IMPLEMENTATION OF K-MEANS CLUSTERING AND AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS TO DETERMINE THE NUTRITIONAL STATUS OF TODDLER

1st Sri Rahmahwati
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
sri.rahmawati_ti19@nusaputra.ac.id

2nd Zulia Nur Permatasri
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
zulia.nur_ti19@nusaputra.ac.id

3rd Alvira Fauziah Rahmah
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, Indonesia
alvira.fauziah_ti19@nusaputra.ac.id

4th Rizal Fadli
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, indonesia
rizal.fadli_ti19@nusaputra.ac.id

5th Ivana Lucia Kharisma
Informatic Engineering Study Program
Nusa Putra University
Sukabumi, indonesia
ivana.lucia@nusaputra.ac.id

Abstract— Problems related to fulfilling toddler nutrition are still homework in Indonesia. The results of the survey data on the nutritional status of children under five in Indonesia (SSGBI) for 2021 tackling the prevalence of stunting in Indonesia reached 24.4%, wasting reached 7.1%, and wasting reached 17.0%. The number of stunting toddlers in Indonesia still exceeds the WHO threshold, which is 20%. Therefore, to reduce the level of malnutrition, it is necessary to record and classify the nutritional status of children under five. This research aim to grouping the data of toddlers based on the age by month and weight by kg using KnA algorithm. By using the silhouette score, 2 clusters has the closest value to 1 compared to the number of other clusters, where cluster 0 has a total of 69 toddlers, while the number of toddlers in cluster 1 is 112 toddlers. This result can be used for posyandu to analyze toddler segmentation.

Keywords: K-Means, Agglomerative, Nutrition

I. INTRODUCTION

Growth is one indicator that is relevant to the nutritional status of children, and is a tool to assess children's health. Malnutrition is generally a condition that occurs as a result of an unbalanced diet where there is no variation in certain nutrients, it could be due to excess (too high), or the wrong proportions. In developing countries, malnutrition causes as many as 10.8 or 54% million child deaths [1]. Problems related to fulfilling toddler nutrition are still homework in Indonesia. Survey data on the nutritional status of toddlers in Indonesia (SSGBI) for 2021 shows that the prevalence of stunting in Indonesia has reached 24.4%, wasting reached 7.1%, and underweight reached 17.0%. The number of toddlers suffering from stunting in Indonesia still exceeds the WHO threshold, which is 20%. Even though it is decreasing every year, the problem of malnutrition in Indonesia is still high. In Sukabumi district itself, based on the Health Service dataset for 2019, the number of toddlers suffering from malnutrition has reached 26,009 [2].

Apart from having an impact on physical growth and mental development, malnutrition also increases the risk of morbidity and mortality in children [3]. The increased risk of morbidity and mortality in children due to malnutrition is related to infectious diseases that often accompany malnutrition such as acute respiratory infections, diarrhea, measles and several other infectious diseases [1]. The World Health Organization (WHO) stated that 54% of under-five deaths in 2002 were caused by malnutrition [4]. Maternal, socioeconomic, demographic, and behavioral factors are factors that influence malnutrition because they are related to nutritional intake and disease in children under five [5]. One form of government programs to control the growth of toddlers is an integrated service post (Posyandu). The role of Posyandu is needed to reduce the amount of malnutrition in toddlers.

This research aim to grouping the data of toddlers based on the age by month and weight by kg [6], because the data provided has no label, we used clustering to create a new label based on the data similarity [7]. In previous research regarding k-means and algomerative algorithms for toddler nutrition, many have been carried out, but on average the results of these studies were in the form of classification based on anthropometric indices, and validation was carried out using validation for classification so that the accuracy results were very low [8]. Therefore, this research was conducted to show the use of clustering on data that does not have labels.

In this research, we used K-Means algorithm combined with the Agglomerative algorithm, the combination of these two algorithms is able to provide maximum results based on validation results using silhouette scores [9][10]. Therefore, the authors conducted research to cluster the nutritional value of toddlers in 3 Sukabumi districts including Posyandu Kampung Pasir Bentik RW 22 and 23 Nagrak Utara Village, Posyandu Tulip RW 07 Parungkuda Village and Posyandu Nusa Indah in Cibatu Caringin Village RW 05 Nagrak Village, Cisaat district.

II. LITERATUR REVIEW

A. Nutritional status

Nutritional status is the condition of the body due to food consumption and utilization of nutrients, where the body requires nutrients for energy sources, maintenance of body tissues, growth, and regulating body processes [11]. In general, growth status in children is calculated based on length, height, weight, and age and is assessed based on a combination of indicators of length/height for age (stunting), weight for length/height (wasting), and weight for age (under-weight). One of the most important risk factors for undergrowth is malnutrition. Adequate nutrition during childhood will significantly promote growth and prevent chronic disease in the future.

B. Clustering

Cluster analysis or *clustering* is a multivariate technique where the analysis process uses a clustering algorithm compared to humans, with the main objective of this analysis being to arrange or sort data based on the characteristics of the data so that data that has the closest similarity to other data will be collected in one cluster. In other words, this analysis is useful in finding labels in previously unknown data [12].

Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with high intra-class similarity and high inter-class similarity. The quality of the clustering results depends on the similarity measure used in the method and its implementation. Clustering can also be useful as a data-preprocessing step to identify related groups in building a model [13].

a) Agglomerative Hierarchical Clustering

The agglomerative hierarchical clustering (AHC) algorithm is a clustering algorithm with a bottom-up approach that builds a hierarchy starting from each node as a singleton cluster and successively merging the nodes into larger clusters until only one cluster remains [14]. Following are the steps of the agglomerative hierarchical clustering method:

1. The first step starts with N clusters, where N is the number of objects or data.
2. 2 objects with similarities (closest distance) are then combined.
3. Repeat Step 2 as much as N – 1 times. (the final result is that all objects will be in 1 cluster). Note the level or identity of the merged cluster where the join is placed.

b) K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm used to solve clustering problems in machine learning. K-Means Clustering groups datasets that do not have labels into different clusters, where the number of clusters is defined by K [15]. Following are the steps in the K-Means Clustering method:

1. Determine the desired number of K clusters.
2. Determine the random value of K clusters for the initial centroid.
3. The distance between the data and each centroid is then calculated using the Euclidean distance formula.

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n} (x_j - c_j)^2} \qquad (1)$$

Where:
d: Distance
x_j: Data to j
c_j: Centroid to j

4. Each data is grouped based on the smallest distance to the centroid.
5. The centroid value is then updated from the cluster average using the equation:

$$C_k = \frac{1}{nk} \sum d_i \qquad (2)$$

With:
C_k= clusters
nk = the amount of data in the cluster
d_i= number of each object included in each cluster

6. Repeat steps 2 to 5 until no cluster members have changed.

C. Related Research

Some of the research that the authors refer to is research on k-means and agglomerative, namely research with the title K-Means Clustering Algorithm for Determining the Nutritional Value of Toddlers written by Eni Irfiani, Siti Sulistia Rani (Bina Sarana Informatics University) in 2018. In this study used the parameters of height and weight in toddlers using the K-means clustering algorithm with a total of 91 toddlers. The toddler data is divided into 5 clusters, namely: Obesity, Overnutrition, Good Nutrition, Undernutrition, Poor Nutrition. The results showed that 7% of children under five were malnourished, 4% under five were malnourished, 35% under five were well nourished, 24% under five were overweight and 30% under five were obese. In this study, the accuracy of clustering results using k-means was not shown with the results from local health workers [16].

The next research is research by Adimas Ketut Nalendra (Nusantara University PGRI Kediri) in 2018 with a journal entitled Measuring the Accuracy of the K-Means Method for Determining the Nutritional Status of Toddlers. This study used data with a total of 50 toddlers and indicators in the form of body weight and height. The number of clusters used is 5 clusters, namely Malnutrition, Undernutrition, Good Nutrition, Overnutrition, and Obesity. The results of the study show an accuracy rate of 34%, or as many as 17 toddler data that are grouped together. This shows that the level of accuracy using the K-Means algorithm in classifying the nutritional status of toddlers is still very low [17].

The next research is a journal with the title IMPLEMENTATION OF CLUSTERING USING K-MEANS METHOD TO DETERMINE NUTRITIONAL STATUS by Stefanny Surya Nagari, Lilik Inayati (Airlangga University) in 2020. This research uses data on toddlers from Ponkesded Mayangrejo, Bojonegoro Regency and is secondary in nature. This study divided the data into 4 clusters, with the result that 23 children under five were malnourished, 17 under five were undernourished, 7 under five were well nourished, and 10 under five were overnourished. The author suggests trying other algorithms as well as good information in order to get a better level of accuracy in classifying the nutritional status of toddlers [12].

The latest research that is a review of the literature in this study is research by M. Venkat Reddy, M. Vuvekananada, R U V N Satish from IIT Tirupati, Andhra Pradesh in 2017 with a journal entitled Division Hieararchical Clustering with K-Means and Agglomerative Hierarchical Clustering. The results of this study say that the agglomerative hierarchical clustering algorithm implements the k-means algorithm efficiently, where the initial centroids of each cluster can be determined precisely rather than randomly selected. By choosing the right centroid value, it will produce efficient results [9].

III. Research Methods

The method used in this research is the combination of kmeans and agglomerative hierarchical clustering algorithm [18].

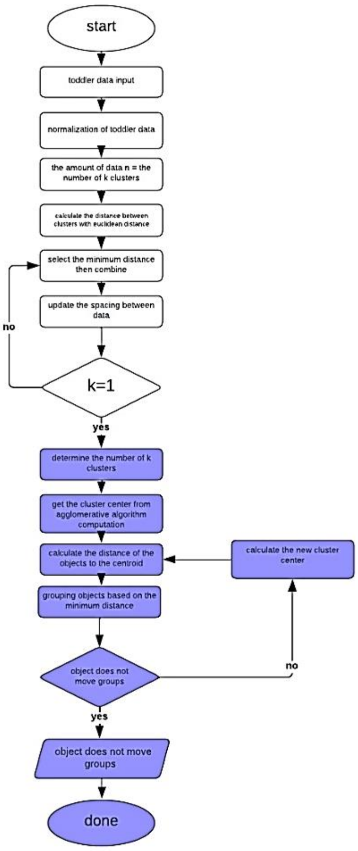
start
toddler data input
normalization of toddler data
the amount of data n = the number of k clusters
calculate the distance between clusters with euclidean distance
select the minimum distance then combine
update the spacing between data
no
k=1
yes
determine the number of k clusters
get the cluster center from agglomerative algorithm computation
calculate the distance of the objects to the centroid
calculate the new cluster center
grouping objects based on the minimum distance
object does not move groups
no
yes
object does not move groups
done

Fig. 1 KnA Algorithm Flowchart.

Based on the process flow in the flowchart above, the main process details consist of the following steps [19]:

a) Input the dataset that has been provided.
b) Normalizing the dataset using the Z-Score method.
c) Perform clustering on the dataset with an agglomerative hierarchical clustering algorithm.
d) Calculating the value of the initial centroid from the results of clustering using agglomerative hierarchical clustering.
e) Determining the number of k clusters using silhouette score.
f) Perform K-Means clustering by initializing the centroid resulting from agglomerative hierarchical clustering.
g) K-Means clustering results.

IV. Results and Discussion

a. Hierarchical Clustering Algorithm Process

Data processing is using agglomerative hierarchical clustering algorithms. This process is done with the help of Google collab software. The processed data is then divided into k clusters, and the average of each clusters is calculated to then be used as the initial center point of the cluster.

b. K-Means Clustering Algorithm Process

After the number of clusters and initial centroid is determined, the next step is to process the data using the k-means clustering algorithm. Data processing with the k-means clustering algorithm was carried out with the help of google collab software, the following are the results of data processing using the k-means clustering algorithm.

The visualisation of clustering on the data showed in the figure below.

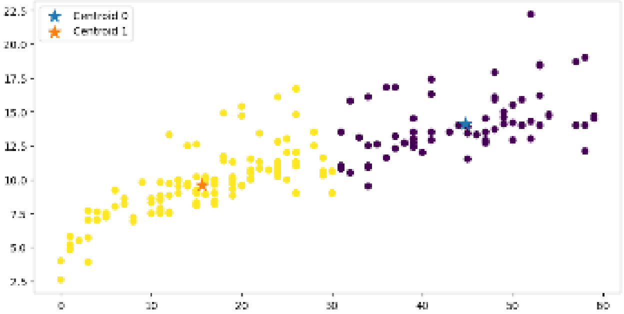

Fig. 2 The Results of Clustering Using 2 k Numbers

On the figure above, the data divided into 2 clusters which are centroid 0 and centroid 1, the centroid 0 showed with blue color, and the centroid 1 showed with yellow color.

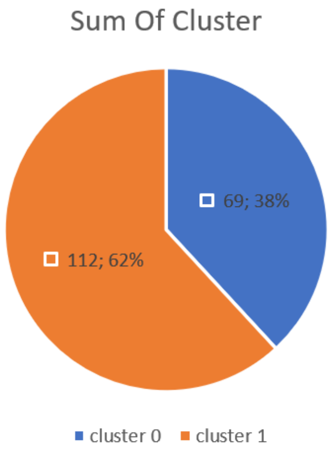

Fig. 3 Sum of Clusters

On the figure above, cluster 0 has a total of 69 toddlers, while the number of toddlers in cluster 1 is 112 toddlers.

TABLE 1. RESULT OF CLUSTERING

Age (Month)	Weight (Kg)	Cluster
57	14.0	0
52	22.2	0
31	13.5	0
24	11.3	1
32	10.5	0
24	11.0	1

26	12.0	1
26	14.8	1
30	9.0	1
25	10.0	1

The table above showed the data of toddlers with the clustering created using KnA algorithm. From the clustering results, based on the closeness characteristics between data, a new label can be determined, namely, Cluster 1 are Toddlers with low to medium weight and age and Cluster 0 are Toddlers with higher weight and age.

c. Model Evaluation

1. Silhuoette coefficient

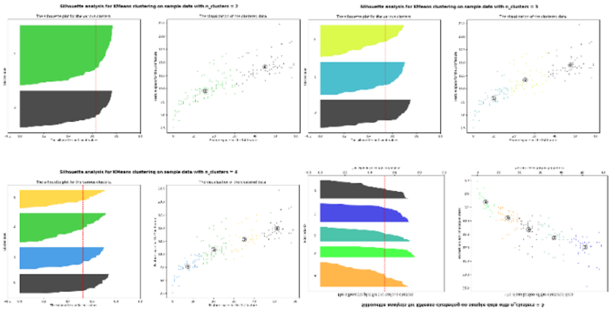

Fig. 4 Silhuoette coefficient

The following is the result of calculating the Silhouette Coefficient method above [8]:
For n_clusters = 2 The average silhouette_score is: 0.628371145500863
For n_clusters = 3 The average silhouette_score is: 0.539225609079401
For n_clusters = 4 The average silhouette_score is: 0.524751661522163
For n_clusters = 5 The average silhouette_score is: 0.518279709569204

From the results of the Silhouette Coefficient method above, it can be seen that the number of clusters 2 has the closest value to 1 compared to the number of other clusters, with a value of 0.628371145500863.

2. Accuration of each algorithm using silhouette score

The accuration of each clusters from each algorithm showed in the figure below.

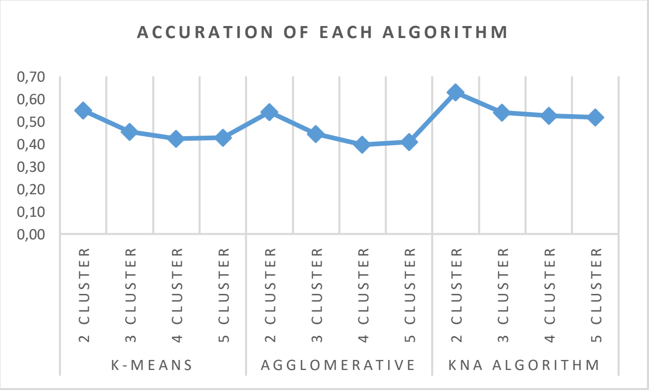

Fig. 5 Accuration of Each Algorithm

The figure above shows the accuration of each k number of clusters of each algorithm, and the KnA algorithm with 2 number clusters gives the highest accuration which is 0.628371145500863 compared to other clusters and algorithm.

V. Conclusion

Based on the results of data processing that has been done using the agglomerative method and k-means with parameters such as age and weight with a total of 181 data, 2 clusters are produced using shilouette score to get the optimal amount of clusters. Cluster 1 are Toddlers with low to medium weight and age and Cluster 0 are Toddlers with higher weight and age. These clusters can be used for Toddler Segmentation, For example, if we find that the group of toddlers with low growth tends to be in Cluster 1, Posyandu can design a special program to improve nutrition and care for toddlers in this group. Meanwhile, the group of toddlers with good growth (Cluster 0) may receive more general attention. By understanding the differences between these groups, Posyandu can work more effectively in providing health services that suit the specific needs of each group of children under five in the community.

Reference

[1] A. M. Fazle Rabbi and S. C. Karmaker, “Determinants of child malnutrition in Bangladesh - A Multivariate Approach,” *Asian J Med Sci*, vol. 6, no. 2, pp. 85–90, 2014, doi: 10.3126/ajms.v6i2.10404.

[2] Dinas Kesehatan, “Jumlah Balita Berdasarkan Kategori Balita Gizi Buruk di Jawa Barat,” https://opendata.jabarprov.go.id/. Accessed: Jan. 04, 2023. [Online]. Available: https://opendata.jabarprov.go.id/id/dataset/jumlah-balita-berdasarkan-kategori-balita-gizi-buruk-di-jawa-barat

[3] Ratoyo, “Strategi pemberdayaan masyarakat dalam penanganan kasus stunting di Kampung Tulung Kakan Kecamatan Bumi Ratu Nuban Kabupaten Lampung Tengah,” *Jurnal Simplex*, vol. 2, no. 3, pp. 63–74, 2019.

[4] R. F. Putri, D. Sulastri, and Y. Lestari, “Faktor-Faktor yang Berhubungan dengan Status Gizi Anak Balita di Wilayah Kerja Puskesmas Nanggalo Padang,” *Jurnal Kesehatan Andalas*, vol. 4, no. 1, pp. 254–261, 2015, doi: 10.25077/jka.v4i1.231.

[5] L. C. Chikhungu, N. J. Madise, and S. S. Padmadas, “How important are community characteristics in influencing children’s nutritional status? Evidence from Malawi population-based household and community surveys,” *Health Place*, vol. 30, pp. 187–195, 2014, doi: 10.1016/j.healthplace.2014.09.006.

[6] J. Homepage, A. Roihan, P. Abas Sunarya, and A. S. Rafika, “IJCIT (Indonesian Journal on Computer and Information Technology) Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper,” 2019.

[7] N. Komang Sri Julyantari *et al.*, “Implementasi K-Means Untuk Pengelompokan Status Gizi Balita (Studi Kasus Banjar Titih) Implementation of K-Means for Clustering the Nutritional Status of Toddlers (Banjar Titih Case Study),” *Jurnal Janitra Informatika dan Sistem Informasi*, vol. 1, no. 2, pp. 92–101, 2021, doi: 10.25008/janitra.

[8] A. Subayu, “PENERAPAN METODE K-MEANS UNTUK ANALISIS STUNTING GIZI PADA BALITA: SYSTEMATIC REVIEW,” 2022. [Online]. Available: www.geospasial.com

[9] V. Makara, M. V. Reddy, M. Vivekananda, and H. -Telangana, “Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,” *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, 2013, [Online]. Available: www.ijcstjournal.org

[10] I. Alpiana and L. Anifah, “Penerapan Metode KnA (Kombinasi K-Means dan Agglomerative Hierarchical Clustering) dengan Pendekatan Single Linkage untuk Menentukan Status Gizi pada Balita,” 2019. [Online]. Available: https://journal.unesa.ac.id/index.php/inajet

[11] C. Auliya, O. W. K. Handayani, and I. Budiono, “Profil Status Gizi Balita Ditinjau Dari Topografi Wilayah Tempat Tinggal (Studi Di Wilayah Pantai Dan Wilayah Punggung Bukit Kabupaten Jepara),” *Unnes Journal of Public Health*, vol. 4, no. 2, pp. 108–116, 2015.

[12] S. S. Nagari and L. Inayati, “Implementation of Clustering Using K-Means Method To Determine Nutritional Status,” *Jurnal Biometrika dan Kependudukan*, vol. 9, no. 1, p. 62, 2020, doi: 10.20473/jbk.v9i1.2020.62-68.

[13] V. Makara, M. V. Reddy, M. Vivekananda, and H. -Telangana, “Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,” *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, no. July, pp. 5–11, 2013.

[14] A. Rosen *et al.*, “PENERAPAN METODE AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK KLASIFIKASI KABUPATEN/KOTA DI PROVINSI JAWA TIMUR BERDASARKAN KUALITAS PELAYANAN KELUARGA BERENCANA,” *Teach Teach Educ*, vol. 12, no. 1, pp. 1–17, 2015.

[15] M. K-means, O. Purwaningrum, Y. Y. Putra, and A. A. Arifiyanti, “Penentuan Kelompok Status Gizi Balita dengan Menggunakan,” vol. 15, no. 2, pp. 129–136, 2021.

[16] E. Irfiani and S. S. Rani, “Algoritma K-Means Clustering untuk Menentukan Nilai Gizi Balita,” *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, vol. 6, no. 4, p. 161, 2018, doi: 10.26418/justin.v6i4.29024.

[17] S. Informasi *et al.*, “Pengukuran Keakuratan Metode K-Means untuk Menentukan Status Gizi Balita,” *E-Journal.Polsa.Ac.Id*, vol. 7, no. 1, 2019.

[18] S. S. Nagari and L. Inayati, “IMPLEMENTATION OF CLUSTERING USING K-MEANS METHOD TO DETERMINE NUTRITIONAL STATUS,” *Jurnal Biometrika dan Kependudukan*, vol. 9, no. 1, p. 62, Jun. 2020, doi: 10.20473/jbk.v9i1.2020.62-68.

[19] “A combination of algorithm agglomerative hierarchical cluster (AHC) and K-means for clustering tourism in Madura-Indonesia,” *Journal of Mathematical and Computational Science*, 2022, doi: 10.28919/jmcs/7086.