

# Implementation Of K-Means Clustering Algoritm To Determine Stunted Status In Children Under Two Years Old

Petrisia Widyasari Sudarmadji<sup>1,\*</sup>, Christa Elena Blandina Bire<sup>1</sup>  
{petrisia.pnk@gmail.com, christa.eren@gmail.com}

Politeknik Negeri Kupang

**Abstract.** The stunting prevalence in Indonesia ranks fourth in the world, meaning that 37% of Indonesian toddlers are stunted. Stunting is a chronic nutritional deficiency problem caused by lack of nutritional intake for a long time, resulting in a disturbance to the child's growth, which is a lower or shorter child (dwarf) than the standard of his age, Especially in the first thousand golden age of life. The first thousand days of life is counted from the 9 months of a child in terms of mothers up to 2 years of life (0-23 months). The Data on the basic health Research (RISKESDAS) in 2018 shows the prevalence of stunting in the national sphere of 30.8%, consisting of a short prevalence of 19.3% and very short at 11.5%, and the highest percentage in 2018 is in Nusa Tenggara province. Eastern (42.6%). NTT ranked first in the province for stunting status and the parameters used to determine stunting status are based solely on the age of weight (BB/U), which is on the card to the Healthy (KMS). The urgency of the problem, is to classify the stunting status specifically based on the standard characteristics of the WHO use the K-means Clustering algorithm. The objective of the study is to obtain the classification results with the algorithm method of the Clustering K-means in determining the stunting status in children under the age of two years. The method of implementing this research is CRISP-DM (Cross Industry Standard Process for Data mining) model.

**Keywords:** Stunting, K-Means, Antropometri, Clustering

## 1 Introduction

The stunting prevalence in Indonesia ranks fourth in the world, which means approximately 9 million or 37% of Indonesian toddlers are stunting (Risksedas 2018). Stunting is a major threat to Indonesian human quality, as well as a threat to the competitiveness of the nation. It is in because stunted children, not just disturbed physical growth (short/dwarf) alone, but also impaired the development of its brain, which certainly will greatly affect the ability of cognitive, productivity and creativity in Productive age. Stunting is a chronic malnourished problem caused by lack of nutritional intake in a long time, resulting in a growth disorder in children namely the height of children's lower or short (dwarf) than the standard of age, especially The first thousand days of life. The first thousand days of life in the count since 9 months of the child in the mother count to 2 years of life (0-23 months). The basic health Research Data (Risksedas) in 2018 showed the prevalence of

stunting in national scope by 30.8%, consisting of a short prevalence of 19.3% and very short of 11.5%, and the highest percentage in 2018 was in the province East Nusa Tenggara (42.6%). Although the percentage number in the East Nusa Tenggara region fell from 51.7% in 2013, the NTT still occupied the first ranked province for stunting status [1].

Short toddler is a toddler with a nutritional status based on length or height according to the age when compared to the raw standard of WHO-MGRS (multicentre Growth Reference study) in 2005, Z-Zcorenya value less than -2SD and categorizes Very short if the Z-scorenya value is less than -3SD [1]. Therefore, the improvement efforts should include efforts to prevent and reduce direct interference (specific nutritional interventions) and attempts to prevent and reduce interference indirectly ( Sensitive nutritional interventions). The specific nutritional intervention is generally done in the healthcare sector, but contributes only 30% while 70% is a contribution of sensitive nutritional intervention involving various sectors such as food security, the availability of clean water and Poverty, education, social and so on. Results from anthropometry stunting inspection, capable of detecting and classifying stunting status [1].

In the Puskesmas located in Kolbano District – South Central Timor – NTT, the parameters commonly used in the determination of stunting status are based solely on the age of weight (BB/U), which is in the card towards healthy (KMS). Then note on the form of monitoring the stunting and nutritional status of infants and match the nutritional status of the toddler based on WHO's reference. But the classification of stunting status is not specifically explained so that the Puskesmas officers and parents do not know in detail whether the toddler belongs to the child who is short, very short, or normal. The explanation above, becomes the urgency of the problem to be meticulously classify the stunting status specifically by grouping data in accordance with the standard characteristics of the WHO (the analysis of the classification result) using K-Means algorithm Clustering which is one of the data mining methods to grouping large data in classes that have similarities. The purpose of this research is to obtain a classification with the method of algorithm K-Means Clustering in determining the stunting status in the child (under two years) in Kolbano-Timor South Central-NTT District and the results of this research can Used as a reference for the Puskesmas in Kolbano sub-district.

In the stage of designing and building the algorithm implementation system K-Means Clustering to determine the Stunting Status in the child Baduta (under two years), has been conducted preliminary studies so that researchers find the modeling that will be applied to Overcome the problems in the effort to obtain the classification result with the method of algorithm K-Means Clustering in determining the stunting status in the child (under two years) in Kolbano-Timor South Central-NTT District and the results of this research Can be used as a reference for the Puskesmas in Kolbano sub-district. The built-in system, prioritizes the needs of users so that supporting data in implementing the K-Means Clustering algorithm to determine the stunting status in the child based on preliminary studies and in the analysis resulting in a The formula of the research flowchart and is described in general in the following architecture:

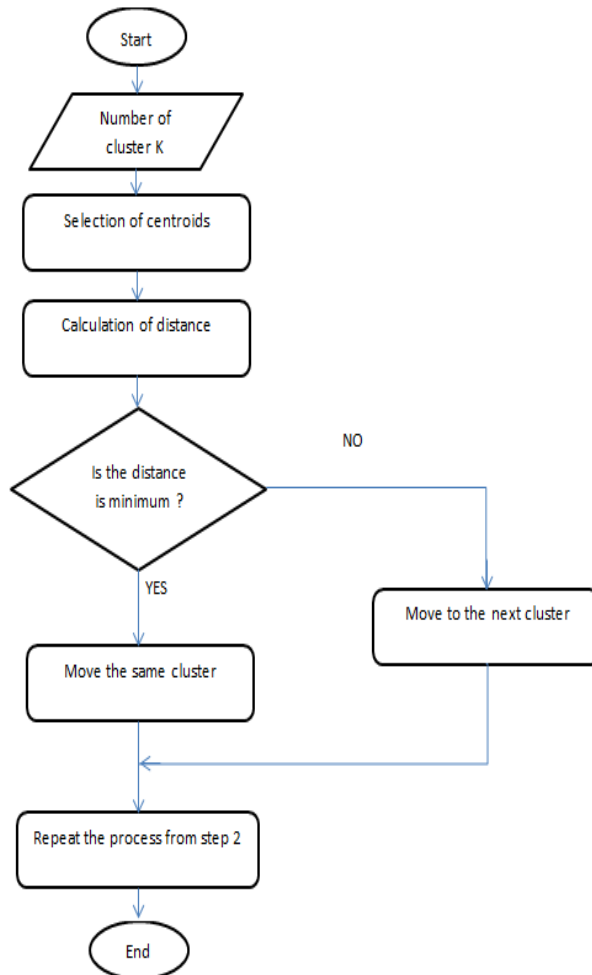


Figure 1. Algorithm K-Means Clustering

### Stunting

Stunting is a chronic malnourished problem caused by insufficient nutritional intake in a long time, due to the provision of food that does not fit the nutritional needs, resulting in growth disorders in children, namely the height of children Low or short (dwarf) from its age standard (Risksedas 2018). Stunting can occur as the fetus is still in the womb and is only visible when the child is two years old. Stunting in shape by growth faltering and the inadequate catch up growth, which reflects the inability to achieve optimal growth, it reveals that groups of infants born with normal weight can experience When fulfilling the next needs is not met properly.

Stunting can be detected when a toddler is weighing in weight and measured in length or height, then in comparison to the standard Z zcore of the WHO and get the result below

normal or physically shorter toddlers in comparison Toddler Seumurannya. In general, Anthropometry is used to see an imbalance of protein intake and energy. Anthropometry index used is weight according to age (BB/U), Height according to age (BB/TB) expressed by standard deviation unit Z (Z-Score) and will generate classification: normal, short and very short, which is a match Stunted (short) and severely stunted (very short). The following is a classification of stunting nutritional status based on the height indicators per age (TB/U) [2]:

- Very short: Zscore <-3.0 SD
- Short: Zscore <-2.0 S. D Zscore ≥-3.0
- Normal: Zscore ≥-2.0 SD
- Above Normal : Zscore > SD

Below is a classification of stunting nutrition status based on TB/U and BB/TB indicators:

- Short – Skinny:-Zscore TB/U <-2.0 and Zscore w/TB <-2.0
- Short – Normal: Zscore TB/U <-2.0 and Zscore BB/TB between-2.0 S. D 2.0
- Short – Grease: Zscore ≥-2.0 s. D Zscore ≤ 2.0

### Data Mining

Data mining is a way to find interesting patterns and knowledge in large amounts of data. The completion of data mining using mathematical techniques, statistics, artificial intelligence and machine learning to identify and extract into information or knowledge that is in the vast database of various databases [3]. A Cluster is a collection of data objects that have similarities between one and the other in the same group and different objects with the other group's data. Clustering or better known as cluster analysis is a process of grouping a set of physical or abstract objects into a single class of the same object [4].

K-Means is the method of data analysis in Data Mining where the modeling process without supervision and is one of the methods that classify data on a partition. In the K-Means method The data is grouped into groups where each group has similar or equal characteristics but with other groups having different characteristics. This method provides differences between data in a single cluster and maximizes differences with other clusters [4]. Generally the K-Means method uses the following algorithms:

1. Specify K as the number of clusters that are in the form. The determination of the number of clusters K is carried out with several factors such as theoretical and conceptual considerations proposed to determine how many clusters.
2. Generate an initial Centroid (cluster center point). To determine the initial centroid is performed randomly from several available objects as many as K clusters, to calculate the next centroid of the cluster to-I.
3. Calculate the distance of each object to each centroid of each cluster. Then calculate the distance between the objects by centroid. in this study using Euclidian Distance.

Where:

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

Xi: Object x to-i

Y: Power y to-i

N: Number of objects

4. Allocate each object to the nearest centroid.

5. Iterate, then specify a new centroid position by using equations.
6. Repeat step 3 if the new centroid position is not equal.

## 2 Experimental methodology

The method in implementing this research is the model CRISP-DM (Cross Industry Standard Process For Data Mining). The CRISP-DM Model was introduced in mid-1990 by a European consortium company. In CRISP-DM, a data mining project has a life cycle divided into 6 phases as shown in the image below [5]:

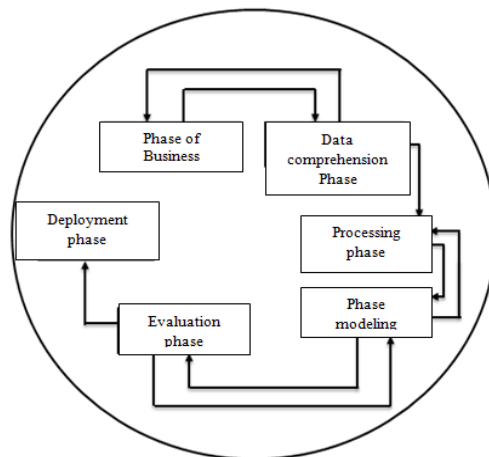


Figure 2. Life cycle Model CRISP-DM

The steps that are done in this study according to the image above are:

1. Business Understanding Phase: Includes business goal determination, assessing current situation, setting data mining objectives, and developing project plan. The business purpose that will be done in this research is how to do the grouping of the nutritional value of stunting based on TB/U and BB/TB of Baduta children (under two years) using the algorithm of K-Means Clustering.
2. Data Understanding Phase: Once the business objectives and project plan are established, the next step performs the initial data collection, data description, data exploration, and data quality verification. This proposed study uses primary data, with the response being the children of the Baduta (under two years) according to the data in the POSYANDU (integrated Service post) of Kolbano subdistrict. Calculation that will be used to determine the status of the nutrition stunting Baduta children (under two years) is the Growth Chart method taken from the site <http://cdc.gov/growthcharts/which> is a reference to the raw standards of WHO.
3. Data Preparation Phase: At this stage the identification and development of the data that has been collected to be able to do grouping and mapping into predefined groups.
4. Modeling phase: In this phase is the selection of models that will be used to grouping the status of the nutrition stunting Baduta children (under two years). The Model or method to be used on this research is the method of K-Means Clustering.
5. Evaluation Phase: Testing will be conducted by comparing the grouping conducted by the algorithm of K-Means Clustering with the grouping conducted by the personnel/medical personnel in Kolbano subdistrict Puskesmas.

6. Deployment phase: This phase is done for knowledge discovery (identification of unexpected and useful relationships) to be applied to business operations in various destinations, including clustering.

### 3 Results and discussion

The population that is in use as the object of this study, consists of 75 populations that are selected using propability sampling techniques. Of these, in the count use the formula Slovin as follows:

$$n = \frac{N}{1 + Ne^2}$$

Where:

n : Sample size

N: Population size

e : looseness of inaccuracy due to error

Then:  $n = 75 / (1 + 75) (0.0025)$

$n = 40$

Based on the formula, the large number of samples is 40 children under two years of age. The following is the data length (PB) and age (U) in the integrated Service post (POSYANDU) at Puskesmas Kolbano.

Table 1. Data length and age of children under two years

<b>Data To -i</b>	<b>Body Lenght (PB) cm</b>	<b>Age (U) month</b>
Baduta 1	60	7
Baduta 2	65	12
Baduta 3	70	10
Baduta 4	80	23
Baduta 5	78	20
Baduta 6	80	15
Baduta 7	65	8
Baduta 8	58	6
Baduta 9	70	7
Baduta 10	65	7
Baduta 11	69	12
Baduta 12	70	12
Baduta 13	80	14
Baduta 14	82	18
Baduta 15	78	23
Baduta 16	80	20

Baduta 17	98,5	22
Baduta 18	71	16
Baduta 19	68	10
Baduta 20	60	6
Baduta 21	70	8
Baduta 22	80	14
Baduta 23	78	16
Baduta 24	82	22
Baduta 25	80	20
Baduta 26	79	22
Baduta 27	82	22
Baduta 28	80	20
Baduta 29	75	18
Baduta 30	70	10
Baduta 31	75	23
Baduta 32	75	23
Baduta 33	72	15
Baduta 34	80	18
Baduta 35	78	17
Baduta 36	71	7
Baduta 37	68	7
Baduta 38	70	10
Baduta 39	65	8
Baduta 40	84	23

The purpose of this study is to conduct the classifying stunting status in children under the age of two in Puskesmas (Public health Center) – Kolbano uses K-Means method.

a. The normalization calculation

Data of children under two years in table 1, not directly able to be due to the number of digits between the variable length of body (PB) and age (U) far enough. To minimize the number of numbers, do not normalize the variable by using the formula:

$$\text{Normalization value} = \frac{(\text{Initial value} - \text{minimum value})}{(\text{Maximum value} - \text{minimum value})}$$

Based on the variable value of the body length and age, the process will be normalized into the range 0 to 1. The process of normalization is needed before calculating the centroid

value of the K-Means. The result of calculating the normalization value based on table 1 is as follows:

Table 2. Normalization of body length (PB) and age (U) values

<b>Data to - i</b>	<b>Body Lenght (PB) cm</b>	<b>Age (U) month</b>
Baduta 1	0,07	0,06
Baduta 2	0,29	0,35
Baduta 3	0,5	0,24
Baduta 4	0,92	1
Baduta 5	0,83	0,82
Baduta 6	0,92	0,53
Baduta 7	0,29	0,12
Baduta 8	0	0
Baduta 9	0,5	0,06
Baduta 10	0,29	0,06
Baduta 11	0,46	0,35
Baduta 12	0,5	0,35
Baduta 13	0,92	0,47
Baduta 14	1	0,71
Baduta 15	0,83	1
Baduta 16	0,92	0,82
Baduta 17	98,5	0,94
Baduta 18	0,54	0,59
Baduta 19	0,42	0,24
Baduta 20	0,08	0
Baduta 21	0,5	1,02
Baduta 22	0,92	0,47
Baduta 23	0,83	0,59
Baduta 24	1	0,94
Baduta 25	0,92	0,82
Baduta 26	0,88	0,94
Baduta 27	1	0,94
Baduta 28	0,92	0,82
Baduta 29	0,71	0,71



Baduta 30	0,5	0,24
Baduta 31	0,71	1
Baduta 32	0,71	1
Baduta 33	0,47	0,53
Baduta 34	0,92	0,71
Baduta 35	0,83	0,65
Baduta 36	0,54	0,06
Baduta 37	0,42	0,06
Baduta 38	0,5	0,24
Baduta 39	0,29	0,12
Baduta 40	1,08	1

b. Value of initial cluster center

After the normalization value of variable body length and age is gained as in Table 2, the next step is to specify the number of groups or clusters. To 40 data of children under two years old in the previous table will be grouped in 4 clusters: Abnormal, Normal, Stunting, Severe Stunting. After the number of clusters is specified, the next step is to specify the value of the initial cluster Centre for each cluster in each variable. To determine the initial centroid value, many methods can be used and in the case of this research method is used to fetch data from the source, randomly. The following is the initial default cluster center value that has been randomly selected;

Table 3. Initial Cluster Center Value

CLUSTER	STATUS	PB(cm)	U(month)
1	abnormal	98,5	0,94
2	normal	1	0,71
3	stunting	0,92	0,53
4	severe stunting	0,47	0,53

c. Grouping data cluster

After each data in the calculated distance in each cluster, then the next step is to group the data according to the Clusternya. Cluster groups of data are retrieved from the shortest distance of the data against a cluster.

d. Determination of new centroid values

Once grouping is done, it is next to calculate the new centroid value by counting based on grouping table and 40 data of children under the age of two years and their distance:

Table 4. New Centroid Value

	New Centroid Value			
	C1	C2	C3	C4
Body Length (PB)	98,5	1	0,92	0,47
Age (U)	0,94	0,71	0,53	0,53

After a new centroid value is gained, the following step is to compare the previous centroid value; And if the value remains the same/unchanged then the iteration process stops. But if the value is not the same, then the data grouping process should be restarted again until there is no more data change. In Figure 3 below, get the result grouping stunting status in children under two years old using K-means. Based on the classifying result in Get 4 clusters where: There are 48% of children Severe Stunting status, 22% child stunting status, 28% of normal child status and 2.50% of abnormal status children.

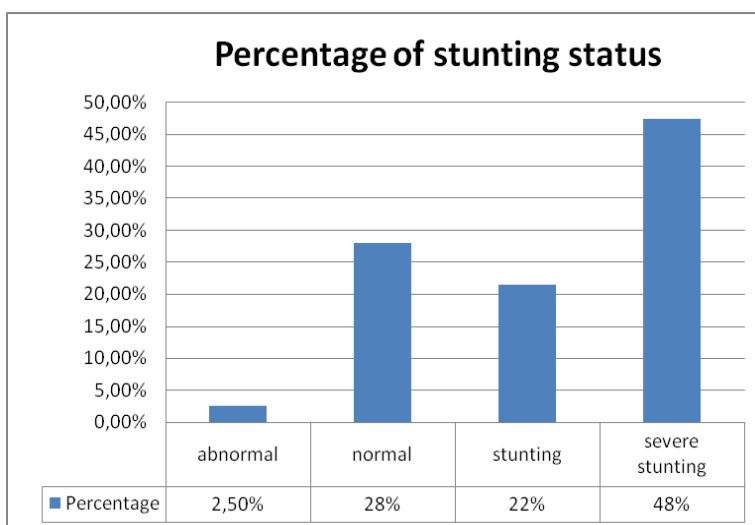


Figure. 3. Percentage result of stunting status chart

#### 4 Conclusion

Based on the results of the study, the percentage of stunting status in children under two years at Puskesmas (Public Health Center) – Kolbano Sub-district, with classifying K-Means Clustering algorithm through two parameters namely body length (PB) and age (U), which is divided into 4 clusters of Abnormal, Normal, Stunting and Severe Stunting, to help the performance on medical personnel and elements of health care practitioners and parents of children under the age of two years in Early detection of stunting status. From the processed

data, there are still 48% of children in severe stunting status as well as 22% of stunted children and these results are compared with grouping results using the Growth Chart (KMS) table in get 28 data that has a group The same. And from this figure can be concluded that the algorithm K-Means Clustering has a value of 70% accuracy is correct. This value can change with the addition of training data.

### **Acknowledgements**

The author wants to thank the Kupang State Polytechnic for financial support through the ROUTINE Research program. The author also thanked doctors, nurses and midwives at Kolbano Puskesmas for the help of data and their contributions in assisting the research process so that all went smoothly, as well as students of the Computer Engineering study Program top network Data mining processing assistance.

### **References**

- [1] RISKESDAS 2018, Ministry of Health Research and development Health, Republic of Indonesia 2018
- [2] anggraeni R, — Classification of infant Nutrition Status by Anthropometry Index (BB/U)  
Using neural networks, | SNASTI Journal, ICCS., pp. 14 – 18, 2010
- [3] E. Prasetyo, — Data Mining Concepts and applications using Matlab, | Publisher Andi Yogyakarta, 2012
- [4] Windha Mega Pradnya Duhita, — Clustering using the K-Means method for Determining the nutritional Status of toddlers, | Journal of Informatics., vol. 15, No. 2, pp. 160 – 174, December 2015
- [5] Eni Irfiani and Siti Sulistia Rani, — K-Means Clustering algorithm to determine Nutritional value of toddlers, | Journal of Systems and Information technology (JUSTIN), vol. 6, No. 4, pp. 16 – 172, October 2018