 districts/municipalities; 10.47%), and (iii) Low Stunting Risk (375 districts/municipalities; 87.20%). The proposed clustering model provides an evidence-based framework for supporting targeted policy interventions and enabling more efficient allocation of public resources according to the characteristics of each risk cluster.*

***Keywords*—** *Child Nutrition, Clustering, Data Mining, K-Means, Public Policy, Stunting, Under-Five Children*.

I. INTRODUCTION

Stunting remains one of the most critical public health challenges in developing countries because it impairs children's physical growth, cognitive development, educational attainment, and future economic productivity [1]. According to the 2023 Indonesia Health Survey (SKI), the national stunting prevalence reached 21.5%, exceeding the 20% threshold established by the World Health Organization (WHO) [2]. To address this challenge, the Indonesian government has set a target of reducing the national stunting prevalence to 14.2% by 2029 [3], highlighting the need for evidence-based policies capable of identifying priority regions for targeted interventions.

The causes of stunting are multifactorial and extend beyond inadequate nutritional intake. Previous studies have shown that poor sanitation, limited access to clean water [4], low parental knowledge of child nutrition, and socioeconomic disparities significantly contribute to the persistence of stunting [5]. Moreover, weak coordination among government sectors and the diverse characteristics of regions across Indonesia reduce the effectiveness of uniform intervention strategies [6]. Consequently, identifying areas with similar risk characteristics is essential for designing targeted and efficient policy interventions.

From a technical perspective, the large volume and heterogeneity of regional data present significant challenges for conventional statistical analysis. Government datasets typically contain thousands of observations with numerous demographic, health, sanitation, and governance indicators, making manual analysis inefficient and potentially inaccurate [4]. As a result, decision-makers often face difficulties in objectively prioritizing regions, which may lead to inefficient allocation of resources and suboptimal intervention outcomes.

Data mining techniques, particularly clustering methods, provide an effective approach for discovering hidden patterns in multidimensional datasets without requiring predefined class labels. Algorithms such as K-Means [7], DBSCAN [8], and Fuzzy C-Means [9] have been widely applied to group objects with similar characteristics across various domains, including public health. By clustering districts and municipalities according to their stunting-related characteristics, policymakers can better identify high-, medium-, and low-risk regions and formulate interventions tailored to the needs of each cluster.

Based on these considerations, this study proposes a data mining-based clustering framework to classify 430 districts and municipalities in Indonesia according to their stunting risk profiles using village-level data collected by the Ministry of Villages, Development of Disadvantaged Regions, and Transmigration. Three clustering algorithms—K-Means, DBSCAN, and Fuzzy C-Means—are evaluated using the Silhouette Score and Davies–Bouldin Index to determine the most appropriate model. The resulting clusters are expected to provide an objective, data-driven foundation for supporting targeted policy interventions and improving the effectiveness of stunting reduction programs in Indonesia.

Nazwa Akilla Zahra, Yazid Hilmi Allamsyah, Abiyyu Rasyiq Muhadzzib, Mario Mora Siregar, Detty Purnamasari, and Milda Safrila Oktiana, "Identifying High-Risk Stunting Districts through Data Mining Clustering for Targeted Policy Making," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 3, pp. 42-46, 2026.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

II. Literature Study

A. K-Means

K-Means is one of the most widely used unsupervised machine learning algorithms for partitioning data into a predefined number of clusters based on similarity. The algorithm iteratively assigns observations to the nearest centroid and updates the centroid positions until convergence is achieved, aiming to minimize the within-cluster variance [7]. Owing to its computational efficiency and scalability, K-Means has been extensively applied to large multidimensional datasets across various domains, including public health. Previous studies have demonstrated its effectiveness in identifying regional patterns of stunting risk, enabling the classification of villages into distinct risk categories that support evidence-based intervention planning [10].

B. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised clustering algorithm that groups data points based on their density distribution [8]. The algorithm relies on two key parameters: the neighborhood radius (Eps) and the minimum number of neighboring points (MinPts) required to form a dense region. Unlike centroid-based clustering methods, DBSCAN can identify clusters with arbitrary shapes, automatically detect outliers, and does not require the number of clusters to be specified in advance. These characteristics make DBSCAN particularly suitable for datasets with heterogeneous distributions, such as regional stunting indicators. Previous studies have demonstrated its effectiveness in discovering complex data structures, although its performance is highly dependent on the appropriate selection of Eps and MinPts [8].

C. Fuzzy C-Means

Fuzzy C-Means (FCM) is an unsupervised clustering algorithm that allows each data point to belong to multiple clusters by assigning a membership degree ranging from 0 to 1 [9]. Unlike hard clustering methods, FCM effectively handles overlapping cluster boundaries, making it well suited for complex public health datasets in which multiple indicators may simultaneously influence regional characteristics. Previous studies have demonstrated the effectiveness of FCM in adaptively grouping regions according to health-related indicators, providing more flexible cluster representations than conventional partitioning methods [9].

D. CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely adopted methodology that provides a structured and systematic framework for executing data mining projects [11]. The methodology consists of six iterative phases that guide the entire analytical process, from problem understanding to model deployment, ensuring that the developed solution aligns with the research objectives. The CRISP-DM workflow is illustrated in Figure 1.

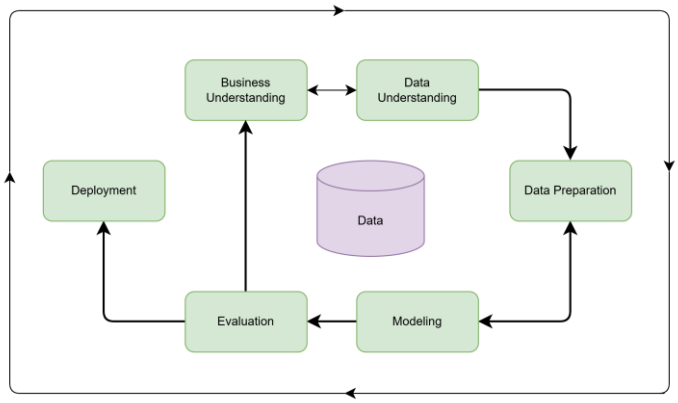

Fig. 1. CRISP-DM

III. Methodology

A. Business Understanding

The objective of this study is to classify 430 districts and municipalities in Indonesia into three stunting risk categories, namely high, medium, and low risk. The resulting clusters are intended to support policymakers in prioritizing intervention programs and allocating public resources more effectively. The success of the proposed approach is determined by its ability to produce meaningful and well-separated clusters that accurately represent regional characteristics related to stunting.

B. Data Understanding

The study utilizes the publicly available "Number of Beneficiaries of Stunting Prevention Services 2023" dataset obtained from the Satu Data Indonesia portal [12]. The dataset contains 75,265 village-level records with 67 variables representing demographic characteristics, healthcare service coverage, sanitation conditions, and village governance indicators associated with stunting prevention.

An initial exploratory analysis was conducted to examine data distributions, identify missing values, detect potential outliers, and understand the characteristics of each variable before preprocessing. This step ensured that the dataset was suitable for subsequent clustering analysis.

C. Data Preparation

Several preprocessing procedures were performed to improve data quality prior to model development. Missing values were replaced with zero because unavailable entries represented the absence of reported beneficiaries rather than unknown observations. Categorical variables were transformed into numerical representations where necessary to facilitate clustering.

To better represent regional conditions, several derived indicators were generated, including the stunting ratio, village governance score, and composite stunting risk index. Since the original dataset was collected at the village level, all variables were aggregated into district- and municipality-level statistics, resulting in 430 observations. Finally, all numerical variables were standardized to remove scale differences and ensure equal contribution during the clustering process.

Nazwa Akilla Zahra, Yazid Hilmi Allamsyah, Abiyyu Rasyiq Muhadzzib, Mario Mora Siregar, Detty Purnamasari, and Milda Safrila Oktiana, "Identifying High-Risk Stunting Districts through Data Mining Clustering for Targeted Policy Making," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 3, pp. 42-46, 2026.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

D. Modeling

Three clustering algorithms, namely K-Means, DBSCAN, and Fuzzy C-Means, were implemented and compared to identify the most appropriate model for grouping districts according to their stunting risk profiles.

For K-Means and Fuzzy C-Means, the optimal number of clusters was determined using the Elbow method, resulting in three clusters representing high-, medium-, and low-risk regions. For DBSCAN, several combinations of the neighborhood radius (Eps) and minimum number of neighboring points (MinPts) were evaluated to identify the parameter configuration that produced the most meaningful clusters. Each clustering model was applied to the standardized dataset using the same input variables to enable a fair comparison of their performance.

E. Evaluation

The clustering performance was evaluated using two internal validation metrics: the Silhouette Score and the Davies–Bouldin Index (DBI).

The Silhouette Score measures the similarity of an observation to its assigned cluster relative to neighboring clusters, where higher values indicate better cluster cohesion and separation.

$$s(i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

The Davies–Bouldin Index evaluates the average similarity between clusters based on intra-cluster dispersion and inter-cluster separation. Lower DBI values indicate better clustering quality.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq 1} \left(\frac{s_i + s_j}{d_{ij}} \right)$$

The clustering algorithm achieving the highest Silhouette Score and the lowest Davies–Bouldin Index was selected as the final model for subsequent analysis.

F. Deployment

The selected clustering model was implemented as an interactive decision-support dashboard using Streamlit. The dashboard visualizes the spatial distribution of stunting risk across districts and municipalities, provides cluster summaries, and presents policy recommendations based on the identified regional characteristics. In addition, the application includes data exploration and result export functionalities, enabling policymakers and other stakeholders to access and utilize the clustering results efficiently.

IV. Result and Discussion

A. Clustering Performance Evaluation

The clustering performance of K-Means, DBSCAN, and Fuzzy C-Means was evaluated using two internal validation metrics, namely the Silhouette Score and the Davies–Bouldin Index (DBI). The comparison of the Silhouette Score for the three clustering algorithms is presented in Figure 2.

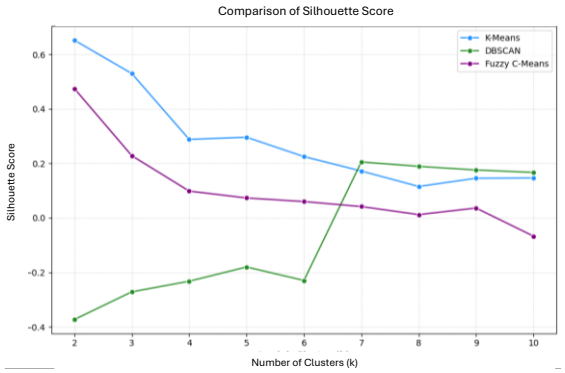

Fig. 2. Comparison of Silhouette Score

As shown in Fig. 2, K-Means achieved the highest Silhouette Score (0.52), indicating better cluster cohesion and separation than DBSCAN (0.20) and Fuzzy C-Means (0.22). The relatively higher score demonstrates that districts assigned to the same cluster exhibit greater similarity while maintaining clearer separation from neighboring clusters. In contrast, the lower scores obtained by DBSCAN and Fuzzy C-Means suggest that their resulting clusters are less compact and contain greater overlap among observations. The comparison of the Davies–Bouldin Index is presented in Figure 3.

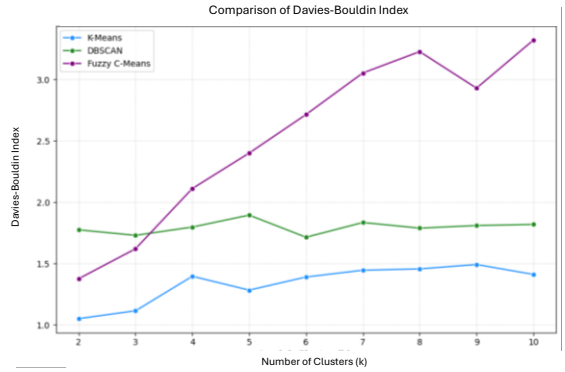

Fig. 3. Comparison of Davies–Bouldin Index

As illustrated in Fig. 3, K-Means produced the lowest Davies–Bouldin Index (1.11), followed by Fuzzy C-Means (1.61) and DBSCAN (1.83). Since lower DBI values indicate better separation between clusters, these results further confirm the superiority of K-Means for the analyzed dataset. The quantitative comparison of all evaluated clustering algorithms is summarized in Table I.

Table I. Clustering Performance Evaluation

Algorithm	Number of Cluster	Number of Districts per Cluster	Silhouette Score	Davies-Bouldin Index
K-Means	3	High: 9 Medium: 45 Low: 376	0.52	1.11
DBSCAN	2	High: 200 Low: 230	0.20	1.83
Fuzzy C-Means	3	High: 41 Medium: 145 Low: 244	0.22	1.61

Nazwa Akilla Zahra, Yazid Hilmi Allamsyah, Abiyyu Rasyiq Muhadzzib, Mario Mora Siregar, Detty Purnamasari, and Milda Safrila Oktiana, "Identifying High-Risk Stunting Districts through Data Mining Clustering for Targeted Policy Making," *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 3, pp. 42-46, 2026.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Based on Table I, K-Means generated three clusters consisting of 9 high-risk, 45 medium-risk, and 376 low-risk districts and municipalities. In comparison, DBSCAN produced only two clusters, while Fuzzy C-Means generated three clusters with a considerably different distribution of regions. The superior Silhouette Score and lower DBI indicate that K-Means provides a more meaningful representation of the regional characteristics associated with stunting.

The better performance of K-Means is likely attributable to the characteristics of the dataset. Following data aggregation, feature engineering, and standardization, the regional indicators exhibit relatively homogeneous distributions that are more effectively partitioned using centroid-based clustering. Conversely, DBSCAN is generally more suitable for datasets containing irregular cluster shapes and varying densities, whereas the advantage of Fuzzy C-Means in handling overlapping cluster boundaries becomes less significant when observations are already well separated.

B. Cluster Characteristics

The characteristics of the three clusters generated by the selected K-Means model are summarized in Table II.

Table III. Cluster Characteristics

Cluster Profile Differences	Cluster Proportion	Key Stunting Characteristics
High-Risk Cluster	2.09%	• Child Malnutrition: 35.89% • Composite Stunting Risk Index: 5.22% • Maternal with Chronic Energy Deficiency (CED): 42.22%
Medium-Risk Cluster	10.47%	• Child Malnutrition: 4.12% • Composite Stunting Risk Index: 4.38% • Maternal with Chronic Energy Deficiency (CED): 28.05%
Low-Risk Cluster	87.44%	• Child Malnutrition: 1.38% • Composite Stunting Risk Index: 1.83% • Maternal with Chronic Energy Deficiency (CED): 16.96%

The High-Risk Cluster consists of nine districts and municipalities (2.09%) characterized by the highest levels of child malnutrition (35.89%), maternal chronic energy deficiency (42.22%), and the composite stunting risk index (5.22%). These findings indicate that nutritional deficiencies and maternal health remain the dominant factors contributing to stunting in these regions. Therefore, policy interventions should prioritize nutritional assistance, maternal healthcare services, and improvements in primary healthcare accessibility.

The Medium-Risk Cluster includes 45 districts and municipalities (10.47%) with moderate values across the major indicators. Although these regions do not exhibit the severe conditions observed in the high-risk cluster, preventive interventions remain necessary. Strengthening maternal and child healthcare services, expanding nutrition education programs, and improving early detection mechanisms are expected to reduce the likelihood of transition into higher-risk categories.

The Low-Risk Cluster, comprising 376 districts and municipalities (87.44%), demonstrates relatively favorable conditions across all evaluated indicators, including lower levels of child malnutrition (1.38%), maternal chronic energy deficiency (16.96%), and the composite risk index (1.83%). Policy efforts for this cluster should focus on maintaining existing achievements while disseminating successful intervention strategies to regions with higher stunting risk.

Overall, the identified clusters reveal clear differences in regional health profiles, indicating that the proposed clustering model effectively captures the heterogeneity of stunting-related conditions across Indonesian districts and municipalities.

C. Spatial Distribution of Stunting Risk

The spatial distribution of the identified stunting risk clusters is illustrated in Figure 4.

Fig. 4. Spatial Distribution of Stunting Risk

Figure 4 shows that most districts and municipalities belong to the Low-Risk Cluster, while only a limited number of regions are classified as high risk. The resulting spatial visualization provides an intuitive overview of regional disparities in stunting risk and facilitates the identification of priority intervention areas. Unlike conventional tabular reports, the clustering map enables policymakers to rapidly identify geographical patterns and allocate resources according to regional needs. Such an approach supports more efficient policy implementation by directing nutritional programs, healthcare services, and infrastructure improvements toward districts with the greatest need.

V. Conclusion And Recommendations

This study proposed a data mining-based clustering framework based on the CRISP-DM methodology to classify 430 districts and municipalities in Indonesia according to their stunting risk profiles. Three clustering algorithms—K-Means, DBSCAN, and Fuzzy C-Means—were evaluated using the Silhouette Score and Davies–Bouldin Index. Among the evaluated methods, K-Means achieved the best clustering performance, with a Silhouette Score of 0.52 and a Davies–Bouldin Index of 1.11, indicating compact and well-separated clusters.

The selected K-Means model successfully classified the study areas into three risk categories. The resulting clusters provide an objective basis for identifying priority intervention areas and demonstrate that regional stunting risk is associated with interconnected indicators, including child nutrition, maternal health, sanitation, and governance. These findings

Nazwa Akilla Zahra, Yazid Hilmi Allamsyah, Abiyyu Rasyiq Muhadzzib, Mario Mora Siregar, Detty Purnamasari, and Milda Safrila Oktiana, “Identifying High-Risk Stunting Districts through Data Mining Clustering for Targeted Policy Making,” *International Research Journal of Advanced Engineering and Science*, Volume 11, Issue 3, pp. 42-46, 2026.

highlight the potential of data mining techniques to support evidence-based policymaking and targeted intervention resource allocation.

This study has several limitations. The analysis relies on secondary village-level data aggregated at the district level, which may not fully capture variations within individual districts. In addition, the clustering approach identifies groups of regions with similar characteristics but does not establish causal relationships among the analyzed indicators.

Future research should incorporate longitudinal datasets to analyze temporal changes in stunting risk, integrate additional socioeconomic and environmental variables, and investigate spatial or hybrid clustering approaches to improve clustering performance and support more comprehensive policy recommendations.

ACKNOWLEDGMENT

This research was supported by the Universitas Gunadarma-Indonesia, Research Center of Universitas Gunadarma, and supported by HPC/DGX team of Universitas Gunadarma (UG-Artificial Intelligence-Center of Excellence/UG-AI-CoE).

REFERENCES

[1] N. O. Nirmalasari, "Stunting pada anak: Penyebab dan faktor risiko stunting di Indonesia," Qawwam: Journal for Gender Mainstreaming, vol. 14, no. 1, pp. 19–28, 2020, doi: 10.20414/Qawwam.v14i1.2372.

[2] S. L. Munira and D. Puspasari, "Survei Kesehatan Indonesia (SKI) dalam Angka," Kementerian Kesehatan, pp. 390–403, 2023.

[3] Strategi Nasional Percepatan Pencegahan dan Penurunan Stunting 2025–2029. [Online]. Available: https://stunting.go.id/strategi-nasional-percepatan-pencegahan-dan-penurunan-stunting-2025-2029/.

[4] Y. Haskas, "Gambaran stunting di Indonesia: Literatur review," Jurnal Ilmiah Kesehatan Diagnosis, vol. 15, no. 2, 2020.

[5] Y. B. Prasetyo, P. Permatasari, and H. D. Susanti, "The effect of mothers' nutritional education and knowledge on children's nutritional status: A systematic review," International Journal of Child Care and Education Policy, vol. 17, no. 11, 2023, doi: 10.1186/s40723-023-00114-7.

[6] J. Sevilla, A. Nugroho, and A. Turymshayeva, "The effectiveness of accelerating stunting reduction policy," Journal of Sustainable Development and Regulatory Issues, vol. 2, no. 2, May 2024, doi: 10.53955/jsderi.v2i2.31.

[7] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," Electronics, vol. 9, no. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[8] R. Adha, N. Nurhaliza, U. Soleha, and Mustakim, "Perbandingan algoritma DBSCAN dan K-Means clustering untuk pengelompokan kasus Covid-19 di dunia," Jurnal Sains, Teknologi dan Industri, vol. 18, no. 2, pp. 206–211, Jun. 2021.

[9] G. S. Nugraha and B. A. Riyandari, "Implementasi Fuzzy C-Means untuk pengelompokan daerah berdasarkan indikator kesehatan," Jurnal Teknologi Informasi, vol. 4, no. 1, pp. 56–62, 2020.

[10] D. M. Cytry, S. Defit, and G. W. Nurcahyo, "Penerapan metode K-Means dalam klasterisasi status desa terhadap keluarga berisiko stunting," Jurnal KomtekInfo, vol. 10, no. 3, pp. 122–127, 2023.

[11] F. N. Dhewayani, D. Amelia, and D. N. Alifah, "Implementasi K-Means clustering untuk pengelompokkan daerah rawan bencana kebakaran menggunakan model CRISP-DM," Jurnal Teknologi dan Informasi, vol. 12, no. 1, Mar. 2022, doi: 10.34010/jati.v12i1.

[12] Satu Data Indonesia, "Jumlah penerima layanan pencegahan stunting tahun 2023." [Online]. Available: https://data.go.id/dataset/dataset/jumlah-penerima-layanan-pencegahan-stunting-tahun-2023/. Accessed: Jul. 27, 2025.