Mathematical Journal of Modelling and Forecasting
Vol. 2, No. 2, December 2024, pp. 1-8

Application of the K-Means Clustering Algorithm to the Case of Stunting Risk Families in Districts/Cities of West Sumatra Province in 2023

Widiyanti[1,*] and Fadhilah Fitri[1]

[1] Departement of Statistic, Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang, Indonesia
[*]wayday045@gmail.com

Abstract. Stunting is one of the indicators of chronic nutritional status that has a long-term effect on child growth; the main contributing factors are households that do not have access to clean drinking water, proper sanitation facilities, and other factors. The adverse effects experienced by stunted children are reduced cognitive ability, learning ability, decreased endurance, and can lead to new diseases such as diabetes, heart disease, and many other diseases. This study uses the K-Means Cluster method to group the Regency / City of West Sumatra Province in 2023 regarding cases of stunting risk families. K-Means Cluster analysis is an analysis used to group data based on similar features or characteristics. From the results of the study, it can be concluded that the clustering of 19 regencies/cities in West Sumatra Province resulted in 2 groups (clusters): cluster 1 consists of 12 regency/city members, and cluster 2 consists of 7 regency/city members. The characteristic results obtained from each cluster formed are cluster 2 shows families with better conditions than cluster 1.

Keywords: Cluster, Families Risk Stunting, K-Means.

1 Introduction

Stunting indicates chronic nutritional status that has long-term effects on child growth [1]. Various factors cause the problem of stunting. The leading causes are households that do not have access to clean drinking water, exclusive breastfeeding, low birth weight babies ≤ 2,500 grams born safely, and households that do not have proper sanitation facilities [2]. The adverse effects of stunting in the long term can reduce cognitive abilities and learning abilities, decrease endurance, and lead to new diseases such as diabetes, heart disease, and many other diseases [3].

Prevalence is one of the biggest nutritional problems (stunting) in toddlers in Indonesia [4]. In West Sumatra Province in 2023, the prevalence of stunting has decreased to 23.6% from 25.3% in 2022 [5]; the basis for setting the stunting prevalence target (Very short & short) in toddlers is Regional Regulation No.6 of 2021 concerning RPJMD 2021-2026. The West Sumatra Provincial Health Office Strategic Plan for 2021-2026 with a target of reducing the prevalence of stunting toddlers to 16% by 2023 [5], and the target of reducing the prevalence of stunting set by the government, namely reducing the stunting rate to 14% by 2024 [6].

Presidential Regulation Number 72 of 2021 concerning the acceleration of stunting reduction is one of the government's commitments to accelerate stunting reduction. As one of the actions in accelerating stunting reduction, the government needs to provide specific interventions (direct causes of stunting) or sensitive interventions (indirect causes of stunting) to families at risk of stunting [7]. It is necessary to cluster districts/cities based on family factors at risk of stunting to make it easier to provide interventions. The K-Means Clustering approach will assess districts based on stunting risk factors.

K-Means Clustering analysis is a statistical method used as an analytical tool to solve a problem. The problem in this study is to group data into several groups (clusters) based on similar characteristics. Clusters are groups of data that have certain similarities. The center point of a cluster is called the centroid. The

centroid is the average presentation of all data in the cluster or in K-Means; each cluster is represented by a centroid [8].

Several researchers use K-Means Clustering, one of which is a research conducted by Dayla, et al. In this study, it can be concluded from the analysis results that the clustering of villages or villages is at risk of stunting. The research here refers to 3 clusters: low, medium, and high. The dataset used was 71 data on urban villages or villages that indicated the risk of stunting. The results showed that there were 32 urban villages or villages that had low risk, with a percentage of 45.07%; moderate risk, as many as 36 urban villages/villages, with a percentage of 50.70%; and high risk, as many as 3 urban villages/villages with a percentage of 4.23% [7]. Windham conducted another study. This study aims to cluster 50 toddlers in Karang Songo village using the K-Means method; the clustering results obtained are as many as 5 groups (clusters), namely cluster 1 - malnutrition, cluster 2 - undernutrition, cluster 3 - good nutrition, cluster 4 - overnutrition and cluster 5 - obesity [9].

Based on the description above, the author is interested in conducting research related to K-Means Clustering Analysis on the case of Stunting Risk Families per Regency / City in West Sumatra Province in 2023. The aim is to group and identify the characteristics of the clusters formed and see which cluster is better.

2 Research Methods

2.1 Data Sources and Research Variables

The data used in this study are secondary data obtained from representatives of the National Population Family Planning Agency (NPFPA) of West Sumatra Province. The data used are data on verification and validation of Stunting Risk Families (verbal FRS) in the Regency / City of West Sumatra Province in 2023. The data used in this problem have four variables.

Table 1. Research Variables

Variable	Symbol	Unit
Drinking Water Source	X1	Family
Latrine	X2	Family
PUS 4 T	X3	Family
Not a Modern Family Planning Participant	X4	Family

2.2 Stages of Analysis

The following are the stages of analysis in the context of applying cluster analysis with the K-Means approach:

2.3 Descriptive Analysis

Descriptive analysis is a statistical technique used to analyze data by describing or describing the data collected as it is without intending to make generalizations [10]. In this study, descriptive analysis was carried out to determine the basic characteristics of the Stunting Risk Family (KBS) verbal data, the minimum, maximum value, the average, and so on.

Calculation with K-Means Algorithm

K-means clustering is a non-hierarchical clustering method that divides data into one or more groups based on k-specified groups. The method is based on the initial value of the center point (centroid), where the initial centroid value affects the next centroid value and the determination of the next cluster value. The calculation is stopped if the previous cluster has the same pattern as the following one.

Here are the steps for completion:

1. Assumption Check

a. Non-Multicollinearity

Multicollinearity is the relationship between independent variables, both positively and negatively related. To test the presence or absence of multicollinearity, the correlation coefficient between the independent variables can be examined. If the correlation coefficient value > 0.8, then there is a multicollinearity problem between one or more independent variables [11].

$$r_{xy} = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{\{(n\sum x^2)(\sum x^2)\}\{(n\sum y^2)(\sum y^2)\}}} \quad (1)$$

With r_{xy} pearson correlation coefficient, n sample size, x first variable and y second variable.

b. Representative Sample

A representative sample is a selected sample that can represent the study's population. In this case, the Kaiser-Mayer-Olkin (KMO) test is used to measure the adequacy of sampling both overall and for each indicator. The KMO test results show that if the KMO value ranges from 0.5 to 1, then the sample can be said to represent the population or is representative [12].

$$KMO = \frac{\sum_i \sum_{i\neq j} r^2{}_{ij}}{\sum_i \sum_{i\neq j} r^2{}_{ij}+\sum_i \sum_{i\neq j} a^2{}_{ij}} \quad (2)$$

With r_{ij}^2 simple correlation coefficient between variables i and j, a a_{ij}^2 partial correlation coefficient between variables i and j.

2. Data standardization

The standardization process is carried out if there is a significant difference in unit size between the variables studied. Significant differences in units can result in invalid calculations in cluster analysis. The most common form of standardization is converting each variable into a standard score (also known as a z score) by subtracting the mean and dividing it by the standard deviation for each variable [12].

$$z = \frac{X-\mu}{\sigma} \quad (3)$$

The scale() function in Rstudio can standardize z scores. This process converts each raw data score into a standardized value with a mean of 0 and a standard deviation of 1 [12].

3. Determining the Optimal Number of Clusters

This study determined the optimal number of clusters using the Silhouette method. The Silhouette method plays a role in determining the quality and strength of the clusters formed. For a point x_i, the average distance to all points in the same cluster is first calculated. This value is set to a_i [13].

$$a_i = \frac{1}{|A|-1} \sum j \in_{A,j\neq i} d(i,j) \quad (4)$$

Where j is another observation in clusters A and A(ij) is the distance between observations i and j. Then, for each cluster that does not contain x_i, the average distance of x_i to all data points in each cluster are calculated. This value is set to b_i.

$$d(i,C) = \frac{1}{|A|}\sum j \in C\ d(i,j) \quad (5)$$

Where d(i,C) is the average distance of observation i to all observations in another cluster C where A≠C.

$$b_i = \min C \neq A\ (i,j) \quad (6)$$

Using these two values, the silhouette coefficient of a point is estimated. The average of all silhouettes in the data set is called the average silhouette width for all points in the data set. To evaluate the quality of a clustering, one can calculate the average silhouette coefficient of all points. The value of the silhouette result lies in the range of values - 1 to 1. The greater the value of the silhouette coefficient close to 1, the better the grouping of data in the cluster. Conversely, if the sihoutte coefficient approaches - 1, the worse the clustering of data in a cluster [14][15].

$$s_i = \frac{b_i - a_i}{\max\ (a_i.b_i)} \quad (7)$$

4. Forming clusters

   Grouping objects into clusters that match the characteristics of each object. Objects are combined based on the level of similarity of object characteristics measured using a similarity measure.

5. Profiling the clusters

   The cluster profiling stage describes the characteristics of each cluster to explain the differences between clusters. each cluster formed will have its own characteristics so that the differences between one cluster and another can be seen. Profiling focuses on describing the characteristics of the cluster after the cluster is identified [12].

3 Results and Discussion

3.1 Descriptive Analysis

Table 2 reflects the disparities in access to basic resources, health behaviors, and participation in family health programs across regions in West Sumatra in 2023. Areas with minimum values tend to show better access or behavior, while maximum values illustrate areas that require more attention in handling the risk of stunting. In general, interventions can focus on increasing access to safe drinking water, sanitation facilities (latrines), counseling related to 4T risks, and encouraging the adoption of modern family planning to reduce family risk of stunting.

Table 2. Descriptive Value of Variables

Variable	Means	Minimum	Maximum
DWS	1.369	1.023	1.984
Latrine	1.331	1.027	2.309
PUS 4T	1.419	1.205	1.638
NMFPP	1.346	1.136	1.546

3.2 Calculation with K-Means Algorithm

Non-Multicollinearity

From Table 4, it can be seen that the correlation value between variables exceeds 0.8; this suspects multicollinearity between variables. After handling 2 times by transforming the data, there is still a suspicion of multicollinearity between the PUS 4 T variable and NMPP. So, the analysis will continue.

Table 4. Correlation Value of Variables

Variable	DWS	Latrine	PUS 4 T	NMPFPP
DWS	1	-0.2718334	-0.7844792	-0.7563520
Latrine	-0.12718334	1	0.2198608	0.1571972
PUS 4 T	-0.77844792	0.2198608	1	0.9732846
NMFPP	-0.7563520	0.1571872	0.9732846	1

Representative Sample

Table 3. KMO Value of Variables

Variable	KMO
DWS	0.91
Latrine	0.51
PUS 4 T	0.61
NMFPP	0.62

Each variable has a KMO value> 0.5. This shows that the sample used in the case of Families at Risk of Stunting (FRS) represents the population or representative sample, so the analysis can be continued.

Data Standardization

Data standardization is done with the scale() function in Rstudio. Data standardization is done because of differences in data units or scales between variables. This process converts each raw data score into a standardized value with a mean of 0 and a standard deviation of 1.

Table 5. Standardized Data

No	Regency / City	DWS	Latrine	PUS.4T	NMPFPP
1	Pesisi Selatan	-0.60356	0.263622	1.151.573	0.916299
2	Solok	-0.33766	0.481828	0.794923	0.561844
3	Sijunjung	-0.7177	0.004987	0.199923	-0.12662
4	Tanah Datar	-0.75303	0.040466	0.573852	0.403189
5	Padang Pariaman	-0.71239	0.269507	0.73467	1.041.384
6	Agam	-0.48561	0.030722	1.044.983	1.269.026
7	Lima puluh Kota	-0.60051	0.42248	0.735501	0.565528
8	Pasaman	-0.4584	0.483173	0.469728	0.535142
9	Kepulauan Mentawai	-0.44872	-0.58223	-123.751	-0.83711
10	Dharmasraya	-0.8193	-0.34055	0.061445	-0.08938
11	Solok Selatan	-0.81769	-0.0356	-0.28116	-0.43474
12	Pasaman Barat	-0.59887	0.23803	0.891462	1.246.521
13	Padang City	-100.688	0.186982	1.602.721	1.523.135
14	Solok City	0.938634	3.440.261	-121.073	-138.059
15	Sawahlunto City	1.787.212	-106.989	-151.557	-177.982
16	Padang Panjang City	1.247.523	-0.90939	-156.883	-147.327

17	Bukit Tinggi City	1.343.206	-0.91808	-0.85282	-0.74987
18	Payakumbuh City	1.253.501	-102.244	-0.59749	-0.39774
19	Pariaman City	1.790.226	-0.98387	-0.99668	-0.79293

Determining Clusters

Table 6. Cluster Analysis Results

Districts/Cities	cluster
Pesisir Selatan	1
Solok	1
Sijunjung	1
Tanah Datar	1
Padang Pariaman	1
Agam	1
Lima Puluh Kota	1
Pasaman	1
Kepulauan Mentawai	2
Dharmasraya	1
Solok Selatan	1
Pasaman Barat	1
Padang City	1
Solok City	2
Sawahlunto City	2
Padang Panjang City	2
Bukittinggi City	2
Payakumbuh City	2
Pariaman City	2

Based on table 6 above, it can be seen that there are 2 clusters, namely clusters 1 and 2. The members of each cluster are, in cluster 1 has 12 Regency / City members consisting of Kab. Pesisir Selatan, Kab. Solok, Kab. Sijunjung, Kab. Tanah Datar, Kab. Padang Pariaman, Kab. Agam, Kab. Limapuluh Kota, Kab. Pasaman, Kab. Dharmasraya, Kab. South Solok, Kab. West Pasaman, and Kota Padang. While cluster 2 has 7 regencies/cities as members consisting of Mentawai Islands Regency, Solok City, Sawahlunto City, Padang Panjang City, Bukittinggi City, Payakumbuh City, and Pariaman City.

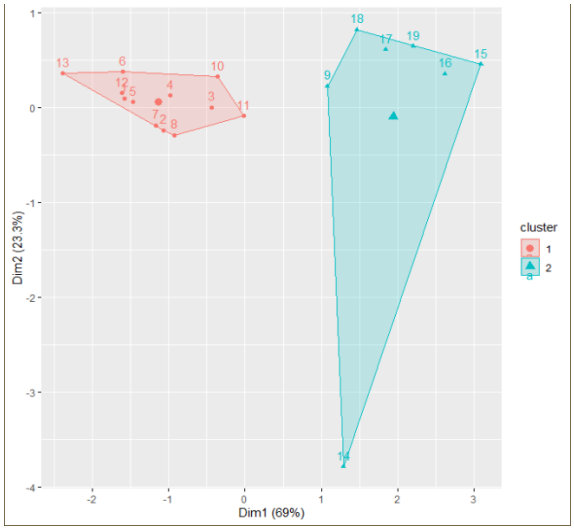

Fig 1. Cluster plot

Figure 1 displays the final clustering results in plot form. The figure shows that cluster 1 and cluster 2 are well separated with no overlap, indicating that the clustering method successfully identified two significantly different groups.

Cluster Profiling

1. Characteristics

Table 7. Cluster Characteristics

Cluster	DWS	Latrin	PUS 4T	NMFPP
1	1.142641	1.379479	1.509918	1.433565
2	1.757253	1.247892	1.263317	1.225729

Based on Table 7, Cluster 1 scores lower on indicators of drinking water sources (DWS), latrine access, PUS 4T risk, and modern family planning participation (NMFPP). This indicates limited access to essential resources and health services, thus requiring interventions to improve the availability of clean water, sanitation, health education, and family planning participation. Meanwhile, Cluster 2 scores better on all indicators, describing families with superior access to resources and better health conditions. The next effort is to maintain the quality of resources and optimize health programs for sustainability.

2. Cluster Profiling

Cluster 1: Families with limited access to essential resources and health services.
Cluster 2: Families with better resource access and relatively superior health conditions.

4 Conclusion

Based on the discussion and analysis above, the Family Risk of Stunting (FRS) served in the Regency / City of West Sumatra Province in 2023 is grouped into 2 clusters. cluster 1 has 12 Regency / City members consisting of Kab. Pesisir Selatan, Kab. Solok, Kab. Sijunjung, Kab. Tanah Datar, Kab. Padang Pariaman, Kab. Agam, Kab. Limapuluh Kota, Kab. Pasaman, Kab. Dharmasraya, Kab. South Solok, Kab. West Pasaman, and Kota Padang. Cluster 2 has 7 regencies/cities, consisting of Mentawai Islands Regency, Solok City, Sawahlunto City, Padang Panjang City, Bukittinggi City, Payakumbuh City, and Pariaman City.

The two clusters formed have their characteristics, of which cluster 2 shows families with better conditions in general with the availability of good drinking water sources, low 4T risk, and better modern family planning participation rates than cluster 1. However, inadequate sanitation facilities, such as latrines, are still

challenging to improve. So it can be concluded that cluster 1 with members of 12 districts/cities, namely Pesisir Selatan Regency, Solok Regency, Sijunjung Regency, Tanah Datar Regency, Padang Pariaman Regency, Agam Regency, Limapuluh City Regency, Pasaman Regency, Dharmasraya Regency, South Solok Regency, West Pasaman Regency, and Padang City, has families that may be vulnerable to stuting.

References

[1] N. Indrayani and M. Nurtyas, "Strategi Komunikasi Dalam Pendampingan Keluarga Risiko Stunting Di Wilayah Kalurahan Wedomartani Kapanewon Ngemplak," *J. Kesehat. Madani ...*, vol. 7, no. 1, pp. 16–20, 2024, [Online]. Available: http://www.jurnalmadanimedika.ac.id/JMM/article/view/382%0Ahttps://www.jurnalmadanimedika.ac.id/JMM/article/download/382/221

[2] S. 'Aina Salsabila, T. Widiharih, and S. Sudarno, "METODE K-HARMONIC MEANS CLUSTERING DENGAN VALIDASI SILHOUETTE COEFFICIENT (Studi Kasus : Empat Faktor Utama Penyebab Stunting 34 Provinsi di Indonesia Tahun 2018)," *J. Gaussian*, vol. 11, no. 1, pp. 11–20, 2022, doi: 10.14710/j.gauss.v11i1.34003.

[3] A. Subayu, "Penerapan Metode K-Means Untuk Analisis Stunting Gizi Pada Balita: Systematic Review," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 2, no. 1, 2022, doi: 10.20885/snati.v2i1.18.

[4] R. F. J, N. Huljannah, and T. N. Rochmah, "Stunting Prevention Program in Indonesia: A SYSTEMATIC REVIEW," *Media Gizi Indones.*, vol. 17, no. 3, pp. 281–292, 2022, doi: 10.20473/mgi.v17i3.281-292.

[5] T. Pipit Muliyah, Dyah Aminatun, Sukma Septian Nasution, Tommy Hastomo, Setiana Sri Wahyuni Sitepu, "Laporan Akuntabilitas Kinerja Pemerinatah," *J. GEEJ*, vol. 7, no. 2, 2020.

[6] Kemenkes RI, "Rencana Strategis (Renstra) Kementerian Kesehatan Tahun 2020-2024 (revisi 2022)," *Kementerian Kesehatan Republik Indonesia*. 2022.

[7] Dayla May Cytry, S. Defit, and G. Nurcahyo, "Penerapan Metode K-Means dalam Klasterisasi Status Desa terhadap Keluarga Beresiko Stunting," *J. KomtekInfo*, vol. 10, no. 3, pp. 122–127, 2023, doi: 10.35134/komtekinfo.v10i3.423.

[8] E. Sartika, S. Murniati, A. Binarto, and E. Habinuddin, "Penerapan K-Means Cluster dan Evaluasi Clustering pada Pesebaran Kasus Covid-19," *Stat. J. Theor. Stat. Its Appl.*, vol. 22, no. 2, pp. 147–156, 2022, doi: 10.29313/statistika.v22i2.1229.

[9] W. M. P. Dhuhita, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita," *J. Inform.*, vol. 15, no. 2, pp. 160–174, 2015.

[10] Soegiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. 2011.

[11] W. J. Corlett and D. J. Aigner, *Basic Econometrics.*, vol. 82, no. 326. 1972. doi: 10.2307/2230043.

[12] J. F. Hair, W.C. Black, B.J.Babin, R.E.Anderson, and R.L.Tatham, "Multivariate Data Analysis. hair.pdf." p. 761, 2019. [Online]. Available: https://www.drnishikantjha.com/papersCollection/Multivariate Data Analysis.pdf

[13] B. Indikator and I. P. M. Dengan, "Analisis perbandingan silhouette coefficient dan metode elbow pada pengelompokkan provinsi di indonesia berdasarkan indikator ipm dengan k-medoids 1,2,3," vol. 13, pp. 13–24, 2024, doi: 10.14710/j.gauss.13.1.13-24.

[14] W. Sartika, S. Suryono, and A. Wibowo, "Information System for Evaluating Specific Interventions of Stunting Case Using K-means Clustering," *E3S Web Conf.*, vol. 202, 2020, doi: 10.1051/e3sconf/202020213003.

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.