The comparison of accuracy between naïve bayes classifier and c4.5 algorithm in classifying toddler nutrition status based on anthropometry index

A Ridwan\(^1\) and T N Sari\(^2\)

\(^1\)Departement of Information System, Universitas Muhammadiyah Kudus, Kudus 59316, Indonesia
\(^2\)Departement of Nurses Profession, Universitas Muhammadiyah Kudus, Kudus 59316, Indonesia

*achmadridwan@umkudus.ac.id

Abstract. Body Mass Index (BMI), one of which is used to indicate the category of nutritional status of children under five whether it is proportional or not. Anthropometry has an important role to play in determining the nutritional status of Toddler. The Anthropometric guidelines for determining the nutritional status of Toddler are the parameters chosen which include an assessment of height, age, and weight. On the other hand this ever-increasing amount of data requires several methods to process and draw conclusions and information from these data. Some methods used to process large data to find patterns contained therein are simple, including the Naïve Bayes Classifier Algorithm and C4.5 Algorithm. Both are methods in the classification of Data Mining Techniques. The Naïve Bayes Classifier Algorithm and C4.5 Algorithm are included in the Ten Most Popular Data Mining Classifications. This Research will apply the Naïve Bayes Classifier Algorithm and C4.5 Algorithm which is used for the classification of Toddler Nutrition Status based on the Anthropometry Index so as to produce the accuracy of both algorithms. While the five classes will be classified as thin, very thin, normal, fat, very fat. The results of this classification process show that the C4.5 algorithm has an accuracy rate of 0.93% better than the Naïve Bayes Classifier Algorithm. The algorithm that has been formed can then be developed and implemented into an application making it easier for stakeholders to make a decision for the Classification of Toddler Nutrition Status based on the Anthropometric Index

1. Introduction

Indonesia is a developing country. One of the government's concerns is in the Health Sector. There is still a lot of homework that needs to be done especially about Toddler Nutrition. Because Toddler Nutrition is a crucial problem to be addressed and needs to get the priority of Toddlers, including a very important level in the phase of human life. If there are health problems at this level, then most likely these health problems will have an impact on the next level of life. Health development in the 2015-2019 period focused on four priority programs, namely reducing maternal and infant mortality rates, reducing the prevalence of short toddler (stunting), controlling infectious diseases and controlling non-communicable diseases. Efforts to improve the nutritional status of the community,
including reducing the prevalence of short toddler, are among the national development priorities listed in the main targets of the 2015 - 2019 Medium-term Development Plan [1].

Nutrition Information System in Indonesia, especially regarding Nutrition Status Monitoring (PSG) for toddlers now exists, but there are still a number of obstacles that hamper such as late reports to the District Health Office. In fact, sometimes there are only recorded cases for a number of Puskesmas, so the number of malnutrition cases does not reflect the real situation. This condition ultimately results in data and information that is incomplete and not timely so that it is less able to support leaders or policy makers in the decision making process precisely and quickly. The amount of data on infant nutrition that continues to increase requires several methods to process and draw conclusions and information from the data, which is expected to be able to improve the quality of data or information as well as the efficiency and effectiveness of data processing. So that in the end it facilitates or helps in policy making, especially in overcoming nutrition problems in Toddlers.

Anthropometric measurements are measurements used to determine a person's nutritional condition. Determination of the nutritional condition includes an assessment of age and weight, body length, or height, and upper arm circumference. However, the large number of toddler makes it difficult to determine which nutritional data for toddler are those whose nutritional status is normal, fat, obese, thin, or very thin. In this study took data in 2013.

During the last decade there has been an explosion in computing and information technology. With that a large amount of data has emerged in various fields such as medicine, biology, finance, and marketing [2]. The method for processing and drawing conclusions and information from these data is Data Mining. Research on nutrition for toddler using the Data Mining classification has been carried out, either a comparison of several Data Mining models classifications or improvements to the Data Mining classification.

In this comparative study, the techniques to be used are the NBC and C4.5 Algorithms, because these two algorithms are among the 10 best Data Mining Algorithms that were explored by the IEEE International Conference on Data Mining (ICDM) in December 2006 expected: C4.5, k-Means, SVM, Apriori, EM, PageRank, Adaboost, kNN, Naïve Bayes, and CART [3].NBC and C4.5 algorithms can be used to handle multi-class problems [4]. Nevertheless, this research will prove the ability of Naive Bayes and C4.5 classifiers in the Classification Of toddler Nutrition Status Based On Anthropometry Index in order that the predicted results will be more accurate information. From this information can help or assist in making special policies in overcoming nutritional problems in toddler

2. Literature Review

2.1. Data Mining

Data Mining (DM) is the core of the Knowledge Discovery in Database (KDD) process, which involves an algorithm that exploits data, develops models and discovers previously unknown patterns. This model is used to understand phenomena from data, analysis and prediction. Accessibility and lots of data make Knowledge Discovery and Data Mining become a problem that is quite important and needed. Based on their duties, Data Mining is divided into 6 groups namely Description, Estimation, Prediction, Classification, and Clustering [5].

2.2. Naïve Bayes Classifier Algorithm

Naïve Bayes Classifier is a statistical classification that can be used to predict the probability of membership of a class. Naïve Bayes Classifier is based on the Bayes theorem which has the same classification capabilities as decision trees and neural networks. Naïve Bayes Classifier is proven to have high accuracy and speed when applied to databases with big data [6].

The basic form of tasks performed by the Naïve Bayes algorithm is just a classification [7]. Naïve Bayes Classifier can also handle class imbalances by additionally applying AdaBoost to the Naïve Bayes Classifier algorithm [8]. Bayes Theorem has the following general form:

\[
P(H | X) = \frac{P(X | H)P(H)}{P(X)}
\]
2.3. C4.5 Algorithm
C4.5 algorithm is the result of the development of the ID3 (Iterative Dichotomiser) algorithm developed by Quinlan [9]. This algorithm is used to build a decision tree that is easy to understand, flexible, and interesting because it can be visualized in the form of images [10]. Previously in the late 1970s until early 1980s J. Ross Quinlan, a researcher in the field of machine learning, made a decision tree algorithm known as ID3 (Iterative Dichotomiser). For ID3, attribute selection measurement is determined by Information Gain, while C4.5 attribute selection measurement is determined by GainRatio [11].

C4.5 Algorithm or Decision Tree resembles a Tree where there are internal nodes (not leaves) that describe the attributes, each branch represents the result of the attribute being tested, and each leaf represents the class. Decision trees can easily be converted to classification rules. In general, tree classification decisions have good accuracy, but successful use depends on the data to be processed. Decision tree is a classification method that uses tree structure representation where each node represents an attribute, the branch represents the value of the attribute, and the leaf represents the top Node class of the decision tree called the root. As for the formula for finding Entropy as below [12]:

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

Information:
S = Case set
n = number of partitions S
pi = proportion of S, to S

2.4 Rapid Miner
Rapid Miner is an open-source Data Mining tool and software available to assist in cluster analysis, data visualization, regression analysis, decision trees, predictive analytics, text mining. Rapid Miner includes software that provides a series of methods and algorithms that help in better data analysis [13].

2.5. The Confusion Matrix Evaluate
To evaluate the classification model based on the calculation of the testing object which is predicted to be true and incorrect. Confusion Matrix contains information about actual (actual) and predicted (predicted) in the classification system. System performance like this is usually evaluated using data in the matrix. This calculation is tabulated into a table called a confusion matrix [14].

The calculation formulation used to calculate accuracy is as follows [15]:

\[
Accuracy = \frac{TP+TN}{TP+FN+FP+TN}
\]
3. Research Methods

Research conducted using the type of experimental research is to do a comparison algorithm between Naïve Bayes Classifier and C4.5 algorithm from the processing of Toddler Nutrition Status Dataset Based on Anthropometric Index then compared to find out the most accurate classification in predicting Toddler Nutrition Status Based on Anthropometric Index.

This study uses data obtained from research results in the Wonosobo District Health Office in the form of a toddlers nutritional status based on the Anthropometry Index sample from the Public Health Centers in each District in Wonosobo which will be classified according to their class. The amount of data obtained is 530 samples, consisting of 5 classes, namely Very Thin, Thin, Normal, Fat, and Very Fat. Nutrition Status data obtained is a collection of data from nutrition status measurements in July to September 2013.

The study was conducted in stages namely data collection, data selection, age grouping for toddlers and grouping of nutritional status according to the Decree of the Minister of Health of the Republic of Indonesia Number: 1995 / MENKES / SK / XII / 2010 [16]. Concerning Anthropometry Standards for the Assessment of Child Nutrition Status where the decision refers to World Health Organization Standards (WHO 2005) [17], followed by Preprocesing data

The initial stage of data Preprocessing is data cleaning and data selection by removing data that has incomplete information. Then the last step is Evaluation carried out by testing the Dataset of Toddler Nutrition Status based on the Anthropometric Index to the RapidMiner Tool Then the Confusion Matrix Evaluation is performed by testing the Naïve Bayes Classifier and C4.5 Algorithm in order to find the Accuracy Value to determine the difference between the Naïve Bayes Classifier and C4.5 Algorithm methods.

4. Result and Discussion

Experiments were carried out using a Computer platform based on Intel Core i3 @ 3.30GHz, 4GB RAM, and Microsoft Windows 10 64-bit Operating System. While the application development environment with the Java Programming Language Netbeans Pro 8.0 and Rapidminer Studio 7, for analysis of results using the Excel Data Analysis application. The nutritional status data for toddlers can be used as a dataset with 530 original records, consisting of 5 classes: Very Thin, Thin, Normal, Fat, and Very Fat. Of these, the data already have complete information in each attribute. Then the data is selected.

This selection is done to classify attributes according to the information needed. From the attributes selected, the attributes are Age , Weight / Age, Height / Age, Weight / Height, and BMI / age. While the Result attribute is the target attribute.

In this study the process of data cleaning and data selection (preprocessing data) is carried out to discard data that has incomplete information. It is possible that a lot of data has incomplete information, such as the absence of age information, weight information, height information, Body Mass Index information. After cleaning the data, clean data is obtained as much as 430 data. The results of Nutrition Status Data after the stages of cleaning data and selection data are as follows:

| NO | Age (month) | Weight (kg) | Height (cm) | BMI | Nutrition Status |
|----|-------------|-------------|-------------|-----|------------------|
|    |             |             |             |     | Weight/ Age      |
|    |             |             |             |     | Height/ Age      |
|    |             |             |             |     | Weight/ Height   |
|    |             |             |             |     | BMI/ Age         |
|    |             |             |             |     | Result           |
| 1  | 17          | 8.8         | 73          | 16.51| Normal           |
|    |             |             |             |     | Very Short       |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
| 2  | 20          | 10.1        | 77          | 17.03| Normal           |
|    |             |             |             |     | Short            |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
| 3  | 19          | 10.5        | 76          | 18.18| Normal           |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
|    |             |             |             |     | Normal           |
After that, in this experiment, the Naïve Bayes Classifier Algorithm and C4.5 for Classification of Toddler Nutrition Status based on the Anthropometry Index are applied to the Rapid Miner tool with the Confusion Matrix Evaluation test, then the accuracy test results obtained are obtained as shown below:

![ACCURACY](image)

**Figure 1.** Comparison of accuracy of naïve bayes classifier and C4.5

In the picture above, the accuracy value of the Naïve Bayes Classifier algorithm model = 88.60% and the accuracy value for the C4.5 model = 89.53% with an accuracy difference of 0.93%.

5. Conclusions

This study compared the Naïve Bayes Classifier and C4.5 Algorithm which was implemented with the dataset of Toddler Nutrition Status based on the Anthropometry Index. Naïve Bayes Classifier algorithm produces an accuracy of 88.60%. And the C4.5 algorithm produces 89.53% accuracy. So that the accuracy of C4.5 Algorithm has an accuracy difference of 0.93% better than the Naïve Bayes Classifier algorithm. The algorithm that has been formed can then be developed and implemented into an application making it easier for Stakeholders to make a decision for the Classification of Toddler Nutrition Status based on the Anthropometric Index.
Acknowledgment
The Author would like to thank the Rector of the University of Muhammadiyah Kudus, Rusnonto, SKM., M. Kes (Epid), and all those who have helped the writer in preparation of this Journal.

References
[1] Mebus G 1998 *Pusat Data dan Informasi Kementerian Kesehatan RI* (Jakarta: ACM SIGAPL APL Quote Quad)

[2] Sowmya R and Sunetha K R 2017 Data Mining with Big Data *Proceedings of 11th International Conference on Intelligent Systems and Control*

[3] Wu X and Kumar V 2009 *The Top Ten Algorithms in Data Mining* (London: Taylor & Francis Group)

[4] Sharma P, Singh D and Singh A 2015 Classification algorithms on a large continuous random dataset using rapid miner tool *2nd International Conference on Electronics and Communication Systems*

[5] Cheng S, Liu B, Shi Y, Jun Y and Li B 2015 *Data Mining and Big Data*

[6] Kotsiantis S B and Pintelas P E 2016 Selective costing ensemble for handling imbalanced data sets *Int. J. Hybrid Intell. Syst.*

[7] Özkan S B, Apaydin S M F, Özkan Y and Düzdar I 2019 Comparison of open source data mining tools: Naive Bayes algorithm example *Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science*

[8] Ridwan A, Andono P N and Supriyanto C 2018 Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma Naive *Teknol. Inf.*

[9] 2017 A Comparative Study of Classification Techniques in Data Mining Algorithms *Int. J. Mod. Trends Eng. Res.*

[10] Gorunescu F 2011 *Data Mining* (Berlin: Springer Berlin Heidelberg) vol 12

[11] Gupta B, Rawat A, Jain A, Arora A and Dhami N 2017 Analysis of Various Decision Tree Algorithms for Classification in Data Mining *Int. J. Comput. Appl.*

[12] Heredia D, Amaya Y and Barrientos E 2015 Student Dropout Predictive Model Using Data Mining Techniques *IEEE Lat. Am. Trans.*

[13] Naik A and Samant L 2016 Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime *Procedia Computer Science*

[14] Caelen O 2017 A Bayesian interpretation of the confusion matrix *Ann. Math. Artif. Intell.*

[15] Ruuska S, Hämäläinen W, Kajava S, Mughal M, Matilainen P and Mononen J 2018 Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle *Behav. Processes*

[16] 2010 Keputusan Menteri Kesehatan Republik Indonesia Nomor : 1995/Menkes/SK/XII/2010 tentang Standar Antropometri Penilaian Status Gizi Anak (Jakarta: Kementerian Kesehatan RI)

[17] 2006 *Multicentre Growth Reference Study Group WHO Child Growth Standards: Length/height-for-age, Weight-for-age, Weight-for-length, Weight-for-height and Body massindexfor-age: Methods and Development* (Geneva: World Health Organization) pp 76-85.