Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 2986~2999
e-ISSN: 2548-6861

Household Clustering in West Java Based on Stunting Risk Factors Using K-Modes and K-Prototypes Algorithms

Muhammad Yusran[1]*, Siti Nuradilla[2], Mega Ramatika Putri[3], Anwar Fitrianto[4], Erfiani[5], Rachmat Bintang Yudhianto[6]
*Statistics and Data Science, IPB University, Bogor, Indonesia
muhammadyusran@apps.ipb.ac.id[1], sitinuradilla@apps.ipb.ac.id[2], megaramatikaputri@apps.ipb.ac.id[3], anwarstat@gmail.com[4], erfiani@apps.ipb.ac.id[5], ydth_2000_rachmat@apps.ipb.ac.id[6]

Article Info

Article history:

Received 2025-10-15
Revised 2025-11-04
Accepted 2025-11-08

Keyword:

Stunting,
Clustering,
K-Modes,
K-Prototypes.

ABSTRACT

Stunting remains one of Indonesia's most persistent public health challenges, with West Java contributing the highest number of cases due to its large population and regional disparities in household welfare. Identifying household groups vulnerable to stunting is essential for designing targeted interventions that integrate nutrition, sanitation, and socio-economic development. This study introduces a data-driven clustering framework using the K-Modes and K-Prototypes algorithms to classify 22,161 households in West Java based on 26 indicators from the March 2024 National Socioeconomic Survey (SUSENAS), encompassing food security, sanitation, drinking water access, economic conditions, social assistance, and demographics. The K-Modes algorithm was applied to categorical data, while K-Prototypes integrated numerical and categorical variables, with parameter optimization performed using a grid search and the Elbow method. Clustering performance was evaluated through the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index, followed by a bootstrapped stability analysis employing the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Results show that K-Prototypes outperformed K-Modes, yielding a higher Silhouette Score (0.6681 compared to 0.2922), higher CH Index (13,890.6 compared to 3,976.1), and lower DBI (0.4607 compared to 1.5274), indicating superior compactness and separation. Stability testing confirmed strong robustness, with mean ARI = 0.959 and mean NMI = 0.932 across 50 bootstrap replications. The optimal five-cluster structure identified distinct socioeconomic groups, with the highest stunting risk found among households with low income, limited housing space, inadequate sanitation, and more children under five. The findings highlight the effectiveness of K-Prototypes in modeling mixed-type data and support the design of evidence-based, regionally adaptive stunting reduction strategies aligned with Presidential Regulation No. 72/2021 on the Acceleration of Stunting Reduction.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Stunting, a condition of impaired growth among children under five, remains a serious nutritional issue in Indonesia. According to the 2024 Indonesian Nutritional Status Survey (SSGI), the national stunting prevalence declined to 19.8%. However, the Indonesian government must exert greater efforts to achieve the 2025 target of 18.8%. Indonesian Stunting Reduction Acceleration Task Force reported that West Java is the province with the highest contribution to national stunting cases [1]. This aligns with the fact that West Java is the most populous province in Indonesia. The large population size and the high heterogeneity of household characteristics in West Java call for targeted stunting reduction strategies that can address the complexity of associated risk factors.

The risk factors contributing to stunting are multidimensional, encompassing nutritional, socio-economic, and household environmental aspects. Factors such as food security, sanitation, access to clean drinking water and hygiene, economic conditions, and demographic characteristics play a crucial role in determining household vulnerability to stunting [2], [3]. Through Presidential Regulation No. 72 of 2021 concerning the Acceleration of Stunting Reduction, the Indonesian government established five pillars of stunting reduction acceleration, which emphasize not only nutritional interventions but also socio-economic and environmental improvements at the household level [4]. A major challenge in implementing these policies lies in the identification of households at risk of stunting. In fact, identifying households with similar risk characteristics is essential for the stunting reduction task force to deliver well-targeted interventions. However, each household exhibits diverse combinations of risk factors. Therefore, a data-driven approach is needed to classify households based on similarities in their stunting-related risk profiles.

Clustering is an unsupervised machine learning method that groups observations based on their similarities. Previous studies have employed clustering techniques to classify stunting-prone areas at the district or regency level, with most relying on the K-Means method. These include: (1) clustering regencies/municipalities in West Java based on child stunting risk factors [5]; (2) spatial cluster analysis of stunting incidence in North Sumatra based on environmental factors [6]; and (3) clustering stunting risk factors using K-Means and Principal Component Analysis (PCA) [7]. While K-Means provides an initial overview of stunting-prone area groupings, it performs optimally only on numerical data through the use of Euclidean distance [8]. This method is less suitable for household-level data, which often contain categorical variables. On the other hand, Hierarchical Clustering methods such as DBSCAN are difficult to implement on categorical data because they rely on density-based distance definitions, while the Gaussian Mixture Model (GMM) is unsuitable as it assumes continuous distributions for all variables [9]. Applying K-Means and Hierarchical Clustering to such data requires transforming categorical variables, which increases dimensionality and reduces interpretability.

In response to the abundance of field data containing categorical variables, [10] developed the K-Modes algorithm, specifically designed for categorical data by employing simple matching dissimilarity and using the mode as the cluster center. This makes K-Modes more efficient, as it eliminates the need for variable transformation that could reduce interpretability. The clustering results from K-Modes are relatively easier to interpret since cluster centers are represented by category modes. Several previous studies have successfully applied K-Modes to various categorical data cases. For example, study [11] utilized K-Modes to analyze the economic background of university students, resulting in clearly interpretable economic clusters. In other hand, study [12] used K-Modes to cluster cases of low birth weight (LBW) in Central Sulawesi based on maternal and environmental characteristics. K-Modes has also demonstrated greater stability than alternative methods when applied to datasets composed entirely of categorical variables [13], [14]. However, many real-world cases across different sectors involve mixed-type data, including both categorical and numerical variables, which cannot be effectively handled by either K-Means or K-Modes alone.

To address this challenge, an advanced clustering algorithm known for its high interpretability on mixed-type data, namely K-Prototypes, has been developed. This algorithm integrates the principles of K-Means for handling numerical variables and K-Modes for categorical variables, enabling the effective and integrated processing of mixed datasets [10]. The key advantage of K-Prototypes lies in its ability to represent cluster results optimally without requiring transformation of categorical variables. According to [15], K-Prototypes outperforms the Two-Step Cluster (TSC) method by producing clusters with higher internal homogeneity. Furthermore, K-Prototypes demonstrates greater stability, as clustering results remain consistent even when applied to large datasets or those dominated by mixed variables. The method also provides interpretable centroid values that effectively represent both numerical and categorical variables [16]. The effectiveness of K-Prototypes has been validated in previous studies, such as study [17], who applied it to cluster stunting prevalence across districts/cities, and [18], who compared Partitioning Around Medoids (PAM) and K-Prototypes for poverty clustering using multidimensional data. These findings are supported by [19], who highlighted that K-Prototypes produces more homogeneous and objective estimation domains when applied to mixed-variable geostatistical data. Given the similar characteristics found in household-level data, K-Prototypes offers a more representative clustering approach for identifying stunting risk groups.

Building upon the issues previously discussed, this study aims to conduct a comparative analysis of clustering methods to identify household-level stunting risk groups in West Java Province. The dataset used is derived from the March 2024 SUSENAS survey, consisting of both numerical and categorical variables. Given this data structure, the K-Prototypes algorithm was selected for its ability to process mixed-type data, while the K-Modes algorithm was employed to assess whether categorical variables alone can sufficiently capture the patterns of stunting risk factors. This comparative approach enables an examination of how variable types influence clustering performance and facilitates the identification of key patterns of stunting risk among households.

The novelty of this study lies in the implementation of a mixed-type data clustering framework enhanced through grid search optimization, which jointly determines the optimal number of clusters (k) and categorical weighting factor (γ) to produce the most representative clustering outcomes. This optimized comparative framework has rarely been applied in household-based stunting risk analyses in Indonesia, particularly using large-scale and heterogeneous datasets such

as SUSENAS. Methodologically, the study contributes by demonstrating how the integration of numerical and categorical variables can improve the precision of household segmentation compared to conventional methods like K-Means or standalone K-Modes. Substantively, the findings are expected to provide new insights into the socio-economic, sanitation, and housing characteristics of stunting-prone households. These insights are anticipated to support the development of more targeted, adaptive, and evidence-based intervention policies for stunting reduction in West Java.

II. METHODS

This study was conducted through a series of systematic stages, including: (1) data collection and variable selection; (2) data preprocessing to ensure the quality and completeness of the information; (3) data exploration to understand the distribution and initial characteristics of the variables used; (4) determination of the optimal parameters, namely, the number of clusters and the weighting of categorical variable contributions, through a grid search procedure; (5) implementation of clustering algorithms, specifically K-Modes and K-Prototypes (6) evaluation and comparison of clustering results using multiple internal validity indices, including the Silhouette Coefficient, and (7) cluster stability analysis to assess the robustness of the clustering structure using a bootstrapping approach combined with Adjusted Rand Index (ARI) metrics. The complete workflow of the research stages is presented in Figure 1 below.

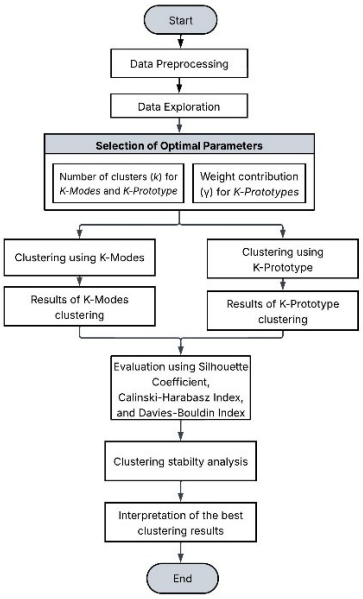


Figure 1. Research Workflow

A. Data Collection and Variable Selection

This study utilized secondary data from the March 2024 National Socioeconomic Survey (SUSENAS) conducted by Statistics Indonesia (Badan Pusat Statistik, BPS). The dataset was selected because it represents the socioeconomic conditions of households and contains indicators relevant to stunting risk factors. The analysis focused on households residing in West Java Province, with an initial total of 26,012 household observations.

In this study, a variable selection process was conducted, resulting in 26 indicators comprising both numerical and categorical data types, which were grouped into six main dimensions. The selected variables were determined based on their relevance to previous research findings and national policy frameworks on stunting risk factors, as stipulated in Presidential Regulation No. 72 of 2021 concerning the Acceleration of Stunting Reduction.

TABLE 1
RESEARCH VARIABLES

Category	Variable	Code	Type
Food Security [20]	Worried about not having enough food	R1701	Categorical
	Ever skipped healthy and nutritious meals	R1702	Categorical
	Consumed only a limited variety of foods	R1703	Categorical
Housing and Sanitation [3]	Ownership status of residential building	R1802	Categorical
	Floor area of the dwelling (m²)	R1804	Numerical
	Main wall material of the largest dwelling area	R1807	Categorical
	Main floor material of the largest dwelling area	R1808	Categorical
	Availability of toilet facilities	R1809A	Categorical
Drinking Water and Hygiene [3], [21]	Source of drinking water	R1810A	Categorical
	Turbidity of the main water source	R1813A	Categorical
	Color of the main water source	R1813B	Categorical
	Taste of the main water source	R1813C	Categorical
	Foam presence in the main water source	R1813D	Categorical
	Odor in the main water source	R1813E	Categorical
	Time required to fetch drinking water (minutes)	R1811B	Numerical
	Availability of soap, detergent, or antiseptic liquid	R1815C	Categorical
Economy and Assets [2]	Ownership of refrigerator	R2001B	Categorical
	Ownership of motorcycle	R2001H	Categorical
	Ownership of car	R2001K	Categorical
	Ownership of land/property	R2001M	Categorical

	Main source of household financing	R2101A	Categorical
Social Assistance [22]	Ever received the Kartu Keluarga Sejahtera (KKS)	R2202	Categorical
	Ever received the Program Keluarga Harapan (PKH)	R2203	Categorical
	Ever received the Bantuan Pangan Non-Tunai (BPNT)	R2207	Categorical
Household Demographics [23]	Number of household members	R301	Numerical
	Number of household members aged 0–4 years	R302	Numerical

B. Data Preprocessing

Before conducting the clustering analysis, the dataset underwent a preprocessing stage to ensure data quality and consistency. This preprocessing included handling missing values, duplicate entries, and miscoded responses to align the dataset with the requirements of the clustering analysis. Data cleaning was performed by removing invalid or non-informative responses such as "menolak menjawab" and "tidak tahu." Additional checks were carried out on specific variables to address potential coding errors. For instance, observations with the value 0 in variable R1815C were removed, as this value was inconsistent with the official metadata (1 = available; 5 = not available). Similarly, responses coded as "tidak tahu" in variable R1811B were excluded since they could not be transformed into a numerical format. After the preprocessing stage, a total of 22,161 observations and 26 variables were retained for subsequent analysis.

C. Data Exploration

Data exploration was conducted to obtain an initial understanding of the patterns of stunting risk factors among households in West Java and to ensure the readiness of the dataset prior to the clustering process. Visualizations in the form of population distribution maps based on shapefiles (SHP), bar charts, scatter plots, and treemaps were utilized to highlight the distribution of variables and the differences in characteristics between urban and rural areas.

D. Clustering Analysis Using K-Modes

The K-Modes method performs clustering analysis exclusively on categorical data to assess whether categorical variables alone are sufficient to represent the clustering structure. The procedure begins by determining the optimal number of clusters (k), which is a critical parameter for obtaining clusters that are both quantitatively robust and interpretatively meaningful [24].

The search space for k was defined in the range of $k = 2, 3, \dots, 10$ using a grid search approach. The K-Modes algorithm employs simple matching dissimilarity as the dissimilarity measure and uses the mode as the cluster centroid [25]. The dissimilarity values serve as the basis for evaluating the variation of the cost function across different k values, which is then used to determine the optimal number of clusters through the Elbow method. The elbow point in the plot of cost function versus k is considered the optimal number of clusters [26]. This dissimilarity measure is computed based on the number of mismatches between pairs of categorical attributes, as formulated in Equation 1 [13].

$$d_{cat}(x_i, \tilde{\mu}_l) = \sum_{j \epsilon C} \delta(x_{ij}, \tilde{\mu}_{lj}) \quad (1)$$

where $\delta(x_{ij}, \tilde{\mu}_{lj})$ is an indicator function that takes the value 0 if the two categories are identical and 1 if they differ. The cost function in the K-Modes algorithm aims to minimize the total dissimilarity across all clusters, which is formally expressed in Equation 2 [25].

$$J(k, \gamma) = \sum_{l=1}^{k} \sum_{x_i \epsilon C_l} d_{cat}(x_i, \tilde{\mu}_l) \quad (2)$$

where x_i denotes the i-th observation, μ_l is the centroid of the l-th cluster, $j \in C$ is the index of categorical features, x_{ij} represents the value of the i-th observation on feature j, μ_{lj} refers to the centroid for numerical features, and $\tilde{\mu}_{lj}$represents the mode for categorical features.

After the optimal number of clusters (k) is determined, the K-Modes algorithm operates through an iterative procedure that begins with the initialization of cluster centers (modes) by randomly selecting data objects as initial representatives [27]. Once the initial modes are set, the dissimilarity between each observation and all cluster centers is computed using the simple matching dissimilarity measure. Each observation is then assigned to the cluster with the smallest dissimilarity value. Following the assignment, cluster centers are updated by recalculating the mode of each attribute within the newly formed clusters. This process of distance computation, data assignment, and mode updating is repeated iteratively until no further object movement occurs between clusters, indicating that convergence has been achieved.

E. Clustering Analysis Using K-Prototypes

Clustering analysis was also performed using the K-Prototypes method, which is capable of handling mixed-type data (numerical and categorical) making it suitable for the characteristics of the dataset used in this study. The K-Prototypes algorithm integrates the principles of K-Means for numerical variables and K-Modes for categorical variables.

Prior to clustering, the optimal combination of the number of clusters (k) and the weighting parameter for categorical variables (γ) was determined. The search space was defined as $k = 2, 3, \dots, 10$ and $\gamma \in \{0.5, 1.0, 1.5, 2.0\}$, explored through a grid search approach. Parameter evaluation was conducted by computing the cost function, which represents the total distance between each observation and its assigned cluster center for every combination of (k, γ). The optimal number of clusters was determined using the Elbow method, based on the point of significant reduction in the cost function value.

In the K-Prototypes algorithm, the cost function combines the squared Euclidean distance for numerical variables with the simple matching dissimilarity for categorical variables. The simple matching dissimilarity follows the same formulation as in Equation 1, while the squared Euclidean distance is defined in Equation 3 [10], [19].

$$d_{num}(x_i, \mu_l) = \sum_{j \epsilon N} (x_{ij} - \mu_{lj})^2 \quad (3)$$

where x_i denotes the *i*-th observation, $\mu_l$represents the centroid of the *l*-th cluster, $j \in N$ and $j \in C$ are indices of numerical and categorical features respectively, x_{ij} is the value of the *i*-th observation on feature *j*, μ_{lj} is the centroid for numerical variables, and $\tilde{\mu}_{lj}$ is the mode for categorical variables.

The overall cost function for K-Prototypes is expressed in Equation 4 [15].

$$J(k, \gamma) = \sum_{l=1}^{k} \sum_{x_i \epsilon C_l} d_{num}(x_i, \mu_l) + \gamma . d_{cat}(x_i, \tilde{\mu}_l) \quad (4)$$

The combined distance function (d_{ij}) integrates both numerical and categorical components, incorporating the coefficient γ as a balancing factor between their respective contributions, as formulated in Equation 5 [28].

$$d_{ij} = \sum_{j \epsilon N} (x_{ij} - \mu_{lj})^2 + \gamma \sum_{j \epsilon N} (x_{ij} - \mu_{lj})^2 \quad (5)$$

Once the optimal parameters were obtained, clustering was performed iteratively by reassigning observations to the nearest cluster until stability was achieved, following three main steps [28]. First, the initialization of centroids was conducted by randomly selecting the initial cluster centers. Next, in the assignment step, the distance of each observation to all centroids was computed using the combined distance function (d_{ij}), and each observation was assigned to the cluster with the smallest distance. Finally, during the update step, the cluster centers were recalculated by computing new means for numerical variables and new modes for categorical variables. These steps were iteratively repeated to minimize the cost function until convergence, that is, when the cluster memberships no longer changed.

The selection of the K-Modes and K-Prototypes algorithms was based on the characteristics of the March 2024 SUSENAS dataset, which consists of both categorical and numerical variables. The K-Modes algorithm was employed to assess the extent to which categorical variables alone can form meaningful clusters, while K-Prototypes was chosen to handle mixed-type data and produce a more representative segmentation of households in terms of socio-economic and environmental conditions.

F. Evaluation and Comparison of Clustering Results

Evaluation was conducted to assess the quality of the clusters formed. The evaluation metrics used in this study were the Silhouette Index, the Calinski–Harabasz Index (CH Index), and the Davies–Bouldin Index (DBI), each of which measures the compactness and separation of clusters using different approaches.

Silhouette Index quantifies how well each object is assigned to its cluster compared to the nearest neighboring cluster. The Silhouette value is defined in Equation 6 [29].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

where $a(i)$ represents the average distance between an object and all other points within the same cluster, while $b(i)$ denotes the minimum average distance between the object and all points in other clusters. The overall Silhouette Index value across all observations is used to validate cluster quality, where a value approaching 1 indicates well-formed and clearly separated clusters

Meanwhile, the Calinski–Harabasz Index (CH) evaluates cluster quality through the ratio of between-cluster variance to within-cluster variance. The CH Index formula is given in Equation (7) [30]:

$$CH(k) = \frac{N - k}{k - 1} \cdot \frac{\sum_{a=1}^{k} d(cc_a, GC)}{\sum_{a=1}^{k} \sum_{x \epsilon C} d(x, cc_a)} \quad (7)$$

where N is the number of observations, k is the number of clusters, cc_a is the centroid of cluster a, and GC is the global centroid of the entire dataset. A higher CH value indicates better-separated and more compact clusters.

On the other hand, the Davies–Bouldin Index (DBI) measures how similar a cluster is to its most similar neighboring cluster by considering internal dispersion and the distance between centroids. The DBI formula is given in Equation (8) [30]:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{d(cc_i, cc_a)} \right) \quad (8)$$

where S_i is the average distance of all points in cluster i to its centroid, and $d(cc_i, cc_j)$ is the distance between the centroids of clusters i and j. A lower DBI value indicates higher compactness and better separation between clusters.

G. Clustering Stability Analysis

Cluster stability analysis was conducted to assess the extent to which the clustering results remain consistent under data variation. This approach is particularly important in the context of heterogeneous socioeconomic data, where small perturbations in the sample may yield different cluster structures.

The stability assessment was performed using the bootstrapping method [31], in which the original dataset of size n was resampled $B = 50$ times with replacement. For each bootstrap replication, clustering was performed using the best-performing method, as identified from the internal evaluation based on the Silhouette, Calinski–Harabasz, and Davies–Bouldin indices. Each replication produced a set of cluster labels $C_1, C_2, \dots, C_B$.

To evaluate the consistency of the clustering structure, each bootstrap result C_b was compared with the original clustering C_0 using two partition similarity measures: the Adjusted Rand Index and the Normalized Mutual Information.

The Adjusted Rand Index (ARI) measures the agreement between two clustering results while correcting for chance. ARI is formulated in Equation (9) [32]:

$$ARI(U,V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11})(N_{00}+N_{10})(N_{10}+N_{11})} \quad (9)$$

where N_{11} denotes the number of pairs of observations assigned to the same cluster in both partitions, N_{00} the number of pairs assigned to different clusters in both, and N_{01} and N_{10} represent the inconsistent pairs.

The Normalized Mutual Information (NMI) quantifies the mutual information shared between two partitions while accounting for their individual entropies, as formulated in Equation (10) [32]:

$$NMI(U,V) = \frac{2I(U;V)}{H(U)H(V)} \quad (10)$$

with

$$I(U;V) = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N}\log\frac{n_{ij}}{a_i b_j} \quad (11)$$

$$H(U) = -\sum_{i}\frac{a_i}{N}\log\frac{a_i}{N} \quad (12)$$

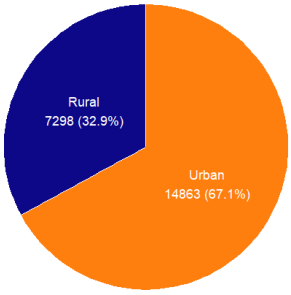
$$H(V) = -\sum_{j}\frac{b_j}{N}\log\frac{b_j}{N} \quad (13)$$

here, n_{ij} denotes the number of observations in cluster i of partition U and cluster j of partition V, while a_i and b_j represent the total members in clusters $i$and j, respectively. For each bootstrap replication, ARI and NMI were computed and subsequently averaged to obtain a global measure of clustering consistency. An average ARI above 0.80 or NMI above 0.85 was interpreted as evidence of a stable and reproducible cluster structure under sampling variation.

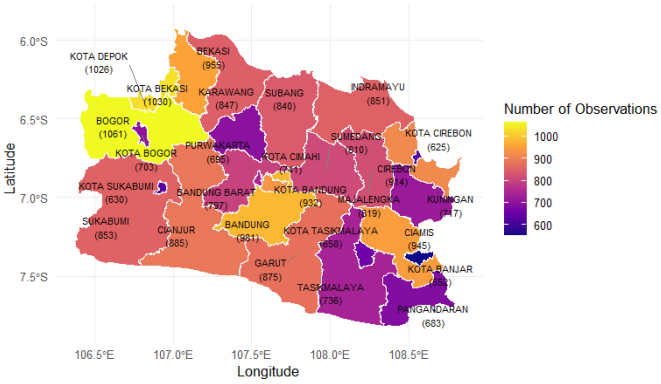
III. Result And Discussion

A. Data Exploration

This study focuses on data exploration and clustering analysis of households vulnerable to stunting in West Java. Six key aspects are examined, namely food security, housing and sanitation, water and hygiene, economy and assets, social assistance, and household demographics. Exploring the condition of each aspect within households is essential, as regional heterogeneity influences the types of stunting interventions that should be implemented.



(a)



(b)

Figure 2. (a) Proportion of Respondents by Type of Residential Area; (b) Distribution of Household Observations by Regency/City

As shown in Figure 2(a), the majority of respondents reside in urban areas (67.1%), while those living in rural areas account for 32.9%. This initial overview can serve as a reference for a more detailed examination of regional conditions based on the aspects presented in Table I. Rural areas often have more limited access to healthcare services, sanitation, and nutritious food; therefore, despite having fewer observations, the potential risk of stunting tends to be relatively higher in these regions.

The distribution map in Figure 2(b) indicates that Bogor, Depok, and Bekasi have high population densities. Population density in a given area is an important factor to consider, as it can lead to overcrowding, competition for resources, and unequal access to nutritional services. On the other hand, areas with lower population densities, such as Pangandaran or Cirebon City, may also face stunting risks due to limited attention from health and sanitation policies and higher economic vulnerability.

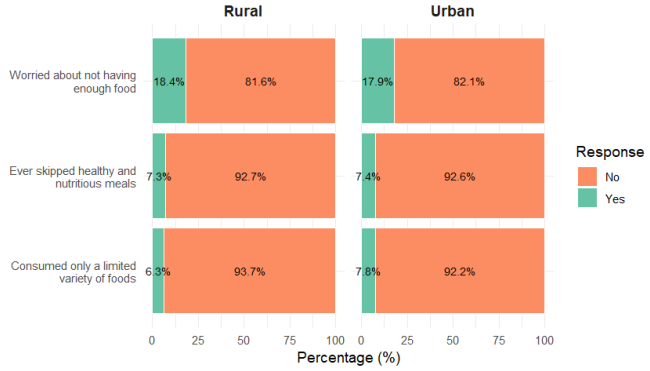


Figure 3. Proportion of Household Food Security by Type of Area

Food security is a crucial aspect in analyzing stunting risk factors. The results of the food security exploration presented in Figure 3 show that rural and urban households exhibit relatively similar levels of vulnerability. About 18.4% of rural households and 17.92% of urban households reported concern about not having enough food. Urban households were more likely to experience limited food variety (7.79%) compared to rural households (6.29%). This indicates that although food access tends to be easier in urban areas, the diversity of food consumed is often lower, which may have implications for

children's nutritional quality. Meanwhile, limited access to healthy and nutritious food was experienced by approximately 7.3% of rural households and 7.4% of urban households. This pattern suggests that regional differences do not significantly affect food insecurity levels, implying that other factors should be considered in forming stunting risk clusters.

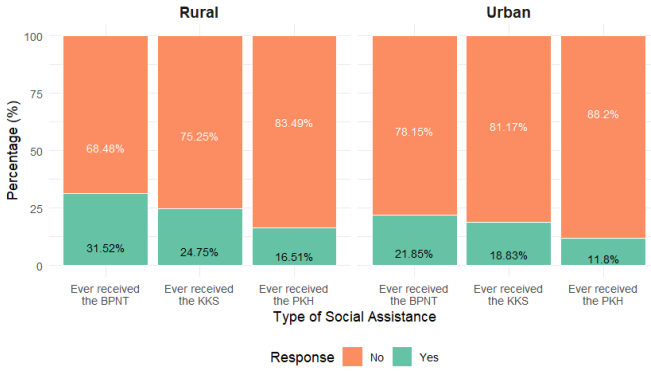


Figure 4. Proportion of Social Assistance Recipients by Type of Area

As shown in Figure 4, the percentage of social assistance recipients is higher in rural areas than in urban areas, although the overall proportion remains relatively small. Social assistance is one of the key interventions in reducing economic vulnerability and stunting risk. The *Kartu Keluarga Sejahtera* (KKS) program is received by only 24.7% of rural households and 18.8% of urban households. Furthermore, the *Bantuan Pangan Non-Tunai* (BPNT) program, which is directly related to food access, is received by 31.5% of rural households and 21.9% of urban households. On the other hand, the *Program Keluarga Harapan* (PKH), which explicitly targets households with children under five and pregnant women (groups vulnerable to stunting), is received by only 16.5% of rural households and 11.8% of urban households.

Although the coverage of social assistance programs in rural areas is relatively higher than in urban areas, the proportions are still limited and have not yet been optimal in reducing economic vulnerability or supporting stunting prevention efforts. This underscores the importance of expanding program coverage and improving effectiveness to ensure that assistance is more accurately targeted toward vulnerable households.

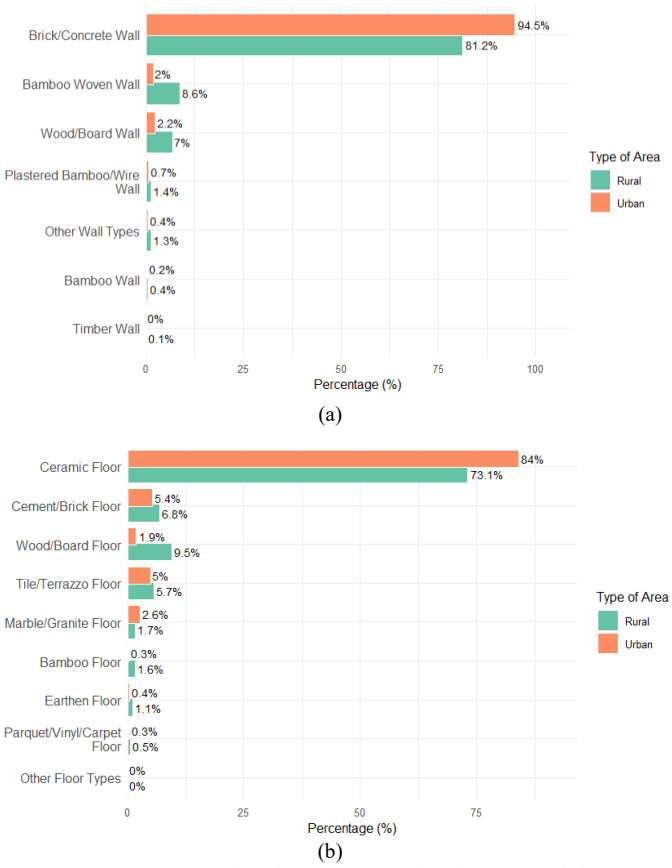


Figure 5. (a) Proportion of Wall Materials and (b) Floor Materials by Area

Figure 5 (a) shows that the majority of houses in both urban and rural areas use brick walls as the primary wall material. However, in rural areas, a considerable number of houses still use wood, boards, or woven bamboo. Based on Figure 5(b), most houses use ceramic flooring, particularly in urban areas, while in rural areas, the use of wood, boards, soil, and other simple materials is more common.

Figure 6. Treemap of Wall and Floor Materials

Similarly, Figure 6 shows that most houses are dominated by the combination of brick walls and ceramic floors compared to other materials. However, there are still houses that use soil floors, wood or board materials, and woven bamboo. These non-permanent structures indicate disparities in housing quality, which may lead to substandard

environmental health conditions and serve as a risk factor for stunting.

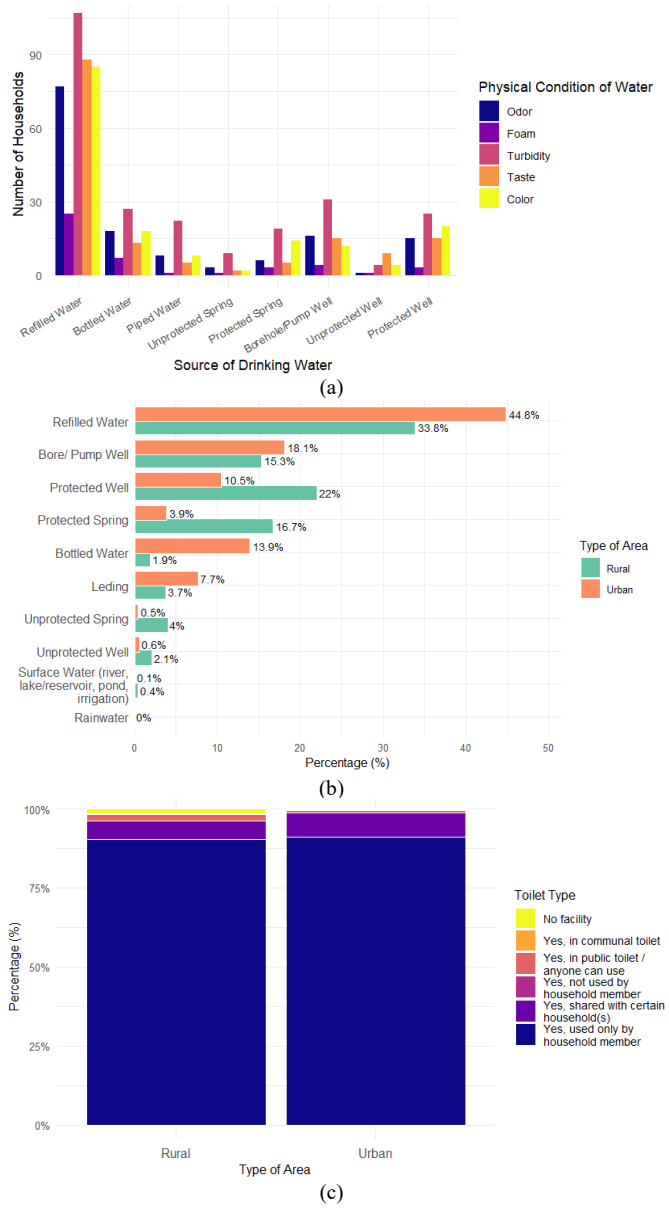


Figure 7. (a) Physical Condition of Drinking Water by Source; (b) Proportion of Drinking Water Sources; and (c) Toilet Access in Urban and Rural Areas

As shown in Figure 7, refilled and branded bottled water are the main sources of drinking water for households, especially in urban areas. However, even in urban regions, the physical quality of water is still reported to have issues such as odor, turbidity, and changes in taste or color. Moreover, some urban households still consume well water or even unprotected spring water. In rural areas, the use of protected wells and springs is also quite common.

In terms of sanitation, most households in both rural and urban areas have private toilets, but rural areas still show a higher proportion of households using communal latrines or lacking toilet facilities altogether. This condition indicates a gap in environmental quality between rural and urban regions. Access to safe drinking water and adequate sanitation facilities plays a crucial role in preventing infectious diseases such as diarrhea, which is a major biological pathway leading to stunting. Therefore, disparities in access to clean water and sanitation may contribute to a higher risk of stunting in rural areas compared to urban ones. This finding is consistent with existing literature emphasizing that inadequate sanitation and unsafe drinking water access are significant risk factors for stunting in Indonesia [3].

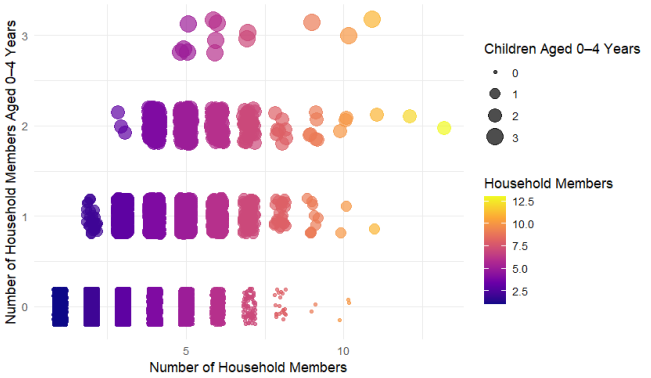
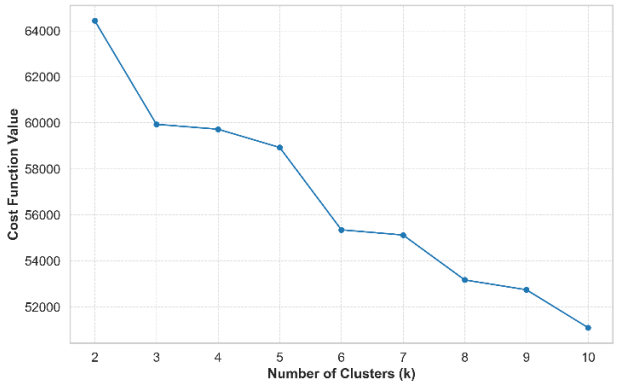


Figure 8. Distribution of the Number of Children Under Five by Household Size

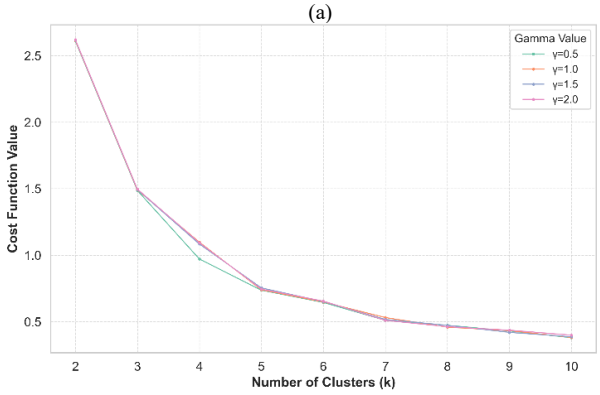
Figure 8 shows that the greater the number of household members, the more children under five years old (0–4 years) are present in the household. Larger households with 8–12 members tend to have greater variation in the number of young children compared to smaller households. This condition implies an increased caregiving burden and potential resource limitations, such as nutrition, healthcare, and parental attention, which can serve as risk factors for stunting.

B. Determination of Optimal Clustering Parameters

Determining the appropriate parameters is a crucial stage in cluster analysis to ensure optimal grouping results, both statistically and substantively. Each method requires different parameters: K-Modes only requires the number of clusters (k), while K-Prototypes requires a combination of the number of clusters and the weighting of categorical variable contributions (γ). The optimal number of clusters was determined using the Elbow method, which identifies the point at which the decrease in the cost value begins to slow down significantly. Figure 9 presents a visualization of the changes in the cost function values relative to the number of clusters (k) for both the K-Modes and K-Prototypes methods, which serves as the basis for determining the optimal number of clusters.



(a)



(b)

Figure 9. (a) Elbow Plot of the Cost Function for K-Modes; (b) for K-Prototypes

Figure 9 (a) illustrates the decline in the cost function value relative to the number of clusters in the K-Modes algorithm. A significant decrease occurs between $k = 2$ and $k = 3$, followed by a slower decline at $k = 4$ and $k = 5$, before dropping again at $k = 6$. After $k = 6$, the cost reduction curve becomes flatter and more stable. This pattern indicates that the elbow point occurs at $k = 6$, where adding more clusters beyond this point no longer results in a substantial reduction in cost. Therefore, the optimal number of clusters for the K-Modes method is determined to be six.

Figure 9 (b) shows the cost function reduction for the K-Prototypes algorithm. The cost decreases sharply from $k = 2$ to $k = 5$, then tends to level off afterward, suggesting that the elbow point occurs at $k = 5$. Although Figure 9 (b) provides a general overview of the cost function's decreasing trend across different values of γ, the differences among γ values at $k = 5$ are not visually distinct in the plot. To clarify these differences, an additional visualization in the form of a bar chart was created to specifically display variations in the cost function values for each γ at the identified optimal number of clusters.

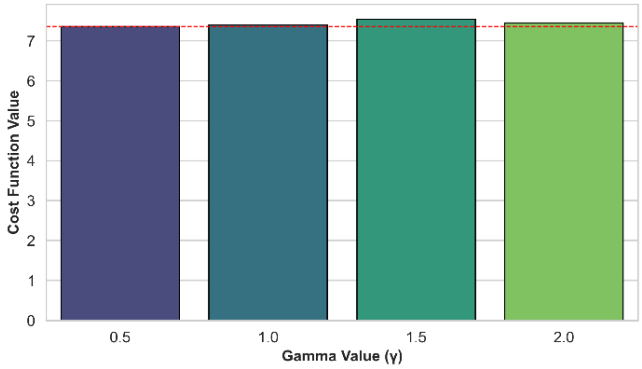
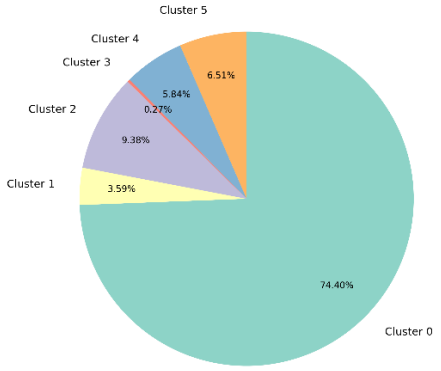


Figure 10. Bar Chart of Cost Function Values at $k = 5$ for Various γ Values

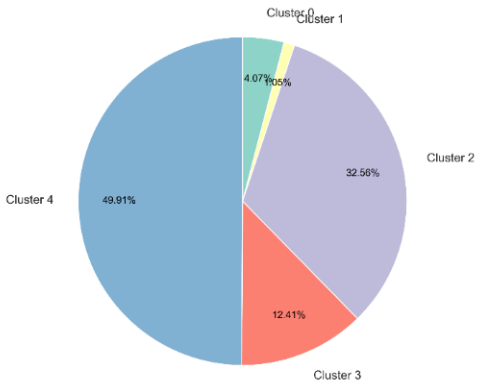
As shown in Figure 10, the parameter $\gamma = 0.5$ produces the lowest cost function value compared to other γ values, although the difference is relatively small. Therefore, the combination of $k = 5$ and $\gamma = 0.5$ was selected as the optimal parameter configuration for implementing the K-Prototypes algorithm in the case of households at risk of stunting in West Java.

C. Clustering Implementation

Clustering was performed using the optimal parameter configurations obtained from the previous stage. Figure 11 presents the distribution of household proportions across each cluster generated by both methods.



(a)



(b)

Figure 11. Cluster Proportions from (a) K-Modes and (b) K-Prototypes Methods

It can be observed that the K-Modes method produces a highly dominant Cluster 0, which accounts for 74.40% of all observations. Meanwhile, the remaining five clusters have much smaller proportions: Cluster 2 accounts for 9.38%, Cluster 5 for 6.51%, Cluster 4 for 5.84%, Cluster 1 for 3.59%, and Cluster 3 is the smallest with only 0.27%. This uneven distribution indicates that most households share very similar categorical characteristics, leading to their concentration within a single large cluster. As a result, variation among the other clusters is less clearly represented, highlighting the limitation of K-Modes in distinguishing household groups when the data are dominated by categorical variables with similar patterns.

In contrast, the results from the K-Prototypes method show a relatively more balanced distribution. Cluster 4 is the largest group, representing 49.91% of the data, followed by Cluster 2 (32.56%), Cluster 3 (12.41%), Cluster 0 (4.07%), and Cluster 1, which is relatively small at 1.05%. Although one cluster still dominates, this distribution is considerably more balanced compared to K-Modes. This pattern demonstrates that K-Prototypes produces a more representative household segmentation, as it simultaneously incorporates information from both numerical and categorical variables, resulting in more proportional clustering and better differentiation of household characteristics.

D. Evaluation and Comparison of Clustering Results

The quality of the clustering results was evaluated using three internal validation metrics: Silhouette Score (SC), Calinski–Harabasz Index (CH), and Davies–Bouldin Index (DBI). These metrics respectively measure the compactness and separation of clusters (SC), the ratio of between-cluster to within-cluster variance (CH), and the average similarity between each cluster and its most similar counterpart (DBI). Table II summarizes the evaluation results for both clustering methods.

TABLE 2
CLUSTERING EVALUATION

Method	Silhouette Score	CH Index	DBI
K-Modes	0.2922	3,976.1219	1.5274
K-Prototypes	0.6681	13,890.6011	0.4607

The K-Prototypes method produced significantly better results than K-Modes across all metrics. A Silhouette Score of 0.6681 indicates that the formed clusters are internally consistent and well-separated from one another. Meanwhile, the high CH Index (13,890.6011) and the low DBI (0.4607) further confirm the robustness and compactness of the clusters. In contrast, the K-Modes method yielded lower SC and CH values, and a higher DBI, suggesting overlapping clusters with suboptimal separation. These findings confirm that K-Prototypes is more suitable for clustering households with stunting risk factors in West Java, as it effectively integrates both numerical and categorical variables in the clustering process.

E. Clustering Stability Analysis

Based on the internal evaluation using three key metrics—Silhouette Score, Calinski–Harabasz Index (CH), and Davies–Bouldin Index (DBI)—the K-Prototypes method with optimal parameters ($k = 5, \gamma = 0.5$) was selected as the best-performing approach for household clustering analysis in West Java. To further assess the robustness and reproducibility of the clustering results, a cluster stability analysis was carried out using the bootstrapping approach with $B = 50$ replications.

Figure 12 presents the distribution of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) values obtained from the 50 bootstrap replications.

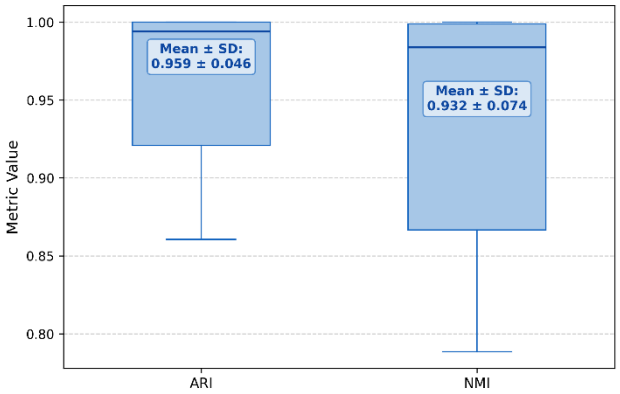


Figure 12. Distribution of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) Based on 50 Bootstrap Replications

Both stability metrics exhibit values concentrated in a high range (close to 1.0), with a mean ARI of 0.959 (±0.046) and a mean NMI of 0.932 (±0.074). The small standard deviations indicate that the clustering results are highly stable and reproducible, even when random variations are introduced into the data. These findings imply that the cluster structures are not artifacts of a specific sample but instead reflect consistent and persistent patterns within the socioeconomic and environmental dimensions of households in West Java.

Therefore, the stability analysis reinforces the internal validation results, confirming that the K-Prototypes algorithm not only achieves superior cluster separation but also maintains high robustness under resampling conditions, making it a reliable and interpretable approach for household-level stunting risk segmentation.

F. Cluster Profile Interpretation

After obtaining the optimal clustering results using the K-Prototypes method with $k = 5$ and $\gamma = 0.5$, the next step is to evaluate the characteristics of each cluster. The interpretation is based on the centroid values of both categorical and numerical variables, which represent the general household profile within each group. Table III presents a summary of the key characteristics for each cluster generated by the K-Prototypes method.

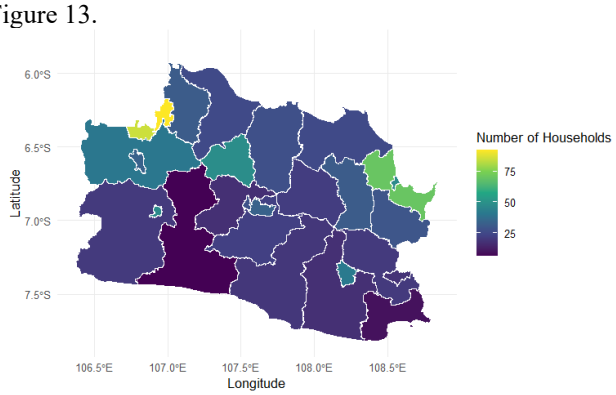
TABLE 3
MAIN CHARACTERISTICS OF EACH CLUSTER

Cluster	Characteristics	Stunting Risk
0	Good food security; lives in an owned house with adequate housing and sanitation facilities; uses refilled drinking water. Middle-income economy with an average floor area of approximately 199 m²; water collection time around 6 minutes; 3–4 household members, with few children under five.	Low
1	Households with good food security, permanent housing, and proper sanitation; uses branded bottled drinking water. Upper-middle-income economy with well-equipped facilities, including a car. The largest floor area (approximately 354 m²), slightly longer water collection time (8 minutes), around 3 household members, and few children under five.	Low
2	Good food security with permanent housing, available sanitation facilities, and the use of refilled drinking water. Middle-income economy with a moderate floor area (approximately 82 m²), relatively short water collection time (5 minutes), around 3–4 household members, and few children under five.	Medium
3	Good food security; lives in a permanent house with adequate sanitation and access to clean refilled drinking water. Upper-middle-income economy with a medium floor area (approximately 128 m²), water collection time of 5.4 minutes, around 3–4 household members, and few children under five.	Medium
4	Good food security; owns a permanent house with adequate sanitation and uses refilled drinking water. Lower-middle-income economy with a smaller floor area (approximately 45 m²), water collection time of around 5 minutes, about 3 household members, and more children under five compared to other clusters.	High

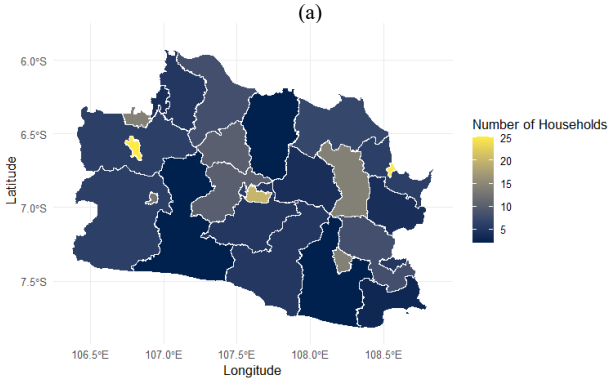
The centroid analysis revealed three levels of household socioeconomic conditions: lower-middle, middle, and upper-middle. Across all clusters, no food insecurity was identified. All households resided in permanent dwellings constructed with high-quality materials, had adequate sanitation facilities, and possessed relatively high household assets (e.g., refrigerators, motorcycles, and in some cases, cars or land). Moreover, none of the households received social assistance. The primary distinctions among clusters were observed in house size, water source, car ownership, and water collection time. Cluster 1 represented the most affluent group, characterized by the largest housing area (approximately 354 m²), the use of branded bottled water, and car ownership. In contrast, Cluster 4 had the smallest average housing size (around 45 m²) and a higher number of children under five compared to other clusters.

From the perspective of stunting risk, most clusters fall into the low to moderate risk categories due to adequate access to food, sanitation, and housing. Cluster 1 exhibited a very low risk and could serve as a model for other households. Clusters 2 and 3 require closer monitoring of child nutrition, as their relatively smaller houses are occupied by three to four household members, despite having proper sanitation. This finding is consistent with Novianti et al. (2023) and Batool et al. (2023), who highlighted that permanent housing with proper sanitation and safe water access serves as a protective factor against stunting

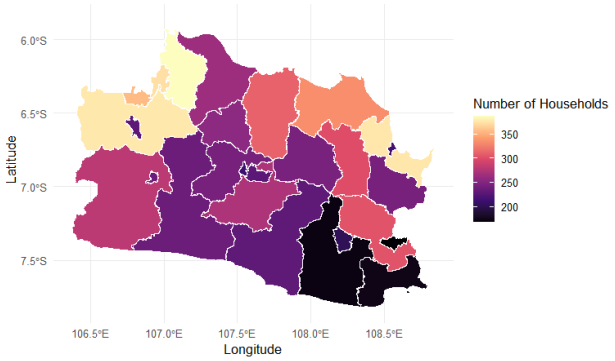
Meanwhile, Cluster 4 showed the highest stunting risk, characterized by lower-middle economic conditions, smaller housing size, and a higher number of young children and household members. High housing density may negatively affect child health. These findings indicate that stunting prevention interventions should not solely focus on improving nutritional intake but must also address structural and environmental household factors. Enhancing access to clean and reliable water, improving housing quality, and strengthening basic sanitation systems are critical components in reducing the risk of stunting. The spatial distribution of households in each cluster is illustrated in Figure 13.



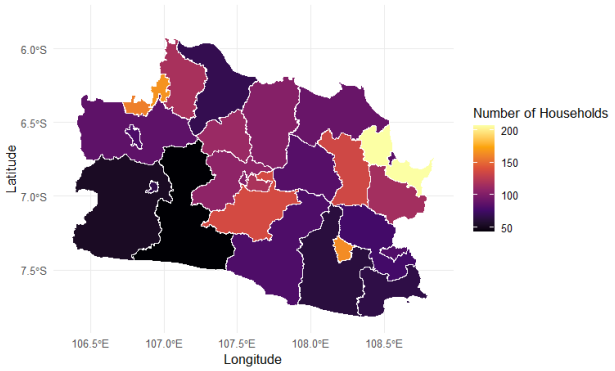
(a)



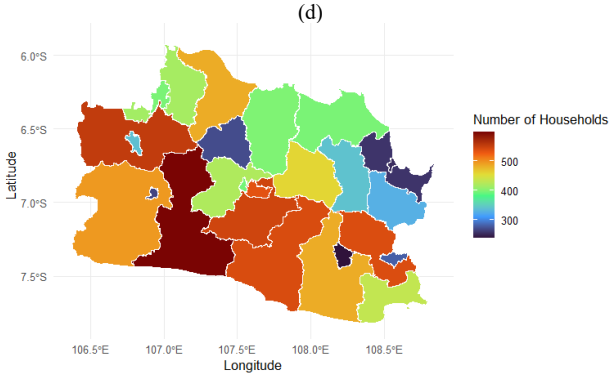
(b)



(c)



(d)



(e)

Figure 13. Spatial Distribution of Household Clusters at Risk of Stunting in West Java: (a) Cluster 0; (b) Cluster 1; (c) Cluster 2; (d) Cluster 3; and (e) Cluster 4

Based on the spatial distribution of household clusters across districts and cities in West Java, a clear geographic pattern emerges between household welfare levels and residential areas. Cluster 4, representing households with relatively lower socioeconomic conditions, is widely distributed and dominates most districts, particularly in the western region, such as Cianjur and Bandung Regencies. In contrast, Clusters 2 and 3 (moderate welfare level) are more concentrated in both eastern and western areas, reflecting regions with moderate socioeconomic status, as well as adequate infrastructure and food security.

Meanwhile, Clusters 0 and 1, which represent households with higher welfare and better environmental conditions, appear in smaller proportions and are mainly located in urban areas or economic centers, such as Bogor City and Cirebon City. Overall, the map highlights that household welfare disparities in West Java exhibit a spatial pattern, where rural areas tend to be inhabited by households with higher socioeconomic vulnerability and greater stunting risk compared to urban areas.

The clustering results provide an empirical basis for formulating more targeted intervention policies based on household characteristics. Households in the high-risk cluster (Cluster 4)—characterized by lower-middle economic status, limited dwelling space, a higher number of children under five, and restricted access to clean water and sanitation—should be prioritized in the implementation of cross-sectoral programs. The most relevant interventions include the Free Nutritious Meals Program (MBG), which aims to improve the nutritional intake of schoolchildren, pregnant women, and nursing mothers in vulnerable households. This program can be integrated with the Bangga Kencana Program implemented by BKKBN through the *Family Assistance Team (TPK)* to monitor and support families at risk of stunting from pregnancy through early childhood. In addition, strengthening the WASH (Water, Sanitation, and Hygiene) program supported by UNICEF and the Government of Indonesia under the framework of Sustainable Development Goal 6 is crucial, particularly in rural areas with limited access to clean water and sanitation facilities as reflected in Cluster 4 characteristics. The combination of these three interventions is expected to enhance household nutritional resilience, improve living environments, and reduce infection risks that are major contributors to stunting.

Households in the medium-risk clusters (Clusters 2 and 3)—with middle socioeconomic conditions, moderately adequate housing, yet limited space and water access—require *promotive–preventive* approaches. The Bangga Kencana Program remains relevant for this group, especially through nutrition education for adolescents and strengthening parenting practices that support child growth and development. In addition, implementing the Sustainable Food House Program (Rumah Pangan Lestari, RPL) can serve as a strategic effort to enhance household food self-sufficiency by utilizing home gardens as sources of healthy and diverse foods. Integrating RPL with family nutrition education can help middle-income households diversify food consumption and maintain nutritional balance for children.

Meanwhile, low-risk clusters (Clusters 0 and 1)—dominated by middle- to upper-income households with complete facilities, good water access, and few young children—can serve as best practice households for nutritional resilience and healthy living behaviors. These households can be engaged in educational and community mentoring programs to model the implementation of Sustainable Food Houses (RPL) focused on urban food sustainability and to act as learning partners for regions or families still at high risk. This network-based household approach enables the diffusion of positive practices from well-off groups to vulnerable ones across social and ecological dimensions.

Overall, these findings align with Indonesia's national policy direction outlined in Presidential Regulation No. 72 of

2021 on the Acceleration of Stunting Reduction, which emphasizes the convergence of multi-sectoral interventions across nutrition, sanitation, clean water, and family development. The data-driven clustering framework developed in this study supports the implementation of *Perpres 72/2021* by providing a scientific foundation for determining more objective and measurable intervention priorities at the household and district/city levels. Thus, the findings not only illustrate stunting risk segmentation but are also expected to make a concrete contribution to strengthening evidence-based policymaking in Indonesia's national stunting reduction efforts.

IV. CONCLUSION

This study explores household-level stunting risk factors in West Java through a data-driven framework integrating visualization and clustering analysis using the K-Modes and K-Prototypes algorithms. The visualization results reveal clear differences between rural and urban areas: rural households tend to have smaller living spaces, longer water collection times, and a higher proportion of young children, while urban households generally have larger dwellings, faster access to water, and more diverse economic assets.

Among the two methods tested, the K-Prototypes algorithm demonstrated superior performance, producing a more balanced cluster distribution and higher Silhouette Score, greater Calinski–Harabasz Index, and lower Davies–Bouldin Index compared to K-Modes. These results indicate more compact and well-separated clusters and confirm the exploratory finding that socioeconomic and environmental household characteristics are the key differentiating factors of stunting risk.

The stability assessment using bootstrapping with 50 replications yielded highly consistent results, with a mean Adjusted Rand Index (ARI) of 0.959 and a mean Normalized Mutual Information (NMI) of 0.932, both with small standard deviations, confirming that the cluster structures are highly stable and reproducible. The optimal five-cluster solution identified three socioeconomic levels—upper-middle, middle, and lower-middle—with the highest stunting risk observed among lower-middle-income households that have smaller living spaces, limited water access, and a larger number of children under five, particularly in densely populated rural districts.

The distribution pattern indicates that regions with a high concentration of low-income households tend to overlap with areas of higher stunting risk. Therefore, improving food security, housing quality, sanitation, and access to clean water in rural areas should be prioritized in stunting reduction programs. The integration of visualization and clustering results provides a comprehensive understanding that stunting interventions must consider both regional context and household characteristics to ensure more targeted and effective strategies.

The clustering framework proposed in this study provides a strong analytical foundation for evidence-based policymaking at the household level. This approach aligns with the multisectoral direction of Presidential Regulation No. 72 of 2021 and reinforces that data-driven and context-specific segmentation can enhance precision, equity, and regional adaptability in Indonesia's stunting reduction strategies.

REFERENCES

[1] TP2S, "Garut Darurat Stunting, Prevalensi Tertinggi di Jawa Barat - TP2S." Accessed: Sept. 18, 2025. [Online]. Available: https://stunting.go.id/garut-darurat-stunting-prevalensi-tertinggi-di-jawa-barat/

[2] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy, and L. M. Neufeld, "A review of child stunting determinants in Indonesia," *Maternal & Child Nutrition*, vol. 14, no. 4, p. e12617, Oct. 2018, doi: 10.1111/mcn.12617.

[3] S. Novianti, E. Huriyati, and R. S. Padmawati, "Safe Drinking Water, Sanitation and Mother's Hygiene Practice as Stunting Risk Factors: A Case Control Study in a Rural Area of Ciawi Sub-district, Tasikmalaya District, West Java, Indonesia," *Ethiop J Health Sci*, vol. 33, no. 6, Dec. 2023, doi: 10.4314/ejhs.v33i6.3.

[4] Kementerian Sekretariat Negara RI, "PERPRES No. 72 Tahun 2021 Tentang Percepatan Penurunan Stunting," Database Peraturan | JDIH BPK. Accessed: Sept. 18, 2025. [Online]. Available: http://peraturan.bpk.go.id/Details/174964/perpres-no-72-tahun-2021

[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," *simkom*, vol. 9, no. 1, pp. 36–46, Jan. 2024, doi: 10.51717/simkom.v9i1.356.

[6] F. Ramadhani, "Spatial Clustering Analysis of Stunting in North Sumatra Based on Environmental Factors Using K-Means Algorithm," *Data Science: J. of Computing and Appl. Informatics*, vol. 9, no. 2, July 2025, doi: 10.32734/jocai.v9.i2-17179.

[7] 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, Jan. 2025, doi: 10.33395/sinkron.v9i1.14311.

[8] I. P. Ica, Martanto, Arif Rinaldi Dikananda, and Dede Rohman, "Use of K-Means Algorithm in Model Improvement Production Data Grouping for Determination Convection Production Strategy," *j. of artif. intell. and eng. appl.*, vol. 4, no. 2, pp. 916–926, Feb. 2025, doi: 10.59934/jaiea.v4i2.775.

[9] A. Ahmad and S. S. Khan, "Survey of State-of-the-Art Mixed Data Clustering Algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

[10] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," 1998.

[11] A. Wirawan and D. Prasetyawan, "Analisis cluster data latar belakang ekonomi mahasiswa untuk rekomendasi penentuan uang kuliah tunggal dengan model K-Modes," *infotech*, vol. 4, no. 2, pp. 234–246, Dec. 2023, doi: 10.37373/infotech.v4i2.898.

[12] C. Aprotama, Yenni Kurniawati, Muhammad Arief Rivano, and Devi Yopita Sipayung, "Application of K-Modes Clustering Method to Identify Low Birth Weight Factors in Central Sulawesi Province," *ujsds*, vol. 3, no. 2, pp. 164–171, May 2025, doi: 10.24036/ujsds/vol3-iss2/357.

[13] K. S. Dorman and R. Maitra, "An efficient K-modes algorithm for clustering categorical datasets," *Statistical Analysis*, vol. 15, no. 1, pp. 83–97, Feb. 2022, doi: 10.1002/sam.11546.

[14] A. Yildiz and E. E. Aksoy, "Investigation of Individual Investment Preferences with K-Mode Cluster Analysis Based on Socio-Demographic Characteristics," *IJARBSS*, vol. 10, no. 7, p. Pages 280-295, July 2020, doi: 10.6007/IJARBSS/v10-i7/7415.

[15] S. Sulastri, B. Susetyo, and I. M. Sumertajaya, "The Clustering of the Aquaculture Fisheries Companies in Indonesia Using the K-Prototypes and Two Step Cluster (TSC) Algorithm," *International Journal of Sciences*, vol. 58, no. 1, 2021.

[16] A. Mohd, L. E. Teoh, and H. L. Khoo, "Passengers' requests clustering with k-prototype algorithm for the first-mile and last-mile (FMLM) shared-ride taxi service," *Multimodal Transportation*, vol. 3, no. 2, p. 100132, June 2024, doi: 10.1016/j.multra.2024.100132.

[17] A. F. H. Marsandy, M. N. Hayati, and M. Fauziyah, "Klasterisasi Prevalensi Stunting Menggunakan K-Prototype pada Data Campuran," *metik. j.*, vol. 8, no. 2, pp. 48–54, Dec. 2024, doi: 10.47002/metik.v8i2.824.

[18] A. Wijayanto, Y. K. Suprapto, and D. P. Wulandari, "Clustering on Multidimensional Poverty Data using PAM and K-prototypes Algorithm : Case Study: Jambi Province 2017," in *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Aug. 2019, pp. 210–215. doi: 10.1109/ISITIA.2019.8937130.

[19] H. Hernández, E. Alberdi, A. Goti, and A. Oyarbide-Zubillaga, "Application of the k-Prototype Clustering Approach for the Definition of Geostatistical Estimation Domains," *Mathematics*, vol. 11, no. 3, p. 740, Feb. 2023, doi: 10.3390/math11030740.

[20] B. Islam, T. I. Ibrahim, T. Wang, M. Wu, and J. Qin, "Current trends in household food insecurity, dietary diversity, and stunting among children under five in Asia: a systematic review," *J Glob Health*, vol. 15, p. 04049, Jan. 2025, doi: 10.7189/jogh.15.04049.

[21] M. Batool *et al.*, "Relationship of stunting with water, sanitation, and hygiene (WASH) practices among children under the age of five: a cross-sectional study in Southern Punjab, Pakistan," *BMC Public Health*, vol. 23, no. 1, p. 2153, Nov. 2023, doi: 10.1186/s12889-023-17135-z.

[22] S. Kishore, T. Thomas, H. Sachdev, A. V. Kurpad, and P. Webb, "Modeling the potential impacts of improved monthly income on child stunting in India: a subnational geospatial perspective," *BMJ Open*, vol. 12, no. 4, p. e055098, Apr. 2022, doi: 10.1136/bmjopen-2021-055098.

[23] I. Siramaneerat, E. Astutik, F. Agushybana, P. Bhumkittipich, and W. Lamprom, "Examining determinants of stunting in Urban and Rural Indonesian: a multilevel analysis using the population-based Indonesian family life survey (IFLS)," *BMC Public Health*, vol. 24, no. 1, p. 1371, May 2024, doi: 10.1186/s12889-024-18824-z.

[24] A. Karimzadeh, S. Sabeti, and O. Shoghli, "Optimal Clustering of Pavement Segments Using K-Prototype Algorithm in a High-Dimensional Mixed Feature Space," *Journal of Management in Engineering*, vol. 37, no. 4, p. 04021022, July 2021, doi: 10.1061/(ASCE)ME.1943-5479.0000910.

[25] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes Clustering," *J. of Classification*, vol. 18, no. 1, pp. 35–55, Jan. 2001, doi: 10.1007/s00357-001-0004-3.

[26] I. Herdiana, M. A. Kamal, Triyani, M. N. Estri, and Renny, "A More Precise Elbow Method for Optimum K-means Clustering," Feb. 09, 2025, *arXiv*: arXiv:2502.00851. doi: 10.48550/arXiv.2502.00851.

[27] D. R. Quinthara, Abd. C. Fauzan, and M. M. Huda, "Penerapan Algoritma K-Modes Menggunakan Validasi Davies Bouldin Index Untuk Klasterisasi Karakter Pada Game Wild Rift," *JSCE*, vol. 4, no. 2, pp. 123–135, July 2023, doi: 10.61628/jsce.v4i2.802.

[28] Z. Jia and L. Song, "Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–13, July 2020, doi: 10.1155/2020/5143797.

[29] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *Int. J. Onl. Eng.*, vol. 19, no. 04, pp. 174–182, Apr. 2023, doi: 10.3991/ijoe.v19i04.37059.

[30] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Benchmarking validity indices for evolutionary K-means clustering performance," *Sci Rep*, vol. 15, no. 1, p. 21842, July 2025, doi: 10.1038/s41598-025-08473-6.

[31] S. Lubbe, "Bootstrapping Cluster Analysis Solutions with the R Package ClusBoot," *Austrian Journal of Statistics*, vol. 53, no. 3, pp. 1–19, 2024, doi: 10.17713/ajs.v53i3.1169.

[32] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, no. 95, pp. 2837–2854, 2010.