International Journal of Applied Sciences and Smart Technologies

Volume 3, Issue 1, pages 131–142

p-ISSN 2655-8564, e-ISSN 2685-9432

Classification of Toddler Nutrition Using C4.5 Decision Tree Method

Kartono Pinaryanto[1,*], Robertus Adi Nugroho[1],

Yanuarius Basilius[1]

[1]*Department of Informatics, Faculty of Science and Technology,*
Sanata Dharma University, Yogyakarta, Indonesia
[*]*Corresponding Author: kartono@usd.ac.id*

(Received 07-05-2021; Revised 15-06-2021; Accepted 16-06-2021)

Abstract

Nutrition is very much needed in the growth of toddlers. It is very important to give babies a balanced nutritional intake at the right stage so that the baby grows healthy and is accustomed to a healthy lifestyle in the future. Children under five years of age are a group that is vulnerable to health and nutrition problems. In determining the nutritional status, it can be done in a system manner using the C4.5 decision tree classification method and entering several variables or attributes. The dataset tested was 853 toddlers. Classification is carried out to determine the nutritional status based on the weight/age (BB/U), height/age (TB/U) and weight/height (BB/TB) categories. The attributes used for the classification of BB/U are gender, weight and age. The attributes used for TB/U are gender, body length or height, and age. The attributes used for BB/TB are gender, weight, body length or height, and age. The average accuracy of the BB/U category is 90.16%, the average accuracy of the TB/U category is 76.64%, and the average accuracy of the BB/TB category is 83.83%.

Keywords: Classification, decision tree, C4.5, nutrition for toddlers

1 Introduction

Nutrients are organic substances required for normal functioning of the body’s systems, growth and health maintenance. It is very important to give babies a balanced nutritional intake at the right stage so that the baby grows healthy and is accustomed to a healthy lifestyle in the future. Children under five years of age are a group that is vulnerable to health and nutrition problems, so that the toddler years are an important period of growth and need serious attention [1]. Based on the results of the 2018 Ministry of Health's Basic Health Research, 17.7% of infants under 5 years of age (toddlers) still experience nutritional problems. This figure consisted of Under-fives who suffered from malnutrition by 3.9% and those suffering from malnutrition by 13.8% [2]. The nutritional status of toddlers can be measured anthropometry, anthropometric indices are often used, namely: body weight for age (BB/U), height for age (TB/U), body weight for height (BB/TB). The weight index based on age (BB/U) is the most commonly used indicator because it has the advantage of being easy and quicker to understand by the general public. The reference standard used for determining nutritional status by anthropometry is based on the Decree of the Minister of Health No. 920/Menkes/SK/VIII/2002, to use the reference book of the “World Health Organization-National Center for Health Statistics” (WHO-NCHS) by looking at the Z-score.

In determining the nutritional status, it has been done manually by the Community health centers, so patients have to come physically to the Community health centers. This is of course very troublesome especially in the current pandemic situation and conditions. Determining nutritional status can be done automatically using a classification approach. One approach that can be taken is to use the C4.5 decision tree method. The C4.5 method is an algorithm that works by applying the concept of a decision tree. A decision tree is a predictive model using a tree structure or hierarchical structure. The concept of a decision tree is to transform data into a decision tree with decision rules.

In previous research on the comparison of the performance of the C4.5 and Naive Bayes algorithms for the classification of scholarship recipients by Choirul Anam and

Harry Budi Santoso, stated that the C4.5 algorithm has better performance than Naive Bayes with the level of accuracy obtained using the C4.5 algorithm of 96.4%, while the accuracy rate of Naïve Bayes is 95.11% [3].

Based on research [4] on the classification of typhoid fever (TF) and dengue hemorrhagic fever (DHF) by applying the C4.5 decision tree algorithm. It can be concluded that by using the k-folds cross validation test, the highest average accuracy value is 91.875% using 32 test data and 128 training data.

From the description above, a study was conducted using the C4.5 decision tree method in determining the nutritional status of children under five. It is hoped that applying the C4.5 decision tree method can help classify the nutritional status of toddlers to determine the growth of children under five.

2 Methodology

The methodology used in this study is as follows (Figure 1).

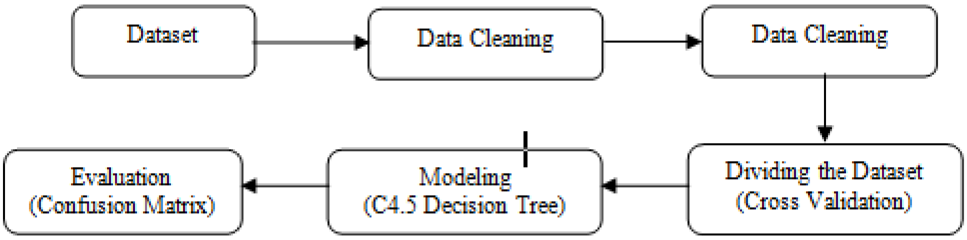

Figure 1. Decision tree classification research methodology C4.5

The research began to prepare the dataset, then the dataset went through the cleaning process and continued with data selection. The next stage, the data will be divided into testing data and training data. Training data will be used to form a decision tree, while testing data will be used to evaluate the system being created. In the next sub-section, it will be explained in detail about the stages that are passed.

2.1. Dataset

The dataset used in this study is the monitoring data on the nutritional status of toddlers, obtained from the Kebong Health Center, Kelam Permai District, Sintang

District, West Borneo in 2017 with a total of 853 toddlers. Monitoring data on the nutritional status of toddlers has three categories, namely the category of body weight according to age (BB/U), height for age (TB/U), and body weight for height (BB/TB). The BB/U category has 4 classification labels namely Best, Good, Bad and Worst. The TB/U category has 4 classification labels namely High, Normal, Short, and Very Short. While the BB/TB category has 4 classification labels namely Fat, Normal, Thin, and Very Thin (Table 1).

Table 1. Categories and Labels

No	Category	Label
1	BB/U	Best, Good, Bad, Worst
2	TB/U	High, Normal, Short, Very Short
3	BB/TB	Fat, Normal, Thin, Very Thin

2.2. Data Cleaning

Data cleaning is a process for cleaning unused data [5]. In this study, some data were deleted because were incomplete. An example of deleted data is that it does not have a BB/TB label, has no PB/TB value, and does not have a TB/PB conversion value.

2.3. Data Selection

In the dataset, there are 19 attributes, including name, date of birth, gender M/F, body weight, PB/TB, measured position, age, age calculation process, conversion of TB/PB, age family, code, code1, code2, nutritional standards Poor BB/U, Nutritional Standards Good BB/U, Short PB/U or TB/U Standards, Normal PB/U or TB/U Standards, Weight Standards BB/TB or BB/TB, and Normal Standards of BB/TB or BB/TB. At the data selection stage, the attributes used for the classification were determined (feature selection). In the selection of attributes, the attributes of Gender M/F, Body Weight, PB/TB and Age were selected. These attributes were selected based on recommendations from the health center. The results of the attribute selection are shown in Table 2.

Table 2. Attributes used by each category

No	Category	Attribute	Label
1	BB/U	Gender M/F, Body Weight, Age	Best, Good, Bad, Worst
2	TB/U	Gender M/F, PB/TB, Age	High, Normal, Short, Very Short
3	BB/TB	Gender M/F, Body Weight, PB/TB, Age	Fat, Normal, Thin, Very Thin

2.4. Dividing the Dataset

The dataset is divided into testing data and training data using k-folds validation. The number of k is chosen by the user where the values of k are 3, 5, 7 and 9 folds. If the value of $k = 3$, then the data is divided into 3 parts, 2 parts used for training data and 1 part for testing data, and likewise for dividing the value of 5, 7 and 9 folds.

2.5. Modeling C4.5 Decision Tree

Every fold is modeled using the C4.5 decision tree method, so that there are n models for each n folds. The C4.5 decision tree method classifies the data by looking for the value of Entropy, Information Gain, Split Info and Gain Ratio. Tree formation begins with finding the highest Gain Ratio value to become the root node, then for leaf nodes it is carried out recursively until a decision tree is formed [6].

The following is an example of a tree formation step:

1. Prepare the data that will be used for the formation of the C4.5 decision tree model. In this example, 9 data on children under five are used for the classification of the BB / U category with the attributes used according to Table 3.
2. Separating data into training data such as Table 4 and testing data as in Table 5 with a total of 3 folds.

Table 3. Dataset

Gender M/F	Body Weight	Age	BB/U
1	8	9	Good
1	7.8	8	Good
1	10.1	8	Good
2	6.1	6	Good
2	4.6	6	Worst
2	10	44	Worst
2	7.3	27	Worst
2	8.9	17	Worst
1	8.1	26	Worst

Table 4. Data Training

Gender M/F	Body Weight	Age	BB/U
1	8	9	Good
1	7.8	8	Good
1	10.1	8	Good
2	6.1	6	Good
2	4.6	6	Worst
2	10	44	Worst

Table 5. Data Testing

Gender M/F	Body Weight	Age	BB/U
2	7.3	27	Worst
2	8.9	17	Worst
1	8.1	26	Worst

3. Calculating entropy using formula (1), information gain using formula (2), split info using formula (3), and calculating the gain ratio using formula (4) for each attribute. The entropy is formulated as

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i * \log_2 p_i. \tag{1}$$

Description of formula (1) follows: S is the set of cases, n is the number of partitions S and p_i is the proportion of S_i to S. The gain is formulated as

$$\text{Gain } (S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy}(S_i). \tag{2}$$

Description of formula (2) follows: S is Sample, A is attribute, n is the number of partitions of the attribute set A, $|S_i|$ is the number of samples on the partition, and $|S|$ is the number of samples in S. Now we formulate the Split Info as

$$\text{SplitInfo}(S,A) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} \times \log_2\left(\frac{|S_i|}{|S|}\right). \tag{3}$$

Description of formula (3) follows: v is the subset resulting from solving using attribute A which has as many as v values. Then, we have the Gain Ratio as

$$\text{GainRatio}(S,A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}. \tag{4}$$

Next, look for the root node candidates by looking for the highest information gain value for each attribute. Determine the root node by finding the highest gain ratio value for each candidate. The highest gain ratio value is found in the weight attribute with a variable value of 4.6, thus the root node of the tree is Weight B. with a value of 4.6. The decision tree formed from the calculation is shown in Figure 2.

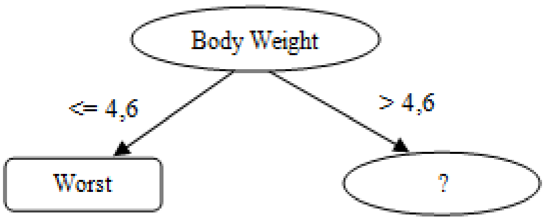

Figure 2. Root Node

4. After getting the root node, then we do a leaf node search. Data with a weight value of 4.6 are deleted / removed from the dataset before searching for leaf nodes (Table 6).

Table 6. The dataset table at node 2

Gender M/F	Body Weight	Age	BB/U
1	8	9	Good
1	7.8	8	Good
1	10.1	8	Good
2	6.1	6	Good
2	4.6	6	Worst
2	10	44	Worst

After it has been removed, it is followed by looking for leaf nodes, and searching for the highest information gain value. The highest information gain value is in the Age attribute with a value of 9, thus the leaf node is Age, if the age is below 9 then the classification label is Good and if it is above 9 then the classification label is Worst. The resulting tree is shown in Figure 3.

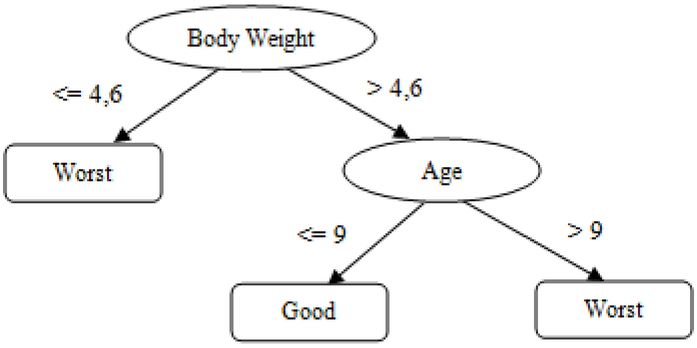

Figure 3. Leaf Node

2.6. Evaluation

Several experiments were carried out to evaluate this system. Each experiment was carried out by dividing the data into 3, 5, 7 and 9 folds. Each experiment was carried out for each category, namely the categories BB/U, TB/U and BB/TB. The experiments are shown in Table 7.

Table 7. C4.5 decision tree experiment

Experiment	Number of Folds
1st	*3-folds*
2nd	*5-folds*
3rd	*7-folds*
4th	*9-folds*

3 Results and Discussion

Based on the experiments, the system is able to classify the nutritional status of the toddler based on BB/U, TB/U and BB/TB. The test results for the BB/U category showed when the number of 3 folds the measured accuracy was 89.52%, when the number of 5 folds the measured accuracy was 90.93%, when the number of 7 folds the measured accuracy was 90.10% and when the number was 9 folds measured accuracy was 90.10%. These results indicate the average level of accuracy is 90.16%. Where the greatest accuracy occurs when using 5 folds (Table 8). This shows that the system can classify the BB/U category well.

Table 8. Results of the BB / U experiment

BB/U		
Experiment	**Number of Folds**	**Average accuracy (%)**
1	3	89.52
2	5	90.93
3	7	90.10
4	9	90.10

While the TB/U category trial showed the average accuracy rate was 76.64% and the highest accuracy occurred at folds 7 (Table 9).

Table 9. Results of the TB/U experiment

TB/U		
Experiment	**Number of Folds**	**Average accuracy (%)**
1	3	75.27
2	5	75.96
3	7	78.32
4	9	77.03

While the BB/TB category trial showed the average accuracy rate was 83.83% and the highest accuracy occurred at folds 7 (Table 10).

Table 10. Results of the BB/TB experiment

BB/TB		
Experiment	**Number of Folds**	**Average accuracy (%)**
1	3	83.27
2	5	83.27
3	7	84.45
4	9	84.34

Based on the test results, we observe that the C4.5 decision tree works well for classifying the categories of BB/U, TB/U and BB/TB using the selected attributes. Although a minority of cases cannot be classified properly.

4 Conclusion

Based on the results of the nutritional classification of children under five using the C4.5 decision tree method, the following conclusions can be drawn:

1. The C4.5 decision tree classification method can be used to classify the nutrition of toddlers quite well.
2. The average accuracy for each category is as follows:
    a. The BB/U category classification has an average accuracy of 90.16%.
    b. The TB/U category classification has an average accuracy of 76.64%.
    c. The BB/TB category classification has an average accuracy of 83.83%.

References

[1] P.T. Juniman. “4 Ancaman Bahaya yang Dialami Balita dengan Gizi Buruk” [Online]. Available: https://www.cnnindonesia.com/gaya-hidup/20180125110614-255-271456/4-ancaman-bahaya-yang-dialami-balita-dengan-gizi-buruk, 2008

[2] Kemenkes. *Hasil Utama Riset Kesehatan Dasar Kementerian Kesehatan 2018* [Online]. Available: https://www.depkes.go.id/resources/download/info-terkini/materi_rakorpop_2018/Hasil%20Riskesdas%202018.pdf. 2018

[3] C. Anam and H.B. Santoso. “Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa,” *Jurnal ENERGY*, **8** (1), 13–19, 2018. [Online]. Available: https://ejournal.upm.ac.id/index.php/energy/article/view/111

[4] U. Febriana, M.T. Furqon, and B. Rahayudi. (2017). “Klasifikasi Penyakit Typhoid Fever (TF) dan Dengue Haemorhagic Fever (DHF) dengan Menerapkan Algoritma Decision Tree C4.5 (Studi Kasus : Rumah Sakit Wilujeng Kediri),” *Jurnal Pengembangan Teknlogi Informasi dan Ilmu Komputer*, **2** (3), 1275–1282, 2017. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1124.

[5] J. Han and M. Kamber. *Data Mining: Concept and Techniques*, Second Edition, Morgan Kaufmann Publishers, 2006.

[6] D.T. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*, John Willey & Sons, Inc., 2005.

This page intentionally left blank