Implementation of C4.5 Algorithm for Classification of Nutritional Status of Toddlers

Alfaeni Syafa Safira
Yogyakarta University of Technology
Yogyakarta, Indonesia

Arief Hermawan
Yogyakarta University of Technology
Yogyakarta, Indonesia

ABSTRACT
Nutrition is an important part in the growth of toddlers. Monitoring of nutritional status is needed. The data mining method used to classify the nutritional status of toddlers using C4.5 algorithm. The nutritional status of toddlers is divided into five classes, namely undernutrition, good nutrition, risk of overnutrition, overnutrition and obesity. There is an imbalance of data in the five classes. This data imbalance is handled using Synthetic Minority Oversampling Technique (SMOTE). From the research that has been conducted, the application of SMOTE in the classification of nutritional status of toddlers can influence the value of the model evaluation. Before SMOTE was applied, the classification model produced 86% accuracy, 87% precision, 86% recall, 85% f1-score, and 33% mean absolute error. After implementing SMOTE, it can increase the accuracy value to 90%, precision to 91%, recall to 90%, f1-score 90%, and can reduce the mean absolute error value to 22%.

Keywords
Classification, Nutritions, Toddlers, C4.5, SMOTE

1. INTRODUCTION
Nutrition is an important factor in the growth and development process of toddlers [1]. The nutritional status of toddlers representation of the size of the fulfillment of the nutritional needs of toddlers obtained from the intake and use of nutrients by the body [2]. Factors that affect the nutritional status of toddlers are environmental health, food availability at the family level, family parenting, family culture, environmental health, and socioeconomics [3]. Good nutrition in toddlers allows optimal brain development, good physical development, and general health at the highest possible level [4]. Meanwhile, nutritional problems in toddlers will have a negative impact, namely the disruption of physical and psychological development, which will cause inhibition of productivity, creativity and reduce the intelligence of toddlers. In addition, the impact of nutritional problems in toddlers can lead to a decrease in endurance and will have an impact on the healthy life span of toddlers, increase the morbidity, and mortality rate of toddlers [5]. Therefore, monitoring and evaluating the nutritional status of toddlers is an important step in efforts to prevent and treat nutritional problems in toddlers by implementing data mining.

Data mining will be applied to the toddler nutrition data in order to optimize the classification of nutritional status of toddlers. Data mining is a method used to uncover valuable information hidden in large data collections, with the aim of finding interesting patterns that were previously unknown [6]. There are several types of data mining techniques which include classification, clustering, and prediction [7]. Classification is one of the concepts in data mining. It refers to a technique for predicting data values by using information derived from other data. The goal is to anticipate the value of an unknown variable by using information from other existing variables. Classification is often referred to as supervised learning because it involves using predefined classes or labels when processing data [8]. One of the classification methods is the decision tree. One algorithm that can be used to create a decision tree is the C4.5 algorithm. The C4.5 algorithm is a further development of the ID3 algorithm proposed by Quinlan. In Algorithm C4.5, information gain and gain ratio are used as measures when selecting attributes [9]. One of the things that can affect the accuracy of a classification model is the problem of data imbalance. Data imbalance causes the resulting model to be more accurate in classifying majority data but less accurate for classifying minority data [10]. Synthetic Minority Oversampling Technique (SMOTE) is one method to balance the data. Synthetic Minority Over-sampling Technique (SMOTE) is a method used to overcome the problem of imbalanced sample data, especially in the majority class. The main function of SMOTE is to improve the performance of classification methods by generating additional samples for the minority class [11]

Therefore, based on the description above, this study aims to examine the classification of nutritional status of toddlers by balancing data with the SMOTE method. The end result is to know how the effect of SMOTE is used to classify the nutritional status of toddlers with the C4.5 algorithm.

2. LITERATURE REVIEW
Several researchers have applied the C4.5 methodology for classification. Some of these studies are outlined in the following sections.

Istiawan et al. [12] stated that the C4.5 algorithm can be used to predict Agricultural Cultivation Areas in Pemali Jratun Watershed. This research uses data obtained from BPDAS Pemali Jratun in 2013, the data includes productivity, slope, erosion, land management, and land criticality. This research resulted in an accuracy rate of 92.47%.

Ledoh et al. [13] stated that the C4.5 algorithm can be used to predict Student Satisfaction Level of Lecturer's Performance in the Covid-19 Pandemic. This research uses data on active students majoring in computer science from 2016-2022 Nusa Cendana University, the data includes Age, Gender, Suitability of Learning Media (SLM), Pedagogic Competence (PeC), Professional Competence (PrC), Personality Competence (PsC), and Social Competence (SC). This research resulted in an accuracy rate of 94.8%.

Tamsir [14] stated that the C4.5 algorithm can be used to classify cat health. This study uses medical record data from March to December 2019, the data includes Condition General and Condition Physical. This research resulted in an accuracy rate of 93.18%.

3. METHODS

3.1 Dataset

The data used in this study came from Puskesmas 2 Ajibarang in 2023 with a total sample size of 160. The variables used were 5 variables which consisted of 1 target variable, namely, nutritional status (0 = Good Nutrition; 1 = Under Nutrition; 2 = Obesity; 3 = Over Nutrition; 4 = Risk of Over Nutrition) and 4 predictor variables. The predictor variables used include X1: JK (0=L; 1=P); X2: age at measurement; X3: weight; X4: height. Dataset information is presented in Table 1.

Table 1. Dataset Information

No	Variable	Description	Type of Data
1	JK	Gender of toddlers	Categorical
2	Age at measurement	Age at the time the toddler was measured	Numerical
3	Weight	Toddler's weight	Numerical
4	Height	Toddler's height	Numerical
5	Nutritional Status	Nutritional status in toddlers	Categorical

3.2 Research Stages

The stages carried out in this research use python programming language tools. The stages of data analysis are as follows:

1. Collecting data at Puskesmas 2 Ajibarang
2. Preprocessing by changing the value of the variables "JK" and "Nutritional Status". In the "JK" variable, the value of "L" was changed to "0" and "P" to "1". For the variable "Nutritional Status" the value of "good nutrition" was changed to "0", then "undernutrition" to "1", "obesity" to "2". "overnutrition" becomes "3", and "risk of overnutrition" becomes "4".
3. Conduct data exploration and descriptive statistical analysis to see the general characteristics of all variables to be analyzed.
4. Randomly splitting the data into training data and test data using the 10-fold Cross-Validation method.
5. The classification process with the C4.5 algorithm will be carried out k=10 times using the partition data as test data and the rest as training data.
6. Measure classification performance by calculating the average of all training results that have been done.
7. Balancing the data by creating artificial synthesized data using SMOTE.
8. Perform classification with the C4.5 algorithm with new training data that has gone through the SMOTE process k = 10 times as test data and the rest as training data.
9. Evaluate the classification performance by calculating the average result of all training processes of the classification model that has been processed with SMOTE.
10. Comparing the model evaluation results between the data that has been processed with SMOTE and the original data to see the impact of SMOTE performance on imbalanced data.

4. RESULT AND DISCUSSION

4.1 Descriptive Analysis

The data used in this study were 5 variables. Variable y is a target variable that has 5 classes, namely undernutrition, good nutrition, risk of overnutrition, overnutrition, and obesity. The x variable is a predictor variable consisting of 4 variables, of which there are 3 predictor variables of numeric type, namely the variables of age at measurement, weight, and height. There is also 1 variable with categorical data type, namely JK (gender). Data exploration aims to see the characteristics of each category. Previously, it has been mentioned that there are data imbalances that occur in the target variable, namely nutritional status. The proportion of imbalances in the five categories is shown in Figure 1 below.

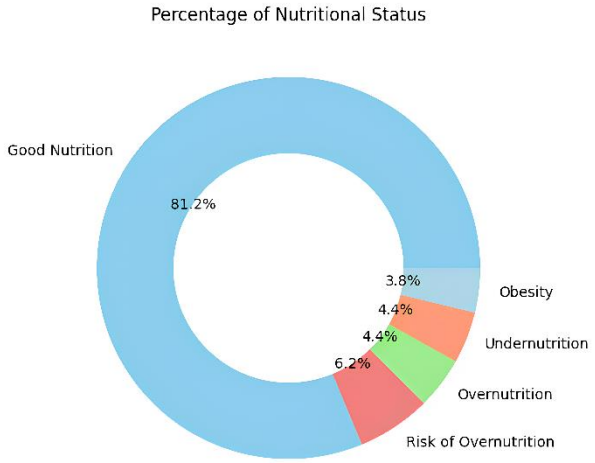

Fig 1: Diagram of Variable Percentage

4.2 Split Dataset

The process of dividing training data and test data in this study uses the k-fold cross validation method with k = 10. The amount of data used is 160 observations with the division of training data and test data illustrated as in Figure 2.

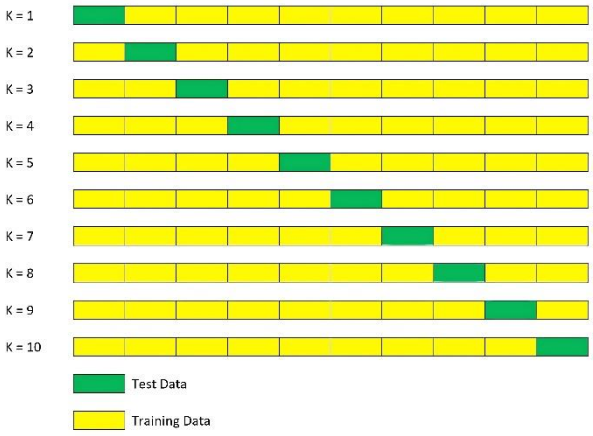

Fig 2: Illustration of Split Training Data and Test Data

4.3 Modeling Classification of Toddlers Nutritional Status Using C4.5 Algorithm

The model formed from each fold will be tested for performance using test data. Testing is done using Confusion Matrix to determine accuracy, precision, recall, f1-score, and mean absolute error (MAE). The C4.5 algorithm classification process using 10-fold cross-validation produces different levels of accuracy, precision, recall, f1-score and mean absolute error (MAE) for each k value used. The results of each fold are presented in Table 2.

Table 2. Evaluation of C4.5 Algorithm

Fold	Accuracy	Precision	Recall	F1-score	MAE
Fold 1	69%	78%	69%	68%	50%
Fold 2	88%	94%	88%	88%	25%
Fold 3	94%	100%	94%	97%	25%
Fold 4	94%	88%	94%	91%	25%
Fold 5	94%	88%	94%	91%	25%
Fold 6	81%	94%	81%	87%	38%
Fold 7	94%	91%	94%	92%	6%
Fold 8	75%	62%	75%	68%	56%
Fold 9	81%	73%	81%	76%	50%
Fold 10	81%	90%	81%	84%	44%

4.4 Modeling C4.5 Classification Using SMOTE

The problem of data imbalance can be overcome by using the SMOTE method applied to the training data. The SMOTE technique is used to overcome data imbalance by generating synthetic data. The new data generated will then be combined with the original data. Table 3 contains a comparison of the original training data with the training data that has been processed with SMOTE

Table 3. Percentage of Data Before and After SMOTE

Categories	Number of Original Data (%)	Number of Data after SMOTE (%)
0 = Good nutrition	130 (81.2%)	130 (20%)
1 = Under nutrition	7 (4.4%)	130 (20%)
2 = Obesity	6 (3.8%)	130 (20%)
3 = Overnutrition	7 (4.4%)	130 (20%)
4 = Risk of overnutrition	10 (6.2%)	130 (20%)
Total	160 (100%)	650 (100%)

After balancing the data using the SMOTE technique, the new training data is used to build the new classification model. This process uses the same method, namely 10-fold cross-validation and the C4.5 algorithm. Model performance evaluation was performed on the test data, resulting in classification performance as shown in Table 4 below.

Table 4. Evaluation of C4.5 Algorithm with SMOTE

Fold	Accuracy	Precision	Recall	F1-score	MAE
Fold 1	94%	94%	94%	94%	6%
Fold 2	89%	90%	89%	89%	31%
Fold 3	92%	92%	92%	92%	22%
Fold 4	89%	90%	89%	89%	12%
Fold 5	97%	97%	97%	97%	5%
Fold 6	89%	90%	89%	89%	23%
Fold 7	85%	86%	85%	84%	38%
Fold 8	98%	99%	98%	98%	6%
Fold 9	86%	86%	86%	86%	37%
Fold 10	89%	89%	89%	89%	18%

4.5 Comparison of Classification Models

Comparison of model evaluation results using SMOTE and without using SMOTE is done by comparing the average results of accuracy, precision, recall, f1-score, and mean absolute error (MAE) on each model which can be seen in Table 5.

Table 5. Comparison of Classification Results Without SMOTE and Using SMOTE

Criteria	Model	
	Without SMOTE	Using SMOTE
Accuracy	86%	90%
Precision	87%	91%
Recall	86%	90%
F1-score	85%	90%
MAE	33%	22%

Table 5 above shows the average performance of the two models based on the results of ten machine learning experiments. The classification model on the data before using SMOTE produces an average accuracy of 86%, while after handling with SMOTE to balance the data, the average accuracy increases to 90%. Precision is the degree of accuracy of the information requested by the user with the results obtained. The precision values obtained before and after SMOTE were 87% and 91%. Recall is the ratio of true positive predictions to the overall true positive data. The recall results obtained before and after SMOTE are 86% and 90%. Furthermore, the f1-score value is an alternative result of accuracy and results obtained. The f1-score values before and after SMOTE are 85% and 90%. The mean absolute error (MAE) value is the absolute average of the forecasting error regardless of positive or negative signs. The mean absolute error (MAE) value before SMOTE was 33% and after SMOTE it became 22%.

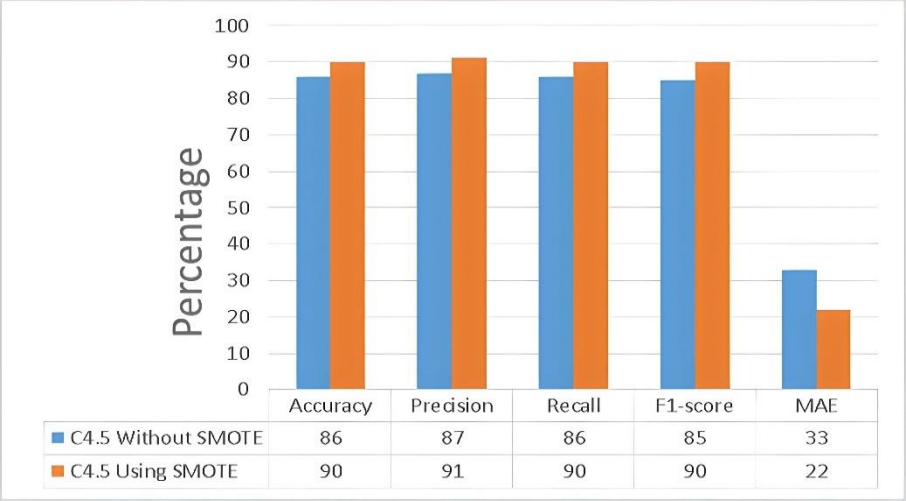

Fig 3: Comparison of Average Model Performance

From the comparison of the average performance of the two models as shown in Figure 3, it can be seen that the use of the SMOTE method has successfully improved the classification results using the C4.5 algorithm. This is evidenced by the increase in accuracy, precision, recall, and f1-score values. The model after SMOTE can also reduce the model prediction error because the mean absolute error has decreased after SMOTE.

5. CONCLUSION

From the results of the toddlers nutritional status classification research using the C4.5 algorithm with the help of python tools evaluated with 10-fold Cross-Validation resulted in an average accuracy of 86%, precision 87%, recall 86%, f1-score 85%, and mean absolute error (MAE) 33%, after the SMOTE process can increase the accuracy value to 90%, precision to 91%, recall to 90%, f1-score 90% and mean absolute error (MAE) decreased to 22%. So the use of SMOTE in this study can optimize the performance of the toddler nutritional status classification process using the C4.5 algorithm.

Based on the research results, suggestions for developing this research are to balance the data using the ADASYN (Adaptive Synthetic Sampling), Borderline-SMOTE, SMOTE-Tomek methods and use other data mining classification methods such as ensemble classification methods (Bagging, AdaBoost), namely combining several methods as a classification solution to get the best results.

6. REFERENCES

[1] W. Aerin and M. Muqowim, “Indonesian Journal of Early Childhood Implementation of Children Nutrition Meeting Through Healthy Eating Program,” vol. 9, no. 1, pp. 48–52, 2020.

[2] R. A. Wardani and V. Virgia, “Analysis of Factors Influencing Stunted Toddlers in the City of Mojokerto,” vol. 14, no. 02, pp. 162–175, 2022.

[3] A. Aziz, A. Hidayat, G. Marini, A. Pangestu, and M. Tyas, “Factors Affecting Nutritional Status in Children Aged 6 – 24 months in Lamongan Regency , Indonesia,” vol. 8, pp. 291–295, 2020.

[4] C. J. Valentine, “Nutrition and the developing brain,” Pediatr. Res., no. October 2019, pp. 190–191, 2020, doi: 10.1038/s41390-019-0650-y.

[5] A. Soliman et al., “Early and Long-term Consequences of Nutritional Stunting : From Childhood to Adulthood,” vol. 92, no. 4, pp. 1–12, 2021, doi: 10.23750/abm.v92i1.11346.

[6] J. L. Pastrana, R. E. Reigal, V. Morales-sánchez, and A. Hernández-mendo, “Data Mining in the Mixed Methods : Application to the Study of the Psychological Profiles of Athletes Data Mining Added to Mixed Methods,” vol. 10, no. December, pp. 1–8, 2019, doi: 10.3389/fpsyg.2019.02675.

[7] M. A. Saleh, S. Palaniappan, N. Ali, and A. Abdalla, “Education is An Overview of Data Mining and The Ability to Predict the Performance of Students,” vol. 15, no. 1, pp. 19–28, 2021.

[8] A. B. E. D. Ahmed and I. S. Elaraby, “Data Mining: A prediction for Student’s Performance Using Classification Method,” World J. Comput. Appl. Technol., vol. 2, no. 2, pp. 43–47, 2014, doi: 10.13189/wjcat.2014.020203.

[9] Y. Wang, “Prediction of Rockburst Risk in Coal Mines Based on a Locally Weighted C4 . 5 Algorithm,” vol. 9, pp. 15149–15155, 2021, doi: 10.1109/ACCESS.2021.3053001.

[10] C. Kaope and Y. Pristyanto, “The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance,” vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[11] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” Mach. Learn., no. 0123456789, 2023, doi: 10.1007/s10994-022-06296-4.

[12] D. Istiawan, L. Khikmah, S. M. Semarang, A. Info, C. Land, and D. Mining, “Implementation of C4 . 5 Algorithm for Critical Land Prediction in Agricultural Cultivation Areas in Pemali Jratun Watershed,” vol. 2, no. 2, pp. 67–73, 2019.

[13] J. R. M. Ledoh, F. E. Andreas, E. S. Y. Pandie, E. Clarissa, and A. Pah, “C4 . 5 Algorithm Implementation to Predict Student Satisfaction Level of Lecturer ’ s Performance in the Covid-19 Pandemic,” vol. 20, no. 4, pp. 126–134, 2023.

[14] N. Tamsir, “Algorithm C4.5 in Classifying Health of Cat,” J. Inf. Technol. Its Util., vol. 4, no. 2, pp. 56–63, 2021, doi: 10.56873/jitu.4.2.4410.