

CLUSTERING OF TODDLER DATA AT PUSKESMAS NANGGALO USING THE K-MEANS ALGORITHM**M. Alif Alfiansyah¹, Dony Novalendry², Syafrijon³, Yeka Hendriyani⁴**

Department Electronic Engineering, Universitas Negeri Padang, Padang, Padang, Indonesia

Abstract (English)

Indonesia is still facing a double nutritional problem in children under five, namely malnutrition and overnutrition, which often occurs in children aged 0-59 months, an important phase in physical and mental development. This study aims to cluster data on the nutritional status of toddlers at the Nanggalo Health Center, Nanggalo District, Padang City, using clustering methods in the data mining process. Using the K-Means algorithm, toddler data is clustered based on body weight (BW) and height (TB) indicators. The Nanggalo Community Health Center was chosen as the research site because the available nutrition data is complete and representative, thus providing a comprehensive picture of the nutritional status of children under five in the area. The clustering results in three main categories: underweight, normal and overweight, which makes it easier to identify groups of under-fives who need more attention. This information is expected to be accessible to parents and health workers to support better decision-making on under-five nutrition management. With this approach, this study is expected to contribute to improving the quality of nutrition services at the Nanggalo Health Center and support the creation of a healthy generation that is accustomed to a healthy lifestyle.

Article History*Submitted: 19 Januari 2025**Accepted: 28 Januari 2025**Published: 29 Januari 2025***Key Words**

Data Mining, K-Means
Clustering, Clustering Toddler
Data, Nanggalo Health Center.

I. INTRODUCTION

Children aged 0-59 months, classified as toddlers, are in a phase of rapid growth and development. During this time, the role of parents in providing knowledge and good nutrition is very important. Good nutrition has a big impact on toddlers, because malnutrition can cause physical and psychological problems in children.

In Indonesia, attention to the growth and development of children under five, especially in terms of balanced nutrition, is still an important issue. The country faces a double nutrition problem: undernutrition and overnutrition. The Ministry of Health's Basic Health Research in 2013 showed that 1 in 3 children are stunted due to chronic malnutrition, which can impair growth and development. Meanwhile, the prevalence of overnutrition is also increasing.

In today's digital era, data and information play a crucial role in various aspects of life, from business to scientific research. Data, which is a description of reality, concepts, or instructions in formal form, can be processed into useful information through various techniques. In the context of under-five nutrition, data mining is an effective method for extracting valuable information from large amounts of available data. Data mining techniques such as estimation, association, classification, clustering, and prediction are used to find patterns hidden in data.

This research uses clustering techniques to group nutritional data of under-fives based on their nutritional status. Clustering is a method of grouping data into clusters that are similar to each other. The K-means algorithm was chosen in this study because of its ability to group data into k clusters based on a high degree of similarity among cluster members.

The Nanggalo Community Health Center in Nanggalo Sub-district, Padang City, West Sumatra, was chosen as the research location because it has complete and representative data on the nutritional status of children under five in the area. In addition, the Puskesmas has various nutrition programs that can be analyzed to improve the quality of under-five nutrition services. This study was conducted from March to June 2024.

By processing data at the Nanggalo Health Center using the clustering method, toddler nutrition data can be grouped based on the nutritional characteristics of each child, such as body weight (BW) and height (TB). This method is expected to make a significant contribution

to the clustering of under-five data, so that the resulting information can be easily accessed by parents, Puskesmas, and Posyandu. Thus, it is expected to support better decision-making regarding children's nutritional patterns, as well as realizing a healthy generation that is accustomed to a healthy lifestyle in the future.

II. METHOD

This study uses data on the growth of toddlers in the Nanggalo sub-district of Padang city reported by the Nanggalo health center in 2023. The data includes toddler growth data such as toddler name, gender, weight (Kg), height (Cm), age (Year), head circumference (Cm), and upper arm circumference (Cm). The tools used are Orange 3.36.2 which is used for processing data, python programming language which is used to run the K-Means Clustering algorithm and tableau which is used to visualize data.

A. Location and Time of Research

This research was conducted at the Nanggalo Health Center which is located on Jalan Padang Perumnas Siteb, Surau Gadang Village and this research was carried out in the time span of March to June 2024.

This research was conducted at the Nanggalo Health Center which is located on Jalan Padang Perumnas Siteb, Surau Gadang Village and this research was conducted from March to June 2024.

B. Population and Sample

The population in this study included all toddlers registered at the Nanggalo Health Center. The criteria for toddlers used as research samples are: (1) Having complete data; (2) Routinely conducting health checks at the posyandu; (3) Not moving areas during the study period.

C. Independent Variables

Independent variables in this study include nutritional data and physical characteristics of toddlers from the Nanggalo Health Center. Details of the independent variables include: (a) age in months, (b) gender, (c) weight in kilograms, (d) height in centimeters, (e) head circumference in centimeters, and (f) upper arm circumference in centimeters. These variables were used to determine their effect on the nutritional status of children under five.

D. Dependent Variable

In this study, the dependent variable is the result of the clustering process that shows the nutritional status grouping of toddlers. This variable includes the Nutrition Cluster, the result of clustering using K-Means, which groups toddlers based on similar nutritional characteristics. Each cluster reflects a different nutritional status category, such as good nutrition, under nutrition, or over nutrition.

E. Research Instruments

This study used health record forms, digital scales, stadiometers, tape measures, and data analysis software to ensure accurate and reliable data[7]. These instruments support the analysis using the K-Means Clustering method, which is expected to produce valid findings and be useful for improving the under-five nutrition program at the Nanggalo Health Center.

F. Data Collection Process

The author made direct observations to the Nanggalo Health Center to collect data related to the problem discussed, and documented the data directly for further analysis.

G. Data Analysis Procedure

The data analysis procedure in this study involves data collection, preprocessing, implementation of the K-Means algorithm, evaluation of results, interpretation, and visualization. By following these steps, the research is expected to produce valid and useful findings to improve the under-five nutrition program at the Nanggalo Health Center[8].

H. Clustering Technique Using K-Means Algoritma

K-Means is a data analysis method in Data Mining that performs unsupervised modeling and clusters data by partitioning. Data is grouped into several groups, where each group has similar characteristics, while different from other groups. The goal is to minimize the differences between data within one group and maximize the differences with other groups. In general, the K-Means method uses the following algorithm[9]:

1. Determine k as the number of clusters in the form. Determination of the number of clusters k is done with several factors such as theoretical and conceptual considerations proposed to determine how many clusters.
2. Randomly generate k initial centroids (cluster center points). To determine the initial centroid is done randomly from several objects available as many as k clusters, to calculate the centroid of the next i-th cluster, using the following formula:

$$v = \frac{\sum_{i=1}^n x_i}{n} ; i = 1, 2, 3, \dots n$$

Where:

- v: centroid in the cluster
 - Xi: i-th object
 - n: the number of objects or the number of objects that are members of the cluster.
3. Calculate the distance of each object to each centroid of each cluster. Then calculate the distance between the object and the centroid, in this study using Euclidian Distance.

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} ; i = 1, 2, 3, \dots n$$

Where:

- x1: i-th object x
 - y: i-th power y
 - n: the number of objects
4. Allocate each object to the closest centroid.
 5. Perform iteration, then determine the new centroid position by using equation.
 6. Repeat step 3 if the new centroid position is not the same. The point merging process is done by comparing the matrix of the set of tasks in the previous iteration with the matrix of the set of tasks in the current iteration. If the results are the same then the k-means cluster analysis algorithm has converged, but if they are different then it has not converged so it is necessary to do the next iteration[10].

III. RESULTS

A. Data Integration

The data obtained was sourced from the official report of toddler growth data of Nanggalo Sub-district from the Nanggalo Health Center Agency of Padang City, West Sumatra Province. This data was collected and integrated into a single .csv file to facilitate further processing and analysis. The data integration process was carried out by combining several datasets from various time periods and different types of measurements. This step is important to ensure that

the data used truly reflects the growth of children under five in the Nanggalo sub-district of Padang City.

B. Testing

The testing stage is a crucial step in our analysis to ensure the accuracy and reliability of the results obtained from clustering and association rule mining techniques. In this process, the data cleaning process is carried out by removing attributes such as KK number, NIK, 3rd child, date of birth, birth weight, father's name, father's NIK, address, RT, RW, and measurement date. In addition to cleaning data on attributes/variables in the initial dataset, data cleaning was also carried out on data with missing values. The preprocessing process was carried out using orange software as shown in Figure 1.

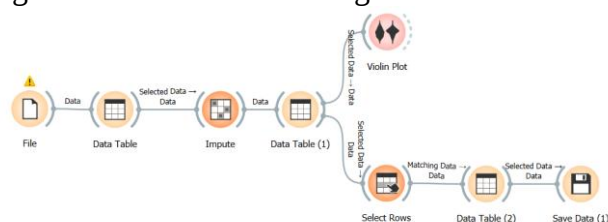


Figure 1 Workflow in Orange 3.36.2 for Preprocessing Data

In this research, data clustering is done using Google Colab to process toddler data containing information such as weight, length, and age. The input Excel file was converted using the pandas library so that it could be processed by Python. The author checked the data format through `df.info()` to ensure the data was ready for processing, then displayed initial visualizations such as graphs of length, weight, and age.

The data was then standardized using StandardScaler so that each feature had a mean of 0 and a standard deviation of 1. After that, the K-Means algorithm was applied to divide the data into four clusters. The clustering results were added to the DataFrame and evaluated using metrics such as Inertia, Silhouette Score, and Davies-Bouldin Score to assess the clustering quality.

The clustering results were visualized with scatter plots using matplotlib and seaborn, with each cluster assigned a different color. The author also calculated the number of members in each cluster, with the second cluster having the highest number of members. Finally, the visualization of the clustering results was made interactive using Tableau, and the dashboard was published to Tableau Public to facilitate further analysis. **ORCID is compulsory for all authors.**

*Table I
Example of clustering results in table form*

Nama Anak	Berat Badan(Kg)	Tinggi Badan(Cm)	Umur(Tahun)	Cluster
Alitah Permata	9	80	2	1
Qiani Ani	11.5	89	2	1
Fathir Pratama	12	93	3	0
Umaiza Akiva	12	101	4	0
Felicia	10	92	5	0

C. Visualization of Clusters

This study uses the K-Means algorithm to group toddler nutrition data at the Nanggalo Health Center into three clusters, which are presented in the form of tables and graphs. In the graph and table there are 3 clusters consisting of several colors as follows:

1) *First Cluster: The first cluster is a cluster that illustrates that toddlers who are in this cluster include thin / short toddlers. Toddlers in this cluster need more supervision from the puskesmas so that there is no malnutrition in toddlers in their growth. This first cluster is denoted by the number 0 in the table and is depicted in purple on the graph.*

2) *Second Cluster: The second cluster is a cluster that illustrates that toddlers who are in this cluster include normal and healthy toddlers. Toddlers in this cluster do not need more supervision from the puskesmas because they are in a healthy condition so that it will not interfere with their growth. This second cluster is denoted by the number 1 in the table and depicted in blue on the graph.*

3) *Third Cluster: The third cluster is a cluster that illustrates that toddlers who are in this cluster include fat / tall toddlers. Toddlers in this cluster do not yet need more supervision from the puskesmas but if their growth leads to excess then supervision by the puskesmas is needed. This third cluster is denoted by the number 2 in the table and depicted in yellow on the graph.*

BB/TB

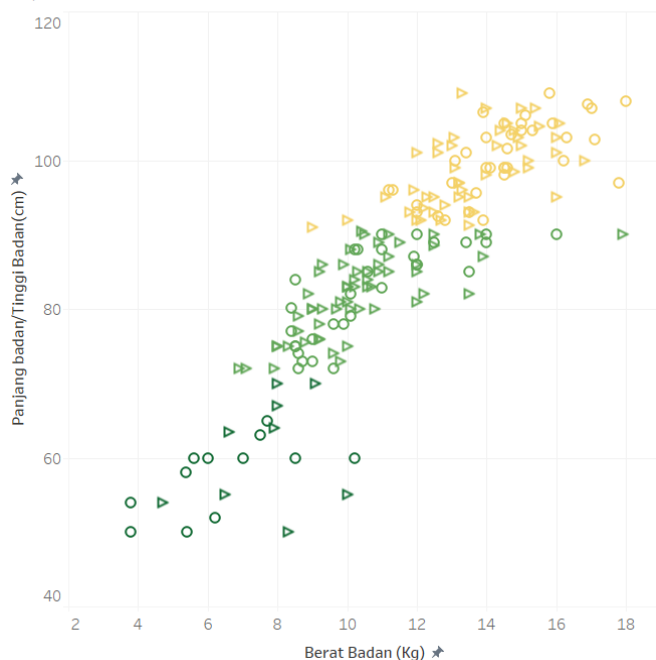


Figure 2 Scatter plot graph based on BB/TB

The figure above shows the results of K-Means clustering on toddler data at Puskesmas Nanggalo, based on weight and height in a scatter plot graph. Toddlers are grouped into three clusters: dark green for short and thin toddlers (cluster 1), light green for healthy and normal toddlers (cluster 2), and yellow for fat and tall toddlers (cluster 3). Each dot in the graph has a different shape to indicate gender, a circle for males and a triangle for females. Toddlers in cluster 1 and cluster 3 require special attention from puskesmas to prevent malnutrition and obesity, while those in cluster 2 only require regular check-ups.

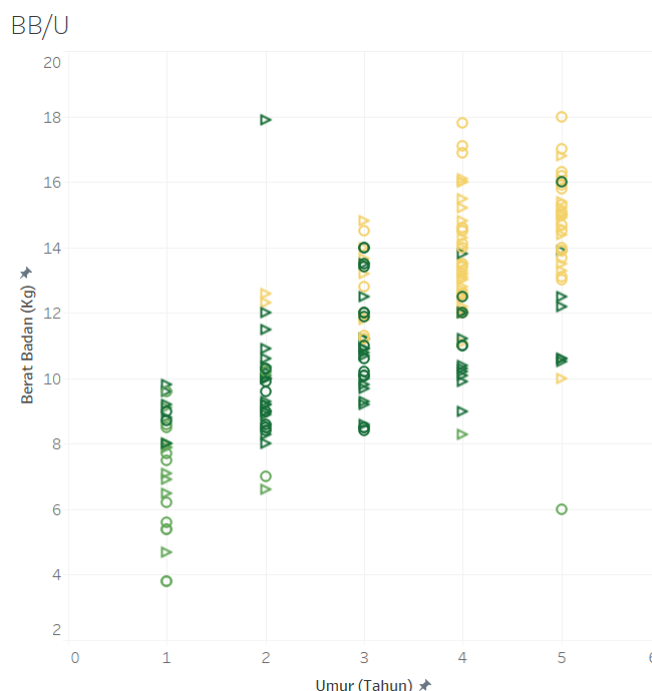


Figure 3 Scatter plot graph based on BB/U

The figure above shows the results of K-Means clustering on Nanggalo Health Center toddler data based on weight and age parameters in a scatter plot graph. Toddlers are grouped into three clusters: yellow for underweight toddlers (cluster 1), light green for healthy and normal toddlers (cluster 2), and dark green for obese toddlers (cluster 3). The shape of the dots on the graph indicates gender, with circles representing males and triangles females. Toddlers in cluster 1 weigh around 4 to 10 kg and are between 1 and 2 years old and require special attention to prevent malnutrition. Cluster 2 consists of healthy toddlers weighing 8 to 14 kg and aged 2 to 5 years who only require regular check-ups. Meanwhile, cluster 3 consists of obese toddlers weighing 12 to 18 kg and aged 3 to 5 years who require special attention to prevent obesity.

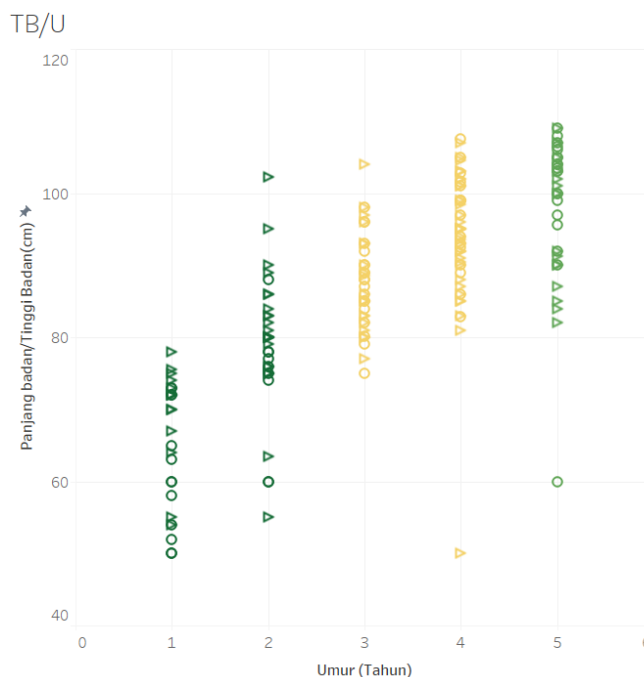


Figure 4 Scatter plot graph based on TB/U

Figure 24 shows the results of K-Means clustering on toddler data at Puskesmas Nanggalo based on height and age. Toddlers are grouped into three clusters: dark green for short toddlers (cluster 1), red for healthy toddlers (cluster 2), and light green for tall and obese toddlers (cluster 3), with circles representing males and triangles females. Toddlers in cluster 1 (50-100 cm tall, 1-3 years old) require attention to prevent malnutrition. Cluster 2 (height 80-100 cm, age around 4 years) only requires periodic check-ups, while cluster 3 (height 80-110 cm, age 5-6 years) requires attention to prevent obesity.

IV. CONCLUSION

Based on the results of research using the K-Means Clustering algorithm on toddler nutrition data at the Nanggalo Health Center, it can be concluded that the majority of toddlers need special attention in monitoring their nutritional status. Of the three nutrition indicators used, namely BB/TB (body weight/height), BB/U (body weight/age), and TB/U (body height/age), it was found that most toddlers were in the underweight or obese category, with relatively few toddlers having ideal nutritional status. In the BB/TB indicator, around 40% of under-fives were recorded as underweight or short, requiring interventions to prevent malnutrition, while only around 10% of under-fives were obese. In the BB/U indicator, as many as 85% of children under five are underweight or obese, indicating a significant problem related to weight and age. Meanwhile, in the TB/U indicator, the majority of children under five experienced height problems (shortness), indicating long-term nutritional problems and the risk of stunted growth.

Overall, although some toddlers were in the normal and healthy nutrition category, the results of this study confirm that most toddlers require special attention from the puskesmas, both to prevent malnutrition and to reduce the risk of obesity. Therefore, increased supervision and more intensive intervention programs are needed at Nanggalo Health Center to ensure optimal development and health of under-fives in the future.

ACKNOWLEDGMENT

This research would not have been possible without the support of various parties. The researcher would like to thank the Nanggalo Health Center, Padang City, West Sumatra, for providing the toddler data used in the study. The researcher would also like to express his

deepest gratitude to Mrs. Dr. Yeka Hendriyani, S.Kom., M.Kom. for her invaluable guidance, especially in the preparation of this final project. Thank you also to Mr. Dony Novalendry, S.Kom., M.Kom. and Dr. Syafrijon, S.Pd., M.Kom for their advice and guidance in the preparation of this final project. Finally, the author would like to thank all those who cannot be mentioned one by one who have helped in the preparation of this final project.

REFERENCES

- [1] 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.
- [2] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," *Hello World J. Ilmu Komput.*, vol. 2, no. 1, pp. 20–33, 2023, doi: 10.56211/helloworld.v2i1.230.
- [3] D. Dona and M. Rifqi, "PENERAPAN METODE K-MEANS CLUSTERING UNTUK MENENTUKAN STATUS GIZI BAIK DAN GIZI BURUK PADA BALITA (STUDI KASUS KABUPATEN ROKAN HULU)," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 7, no. 2, pp. 179–191, Jul. 2022, doi: 10.36341/rabit.v7i2.2171.
- [4] C. Zai and T. Komputer, "IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA."
- [5] H. Pohan, M. Zarlis, E. Irawan, H. Okprana, and Y. Pranayama, "Penerapan Algoritma K-Medoids dalam Pengelompokan Balita Stunting di Indonesia," *JUKI J. Komput. dan Inform.*, vol. 3, no. 2, pp. 97–104, 2021, doi: 10.53842/juki.v3i2.69.
- [6] M. Imron, U. Hasanah, and B. Humaidi, "Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 3, no. 1, pp. 12–22, 2020, doi: 10.47738/ijiis.v3i1.3.
- [7] H. Satria Tambunan, Z. Almaida Siregar, A. Perdana Windarto, F. Rizki, and S. Tunas Bangsa, "Penerapan Data Mining Klasifikasi Gizi Bayi Dengan Algoritma Decision Tree C4.5." [Online]. Available: <https://ejurnal.pdsi.or.id/index.php/zahra/index>
- [8] A. K. Wahyudi, N. Azizah, and H. Saputro, "DATA MINING KLASIFIKASI KEPRIBADIAN SISWA SMP NEGERI 5 JEPARA MENGGUNAKAN METODE DECISION TREE ALGORITMA C4.5." [Online]. Available: <https://journal.unisnu.ac.id/JISTER/>
- [9] W. Widyawati, W. L. Y. Saptomo, and Y. R. W. Utami, "Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan," *J. Ilm. SINUS*, vol. 18, no. 1, p. 75, 2020, doi: 10.30646/sinus.v18i1.448.
- [10] E. Irfiani, S. Sulistia Rani, S. Nusa Mandiri Jl Kramat Raya No, and J. Pusat, "Algoritma K-Means Clustering untuk Menentukan Nilai Gizi Balita," vol. 6, no. 4, pp. 17–27, 2018.
- [11] C. Satria and A. Anggrawan, "Aplikasi K-Means berbasis Web untuk Klasifikasi Kelas Unggulan," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 1, pp. 111–124, 2021, doi: 10.30812/matrik.v21i1.1473.
- [12] L. Widyawati and V. Lusiana, "Penerapan K-Means Clustering Untuk Mengelompokkan Data Transaksi Penjualan (Studi Kasus pada Wijaya Hijab)."
- [13] A. Triningsih and H. Supriyono, "APLIKASI DATA MINING BERBASIS WEB MENGGUNAKAN METODE K-MEANS CLUSTERING UNTUK PENGELOMPOKAN PENJUALAN TERLARIS PRODUK KACAMATA."
- [14] ALFIAN MA'ARIF, "buku python," *BUKU AJARPEMROGRAMAN LANJUTBAHASA PEMROGRAMAN Pythonolehalfianma'arifprogr. Stud. Tek. ELEKTROFAKULTAS Teknol. Ind. AHMAD DAHLANYOGYAKARTA2020*, pp. 2–62,

- 2020.
- [15] F. Handayani, “Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar,” *J. Teknol. dan Inf.*, doi: 10.34010/jati.v12i1.
- [16] S. Narulita, P. Prihati, A. Tigor Oktaga, and A. Eka Widyantoro, “Performansi Algoritma Clustering K-Means untuk Penentuan Status Malnutrisi pada Balita,” no. 1, p. 2023.