Decision Tree Algorithm Using Particle Swarm Optimization To Improve The Accuracy Of Detection Malnutrition In Toddler

Candra Agustina[1], Nani Purwati[2], Gunawan Budi Sulistyo[3], Noor Hasan[4], Paulus Tofan Rapiyanta[5]
{candra.caa@bsi.ac.id[1], nani.npi@bsi.ac.id[2], gunawan.gnw@bsi.ac.id[3], noor.nhs@bsi.ac.id[4], paulus.pty@bsi.ac.id[5]}

Universitas Bina Sarana Informatika Jakarta, Indonesia[12345]

Abstract. Malnutrition in Indonesia is still relatively high, it is recorded that there are 19.6% of children under five years old who suffer from malnutrition throughout Indonesia. Malnutrition will give impacts tochildren's health in the future. Therefore, the action to detect the malnutrition occurrence should be conducted as early as possible; thus, the patient will immediately getthe right health care. Many methods have already implemented to determine whether a toddler suffers from malnutrition or not. One of them is by using data mining techniques to create a grouping. Toddlers will be categorized into 4 groups namely Good Nutrition, Lack of Nutrition, Over Nutrition and Malnutrition. The data used are Toddler data, which is consisted of 4 predictor attributes and 1 result attribute. In the previous research the algorithm used was C 4.5 that was compared to*Back-propagation*. The result of the data processing by using C 4.5 algorithm is 88.24% and Kappa with the amount of 0.725. In order to improve the accuracy of the C 4.5 algorithm, the algorithm of Particle Swarm Optimization is implementedfor the optimization. Having implemented Particle Swarm Optimization, the accuracy is obtained in the amount of 98.04% and Kappa 0.954. Accordingly, the Particle Swarm Optimization increases the accuracy of C 4.5 by 9.80%. The feature selection, which is conducted, indicates that the attribute of family status must be omitted to obtain higher amount of accuracy.

Keywords: Nutrition status, C 4.5 Algorithm, Particle Swarm Optimization.

1. Introduction

Malnutrition in Indonesia is still relatively high; from the data taken from the Ministry of Health retrieved fromkemkes.go.id, it is stated that the number of children under five years old, which is subsequently mentioned as toddlers, with malnutrition and lack of nutrition status are still recorded relatively in high numbers. In the year of 2010, there were 17.6% of children suffering from malnutrition, this number increased in 2013 reaching the number of 19.6%. Malnutrition is a very crucial case because the growth of toddlers is greatly influenced by the nutritional intake given to them. The impact of this lack of nutrition for toddlers can cause health

ICSTI 2018, October 19-20, Yogyakarta, Indonesia

DOI 10.4108/eai.19-10-2018.2281287

problems in the future. The possible health problems that may occur are osteoporosis, heart diseases, diabetes of type 2, respiratory disorders, and obesity.[1]

The early action to detect this occurrence needsto be done, so that the case is not getting worse; therefore, it is necessary to develop the quickest and the most precise system for analyzing it. One of the proposed methods is by using machine learning based on data mining. It is expected that by implementing this new method, the malnutrition detection will be easily done so that the patients get immediate health care to prevent the bad impacts in the future.

2. Nutrition Status

The nutrition status is a visible result of a person resulting from the input and the excretion of the nutrients to the body. The nutritional substances of one's body are derived from the foods that are consumed based on the categories and indicators used. [2]. The nutritional status of the children can be measured by knowing their age, weight and height. These three variables can be presented with anthropometric standards (BB/U), (TB/U) and BB/TB). For more detail information, itis presented in Table 1.

Table 1. The Elaboration of Toddlers' Nutrition Status Categories [3]

Indicators	Nutritional Status	Z-Score
	Malnutrition	<-3,0 SD
BB/U	Lack of Nutrition	-3,0 SD s/d -2,0 SD
	Good Nutrition	-2,0 SD s/d 2,0 SD
	Over Nutrition	>2,0 SD
	Very Short	<-3 SD
TB/U	Short	-3,0 SD s/d -2,0 SD
	Normal	>= -2 SD
BB/TB	Very thin	<-3,0 SD
	Thin	-3,0 SD s/d -2,0 SD
	Normal	-2,0 SD s/d 2,0 SD
	Fat	>2,0 SD

BB/ U presents the weight of toddler at a certain age.
TB/ U presents the height of toddler at a certain age.
BB/ PB presents the weight of a toddler compared to the height of his body.

3. Data Mining

Data mining is a process for finding and reading the patterns based on the data; hence, it holds a function as a tool to help finding important information from the data [4]. The role of the descriptive data mining is to reveal the pattern of a set of data so that the user can easily interpret it. The predictive data mining uses several variables to estimate the value of other variables, for example classification, regression, etc[5]. Data mining is a part of Knowledge Discovery Data (KDD), which is a process of information that can be used by the users; moreover, the information has not previously known before and it is hidden in a group of data [6].

The stages of Knowledge Discovery Data process shown in Fig 1, consist of:

1. Data Selection

In this stage, the user selecteda set of data that is available. The selected data will be used for data mining process. The data used is stored separately from the operational database.

2. Pre-processing / Cleaning
   The processes were done by removing repetitive data, examining the inconsistent data, and also creatingdata revision. In this process, the user also added existing data with data or other related information.
3. Transformation Coding
   This process is conducted by transforming the data used for research so that the data can be used in data mining.
4. Data Mining
   In this stage the data is processed to search the patterns orthe information containing the patterns or the information by using a particular method.
5. Interpretation
   From the data-mining step, the information pattern will be discovered which is then translated into a form that is more easily understood.[8]

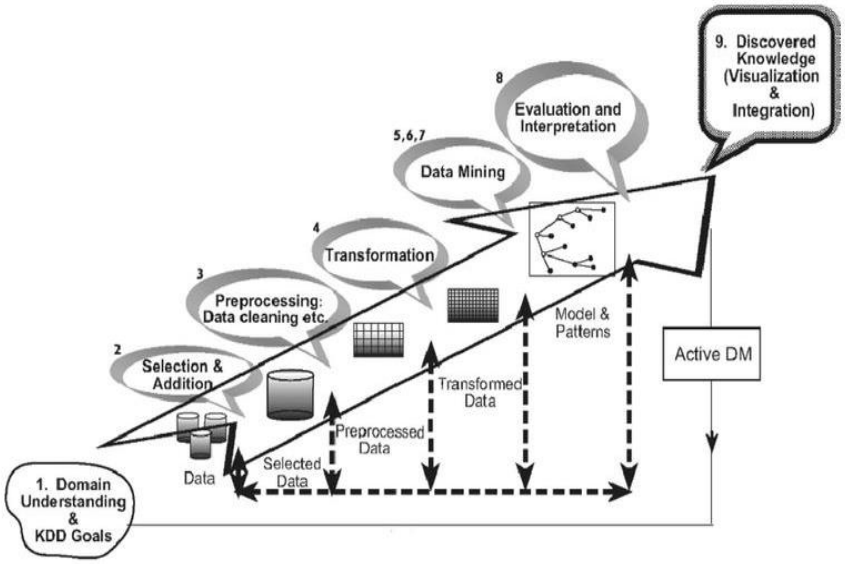

Fig 1. Knowledge Discovery In Databases [7]

4. Algorithm C.45

ID3 Algorithm has been previously known before the public knows C.45 Algorithm. Afterwards, Quinlan developed ID3 (Iterative Dichotomizer) and named it with C.45 Algorithm, a new algorithm that is based on supervised learning.

The C.45 algorithm is also called a decision tree because it has similar structure to a tree. The algorithm has an internal node to describe all attributes; the branches will describe the results of the attributes that are tested, while each of the leaves is presenting the class. The decision tree works from the top of the root. When it is used to process the tested data, for

example is X that is the previous X datais not known yet, then the decision tree works by tracing from the top of the root to the node at each value of the attribute in accordance with the data X tested, this means to check whether it has been appropriate or not with the rules of thedecision tree. Next the decision tree will predict the class of the tuple X.

In implementing this algorithm there are several steps that are need to be done, namely:

1. Preparing Data Training
   This data is taken from the data that already exists and has been classified into certain class groups.
2. Determining Roots
   The root is taken from the selected attribute by calculating the gain value of each attribute. The highest Gain value becomes the main root. In order to know the Gain value, we must know its Entropy value first. The formula for calculating entropy is presented as follows:

$$Entropy(S = \sum_{i=1}^{n} -pi * \log_2 pi \quad (1)$$

   Information:
   S : the set of cases
   A : attribute
   n : the number of partitions S
   pi : the proportion of Si towards S

3. Calculate the Gain Value

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

   Information:
   S : The set of cases
   A : attribute
   n : The number of attribute partition A
   |Si| : the number of cases on the i-th partition
   |S| : the number of cases in S

4. Repeat Step 2 until all tuples are partitioned
5. Partition process will stop if it experiences the following conditions:
   a. All tuples in node N get the same class.
   b. All attributes are already partitioned.
   c. There are no tuples in the empty branch.

5. Swarm Intelligence

The Swarm Intelligence alsocalled,as Swarm Optimization Algorithm is an optimization method based on distributed agent that serves to solve a problem. Swarm Intelligence studied the behavior of a set of social organisms [9].

6. Particle Swarm Optimization

Particle Swarm Optimization is a simple model of the evolution theory that is based on principles derived from the behavior of a group of animals such as birds, fish as well as a bunch of bees in searching thesources of food. When they are firstly looking for food, they did not know where the location of the food source was, but the animals can finally reach the best locations of food sources by communicatingone another[10]

Particle Swarm Optimization has 3 steps that must be conducted, which are:

1. Evaluating the fitness values of the existing particles.
2. Renewing the value of fitness and the best individual position that is commonly called as global best.
3. Updating the speed and position of each particle.[9]

7. Research Metodology

7.1 Data Collection

Toddler data was obtained from the previous research taken from Puskesmas Mranti Purworejo in 2014. The study was conducted towards the data of children aged 0-59 months, with the amount of number 261 toddlers. The attributes used are the names of children, gender, age, weight, height, economic status, nutritional status. Nutrition status is a class that will be determined based on other attributes.

7.2 Research Framework

The raw data is divided into 2 sections (Fig 2): There are 80% of training data and 20% of testing data. Then, C 45 algorithm processed the training data with the support of software Rapidminer. Afterwards,the validation is conducted by using data testing. The results obtained are in the form of accuracy level and kappa. Then, Particle Swarm Optimization algorithm is inserted into the model. The next stage is the reevaluation of the accuracy and Kappa.

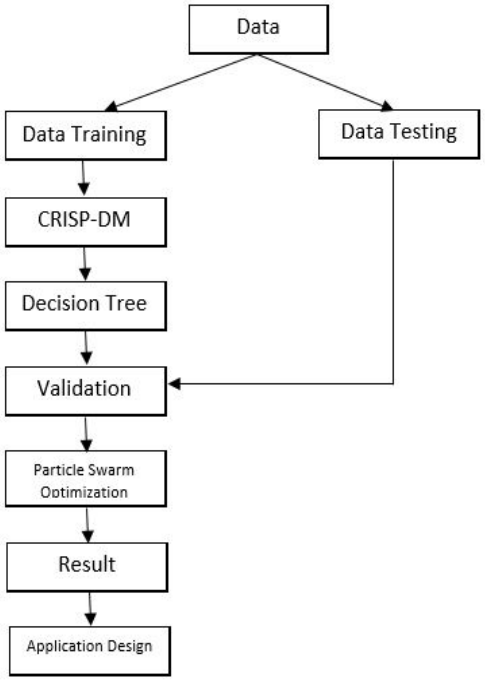

Fig 2.Research Framework

8. Experimental Result

8.1 C.45 Algorithm Implementation

The software used is Rapidminer and it is designed as follows in Fig 3

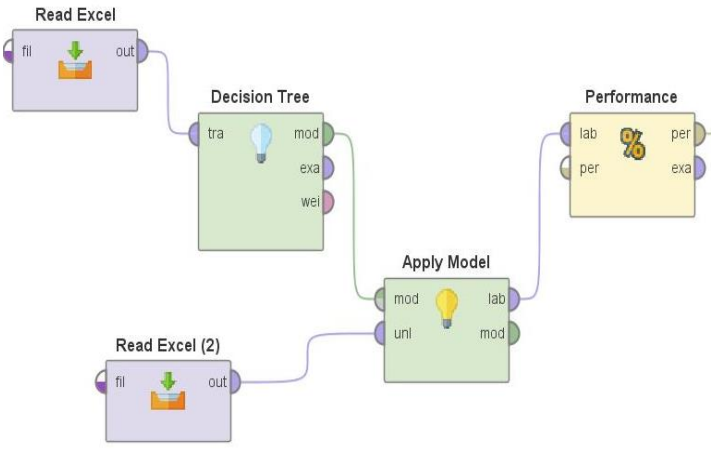

Fig 3. C.45 Design

The results of the data processing using Algorithm C.45 is presented in Table 2 [11].

Table 2. Confusion Matrix C.45

	True Good	True Lack of	True Over	True Bad
Pred Good	35	2	1	0
Pred Lack of	0	9	0	0
Pred Over	1	0	0	0
Pred Bad	0	2	0	1

From the Table 2, it can be obtained the following results:

1. The accuracy rate reaches 88.24% and Kappa reaches 0.725
2. From the total 38 toddlers who are predicted to be good-nutrition, apparently,35 of them are well-predicted, but there are 2 toddlers belong to the group lack of nutrition and 1 toddler belongs to over nutrition, this means that the accuracy reaches 92.11%
3. From the total of 9 toddlers who are predicted to be lack of nutrition,all of them belong to the right group as the prediction, so the accuracy is 100%
4. One toddler who is predicted belongs to the over nutrition, apparently belongs to the good nutrition in fact. It means that the accuracy rate is 0%
5. From the total of 3 toddlers who are predicted to suffer malnutrition, 2 of them belong to the lack of nutrition class and 1 toddler is correctly predicted belongs to the malnutrition class. Thus, the accuracy obtained is 33.33%

8.2 The Application of Particle Swarm Optimization

The design in Rapidminer can be described in Fig 4, Fig 5 and Fig 6

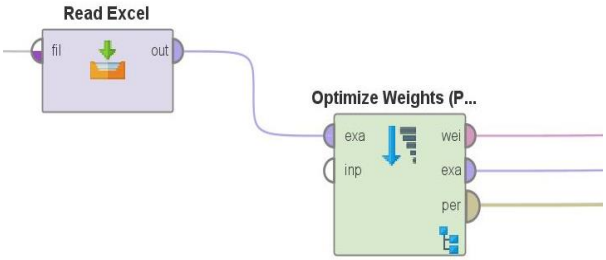

Fig 4. Process

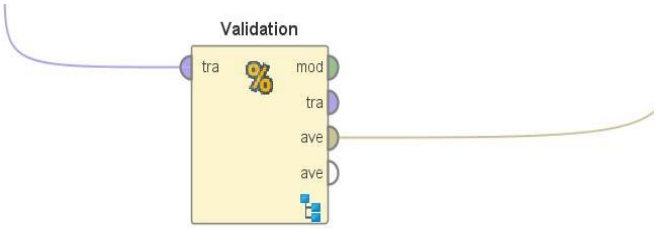

Fig 5. Optimize Weights (PSO)

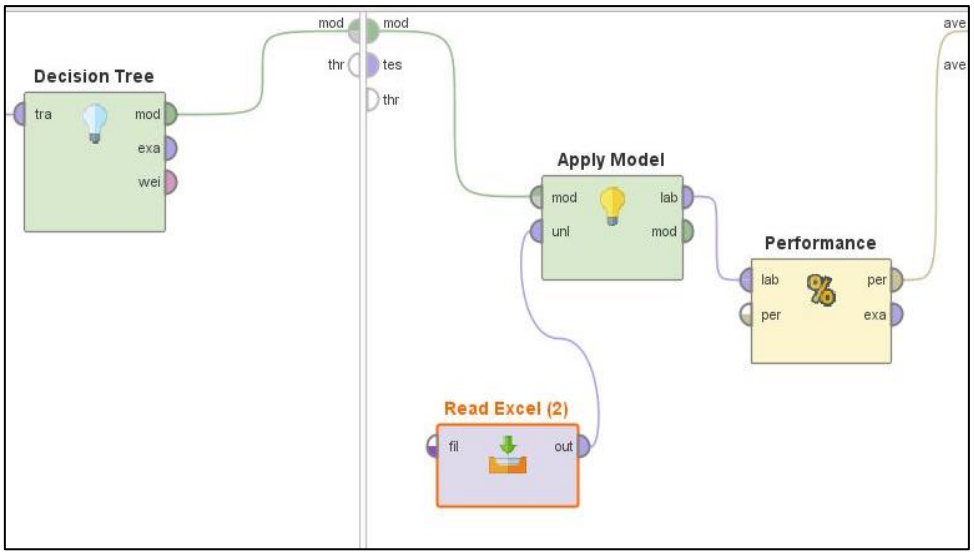

Fig 6. Validation

Data processing is done using rapidminer software, modeling is carried out according to figures 4,5, and 6. The first time the training data is connected to optimize weight, then the validation weight is included in the optimize weight menu. Validation will contain 2 windows, training data and testing data. In the training data window, a decision tree model is entered, while window 1 is shown in Figure 6.
Results obtained in Table 3

Table3. Confusion Matrix C 45 based on Particle Swarm Optimization

	True Good	True Lack of	True Over	True Bad
Pred Good	36	1	0	0
Pred Lack of	0	12	0	0
Pred Over	0	0	1	0
Pred Bad	0	2	0	1

The table data processing with C.45-based Particle Swarm Optimization yields accuracy of 98.04% and Kappa 0.954.

1. From the total 37 toddlers who are predicted to be good-nutrition, apparently 36 of them are well predicted, but there is 1 toddler belongs to the lack of nutrition group, it means that the accuracy reaches 97.30%
2. The toddlers who are predicted to be lack of nutritionare 12 toddlers and the results are accurate, all of them belong to the right group and the accuracy is 100%.
3. The toddlers who are predicted to be over nutrition is 1 toddler and the result is accurate, it gains 100% accuracy.
4. The toddlers who are predicted to be malnutrition is 1 toddler and the result is accurate, it obtains 100% accuracy.

5. From the total 36 toddlers with good status, all of them have been predicted belong to the good group.
6. From the total of 13 Toddlers with lack of nutrition status, there is 1 toddler who is incorrectly predicted.

The comparison of C 45 algorithm result and C 4.5 PSO can be seen in Table 4:

Table 4. Algorithm Comparison

Algorithm	Accuracy	Kappa
C4.5	88,24%	0,725
C4.5 -PSO	98,04%	0,954

The result of the weighing process of the attribute is presented in the table 5.

Table 5. PSO Result

Attribute	Weight
JK	0,762
BB	0,854
U	1
SK	0,096

The family status attribute is removed because it only weighs 0.096.

8.3 Application Design

Use Case shown in figure 7.

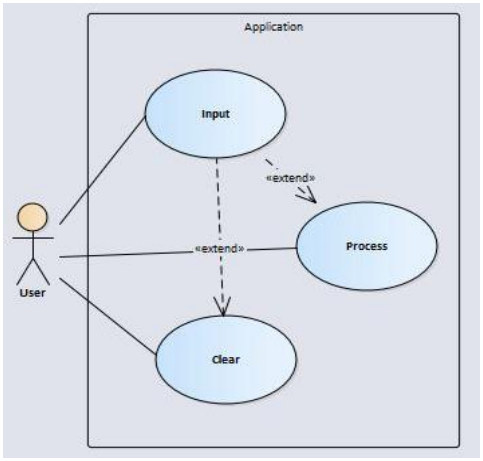

Fig 7. Use Case

User input the form in the application, such as sex, weight, and age, then click the process button. The system processes the data, and show the result. Or the user can click the clear button to clear the textbox.

Table 6 Scenario Use Case

User	System
1. User Input Sex	
2. Input weight	
3. Input Age	
4. Click Process	
	5. Process the data
	6. Display The Status
7. Click Clear	
	8. Clear Text Box

User Interface shown in figure 8.

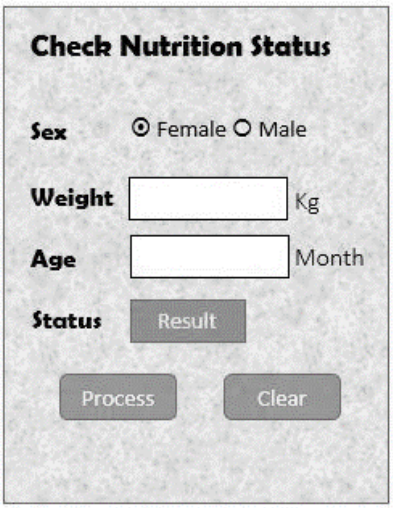

Fig 10. User Interface Design

The Application is built using Visual Basic Programming. The application can be improved at any software such as Android platform.

9. Conclusion

According to this research, it can be summed up that Particle Swarm Optimization is able to increase the accuracy of Decission Tree with the amount of 9.80%. The weighing result shows the family status attribute is worth 0.096 meaning that it must be omitted from Data Set to improve the accuracy.The application can display the result in short time.
It suggested for further research to implement other optimization algorithm in order to find the highest accuracy.

References

[1] R. Candraswari, "Dampak Gizi Buruk Pada Kesehatan Anak yang Harus Diwaspadai," *halosehat.com,* pp. -, 12 April 2018.

[2] Utami Wahyuningsih, Ali Khomsan, Karina Rahmadia Ekawidyani, "Asupan Zat Gizi, Status Gizi, Dan Status Anemia Pada Remaja Laki-Laki Pengguna Narkoba Di Lembaga Pemasyarakatan Anak Pria Tangerang," *Jurnal Gizi Dan Pangan,* vol. 9, no. 1, pp. 23-28, 2014.

[3] K. Kesehatan, "Status Gizi Anak," Kepmenkes, Jakarta, 2010.

[4] Witten, I.H. Frank E, Hall A, Data Mining Practical Machine Learning Tools and Techniques (3rd ed), USA: Elsivier, 2011.

[5] F. Gorunescu, Data Mining Concepts, Models and Techniques, Verlag Berlin Heidelberg: Springer, 2011.

[6] M. Bramer, Principle of Data Mining Second Edition, London: Springer, 2013.

[7] Oded Maimon, Lior Rokach, Data Mining and Knowledge, New York: Springer, 2010.

[8] Kusrini, Luthfi Taufiq Emha, Algoritma Data Mining, Yogyakarta: Andi Offset, 2009.

[9] H. D. Purnomo, "Soccer Game Optimization : Fundamental Concept," *Metris,* vol. 15, no. 2, pp. 25-30, 2014.

[10] Argha Roy, Diptam Dutta, Kaustav Choudhury, "Training Artificial Neural Network Using Particle Swarm Optimization Algorithm," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 3, pp. 430-434, 2013.

[11] Nani Purwati, Candra Agustina, Gunawan Budi Sulistyo, "Komparasi Algoritma C 4.5 Dan Backpropagation Untuk Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri BB/U dan BB/PB," *Sentra Penelitian Engineering Dan Edukasi,* vol. 9, no. 3, pp. 26-33, 2017.