The Effect of the SMOTE Method on the Classification of Toddler Nutritional Status Using the Naïve Bayes Method

Dewi Sartika
Informatics Management Department, Faculty of Computer Science
Sriwijaya University
Palembang, Indonesia
dewisartika@unsri.ac.id

***Abstract*— The first five years of life are a golden age for growth and development, so fulfilling nutritional intake during this period is very important to avoid stunting or growth failure. The problem of stunting is still the focus of the government because it is related to nutrition which is one of the key aspects for the development of qualified resources as well as in national development. According to the report of the Ministry of Health in 2023, it was stated that the results of the 2023 Indonesian Health Survey showed that there had been a decreasing in the prevalence of stunting over the past 10 years but it had not been able to meet the target of the 2020-2024 National Medium-Term Development Plan of 14% in 2024. This study will classify the toddler's nutritional status using the Naive Bayes method. This method uses a probability technique with Bayes' theorem which is based on the assumption of mutually independent and equal conditions. The calculation of the Naive Bayes probability in this study uses the Multinomial distribution because the data used is discrete data. The total numbers of toddlers' nutritional status data obtained was 245 data, with 4 invalid data. Based on the data set owned, the number of samples for each class label had an unbalanced number. One method could be used to handle this unbalanced data is the random oversampling method, Synthetic Minority Oversampling (SMOTE). SMOTE will create synthetic data randomly to balance minority data samples. The analysis and testing results showed that in Multinomial Naive Bayes with the 10-cross validation technique, the g-means value obtained on the original data set was 44.98% while in the balanced data set the g-means value was 80.06%. In Multinomial Naive Bayes with the split validation technique, the g-means value obtained on the original data set was 44.20% while in the balanced data set was 80.06%. This showed that there was an increase in the g-means value of 35%. It can be stated that the SMOTE method effectively improves the overall capability of the Multinomial Naive Bayes model.**

Keywords*— *Multinomial Naive Bayes; SMOTE; Stunting.

I. INTRODUCTION

Stunting is a problem of growth failure due to inadequate nutritional intake [1]. Stunting is common in infants under five years old or in toddlers. Infants with stunting status are characterized by weight and height that are not in accordance with their age. Stunting condition can affect intelligence levels and motor development [2]. In fact, the first five years of life are the golden age for growth and development. Therefore, stunting has a long-term impact, namely causing a decline in the quality of human resources and degenerative disease problems [3]. Based on this, reducing stunting rates becomes the main focus of the Indonesian government nowadays, because although stunting rates have decreased every year, they are still far from the expected target [4], [5]. The problem of stunting is still a focus of the government because it is related to nutrition which is one of the key aspects for developing quality resources as well as in national development.

In this study, the nutritional status of toddlers will be classified using the Naive Bayes method. This method is used because the calculation is simple [6], it can work quickly [7], it has a high level of accuracy [8], and it can work even with a small data set [9]. The data used in this study is the 2022 toddlers' nutritional data obtained from the XYZ Health Center in Palembang. In the data, there is one class label that is more dominant than the other class labels, so the data become unbalance.

Data imbalance will result in the minority class label being ignored and the classification results tend to lean towards the majority class label. It can cause bias in data classification, thus it can affect the performance of the classification model [10]. The technique used to balance the data is called resampling. Resampling is a data preparation stage (pre-processing) by adding or reducing samples to the data set. There are three resampling techniques, namely oversampling, under-sampling, and hybrid, namely a combination of oversampling and under-sampling techniques. The oversampling technique is more widely used because it does not remove important information from a data set [11]. In previous studies, a comparison of random oversampling, Synthetic Minority Oversampling (SMOTE) and borderline SMOTE methods was conducted on telecommunications company churn data. The results obtained showed that the SMOTE method produced better balanced data than other methods [12]. Therefore, this study will apply the SMOTE method. This method is hopefully being able to balance toddlers 'nutritional status data so that it can increase the level of accuracy of the classification carried out because

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

the data pattern becomes easier to understand. Next, it will analyse whether the SMOTE method has an influence on the performance of the Naive Bayes method in classifying the nutritional status of toddlers or not.

II. RESEARCH METHODOLOGY

A. System Overview

The stages carried out in this study can be seen in Figure 1. The first step taken is to prepare the data set. The data used in this study is toddlers' nutritional status data. Variables whose values are in the form of categories are changed to numeric, which are then saved as a document in csv format. Furthermore, the data set is balanced with the oversampling technique using the SMOTE method so that the number of data sets increase. The next stage is to create a classification model using the Multinomial Naive Bayes method with cross validation and split validation on the data set before and after oversampling. The model that has been created is then tested and analyzed by calculating accuracy, precision, recall, f-measure and g-means to determine the classification performance. Based on the results obtained, it can be concluded that whether the SMOTE method affects the performance of toddlers' nutritional status classification using the Multinomial Naive Bayes method or not.

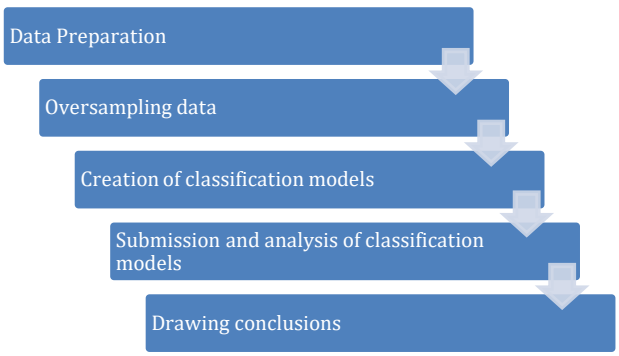

Fig. 1. Research Stages

B. Implication Research

In previous studies, the application of the SMOTE method was carried out on a large dataset. In addition, each study performed classification using one of the validation techniques. Studies [11], [12], [13] used split validation, while studies [10], [14] used k-cross validation. In this study, data balancing will be carried out with a small amount and the analysis of the Multinomial Naive Bayes classification model will be carried out with both data validation techniques, namely split validation and k-cross validation.

C. Dataset Description

In this study, the variables used in determining the nutritional status of toddlers are gender, birth weight, birth height, age, weight, height, and weight gain status obtained during identification of toddlers. There are also variables obtained from the calculation of the comparison of weight and height with age. The distribution and visualization of toddlers' nutritional status data can be seen in Figure 2 and Figure 3. Based on the image, it is clear that the data set that will be used in this study is not balanced. In Figure 3, the class 1 label is for toddlers with malnutrition, class 2 is for good nutrition, class 3 is for the risk of over nutrition, class 4 is for over nutrition, while 5 is for obesity.

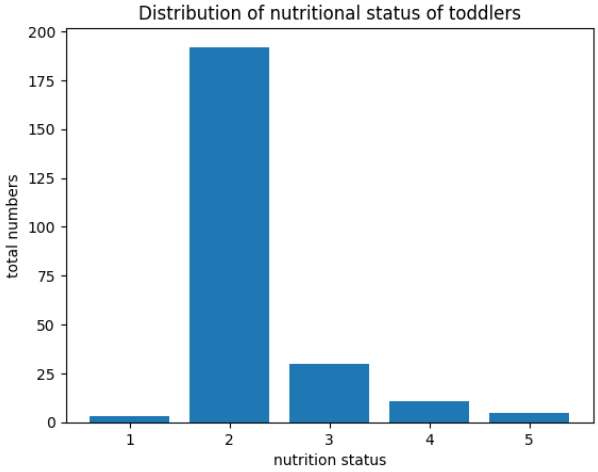

Fig. 2. Data Distribution

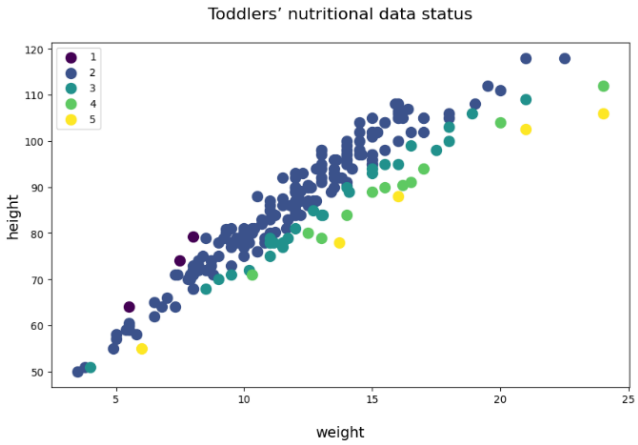

Fig. 3. Data Visualization

D. SMOTE

SMOTE is a data resampling technique where the number of minority data samples will be increased to balance the number of majority data samples. This technique is good for small data sets, so it is expected not to reduce important information in it. How SMOTE works is to find the nearest neighbors as many as k on each minority data to create replica/synthetic data as much as the percentage needed between the minority data and its k nearest neighbors randomly by calculating the difference in their vectors [13]. Determining k is at least n-1, where n is the number of minority data. The calculation of the SMOTE method uses formula (1).

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \quad (1)$$

with x_i is the original point of the minority data, x_j is the nearest neighbouring point of x_i and λ is a random value between 0 and 1 that will determine how far the synthetic point x_{new} is from x_i towards x_j [14].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

E. Naïve Bayes

Naive Bayes is one of the supervised learning methods that can be used to predict labels or classes of data by calculating the probability of each training data label [15]. This method is not only easy, but also fast so it is often used [10]. This method uses probability techniques with Bayes' theorem which is based on the assumption of conditional independence and equality. Independence means that assuming all attributes are free to be targeted and equality means that an event where all attributes are considered equally important [16]. This assumption refers to the idea that the effect of one attribute value on a particular class is unrelated to the value of other attributes. Naive Bayes is implemented with low complexity because it does not require a lot of data for training and does not require model optimization [17], [18]. The calculation of Naive Bayes probability can use the Gaussian distribution formula for continuous data, while discrete data can use the Multinomial distribution. Based on previous research, it was stated that the Multinomial Naive Bayes method was able to create a classification model with a fairly high average performance [19]. The Multinomial distribution is used to determine the probability that is categorized in more than two groups. In general, if an experiment can produce one of k possible outcomes E_1, E_2,..., E_k, with probability p_1, p_2,..., p_k, then the Multinomial distribution will provide the probability that E_1 occurs x_1 times, E_2 occurs x_2 times, ..., and E_k occurs x_k times in n independent experiments, where:

$$x_1 + x_2 + \cdots + x_k = n \quad (2)$$

$$p_1 + p_2 + \cdots + p_k = 1 \quad (3)$$

With probability distribution,

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k; n) \quad (4)$$

Any particular sequence that produces x_1 outcomes for E_1, x_2 for E_2, ..., x_k for E_k will occur with probability $p_1^{x_1}, p_2^{x_2}, \ldots, p_k^{x_k}$. The total number of sequences that produce similar outcomes for n trials is equal to the number of partitions of the n items into k groups with x_1 in the first group, x_2 in the second group, ... and x_k in the k-th group. This can be done in:

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! x_2! \ldots x_k!} \quad (5)$$

Since each trial will produce k outcomes E_1, E_2, …, E_k with probabilities p_1, p_2, …, p_k, then the probability distribution of the random variables $x_1, x_2, \ldots, x_k$ which states the number of occurrences of E_1, E_2, …, E_k in n trials is:

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k; n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \quad (6)$$

with $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} p_i = 1$

F. Data Validation

The selection of model validation methods in the classification process plays an important role in overcoming possible over fitting. Commonly used model validation methods are cross validation and split validation. Cross validation is suitable for limited data sets [16]. There are several methods of cross validation, but the one used in this study is k-fold where the data will be segmented into parts of the same sizes, one part will be used as testing data and the other as training data. Split validation is a model validation method that randomly divides training data and testing data according to a predetermined proportion [20].

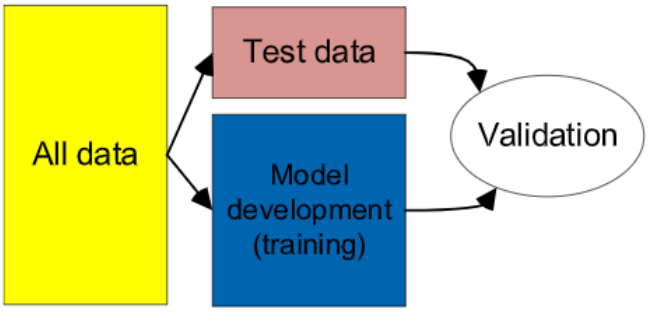

Fig. 4. Illustration of Split Validation [15]

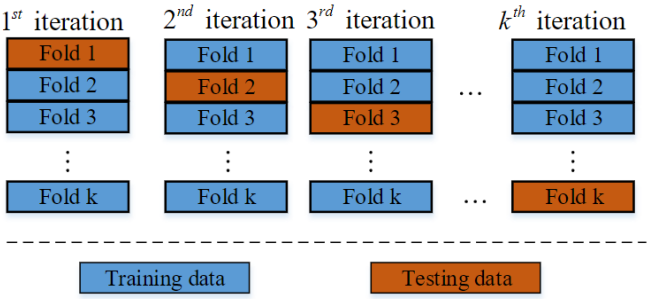

Fig. 5. Illustration of Cross Validation [18]

F. Classification Performance

Classification performance measurement is needed to evaluate the classification model that has been created. Confusion Matrix is a classification performance measurement by calculating the number of correct and incorrect class labels. In this study, data is classified into more than two class labels, so it can be stated as multi-class. The calculation of the Confusion Matrix in multi-class can be stated in formula (7). Other classification performance criteria are calculated based on the results of the Confusion Matrix [21], namely recall (8), precision (9), accuracy (10), f-measure (11), specificity (12) and g-means (13).

$$M(r,c) = \sum_{i=1}^{m} \left(I(y_i = r) I(h(x_i) = c) \right), \forall_r, c \in \{0, \ldots, q-1\} \quad (7)$$

With M, being the value of Confusion Matrix for multi-class, r and c are the rows and columns of M, m is the number of data sets. $I(.)$ is the indicator function, x_i is the i-th input of classifier $h(.)$, y_i is the true label of input x_i and q is the number of classes [22].

$$recall = \frac{TP}{TP+FN} \quad (8)$$

$$precision = \frac{TP}{TP+FP} \quad (9)$$

$$accuracy = \frac{TP+TN}{n} \quad (10)$$

$$fmeasure = \frac{2TP}{2TP+FP+FN} \quad (11)$$

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

$$Specificity = \frac{TN}{TN+FP} \quad (12)$$

$$gmeans = \sqrt{recall - specificity} \quad (13)$$

where *n* is the number of data, True Positive (TP) and True Negative (TN) are the number of class labels that are correctly classified, while False Positive (FP) and False Negative (FN) are the number of class labels that are not correctly classified. Previous research stated that accuracy measurement was less appropriate if it was used as an evaluation of model performance with imbalanced data because the minority class did not affect the level of accuracy [23].

III. RESULT AND DISCUSSION

At data preparation stage, 245 toddler nutritional status data were obtained, with 4 invalid data. After oversampling, there was an increase of data set, from 241 became 430 data. The number of toddler nutritional status data obtained was 245 data, with 4 invalid data. Toddler nutritional status data with variables are gender (male and female), age (in months), birth weight and measured weight (in Kg), birth height and measured height (in Cm), age weight (very less, less, normal, and higher risk), age height (very short, short, normal, and tall), weight gain status (no increase, no previous data and increase) and class labels in the form of toddler nutritional status (undernutrition, good nutrition, risk of overnutrition, overnutrition, and obesity). Samples of raw data obtained can be seen in Table I while samples of processed data sets can be seen in Table II.

TABLE I. RAW DATA SAMPLE

Variable	Score
Sex	Male
Birth Weight	2,7
Birth Height	48
Age	4 years - 6 months - 6 days
Weight	15
Height	101
Age Weight	Normal
Age Height	Normal
Gain Weight	O
Nutritional Status	Good Nutrition

TABLE II. DATA SET SAMPLE

Variabel	Nilai
Sex	1
Birth Weight	2,7
Birth Height	48
Age	54
Weight	15
Height	101
Age Weight	3
Age Height	3
Gain Weight	2
Nutritional Status	2

In the next stage, namely data oversampling, creating, testing, and analysing classification model was carried out by using Python language which was run on the Google Collaborator platform. The libraries used were io, pandas, scikit-learn (sklearn), imbalanced-learn (iblearn), mathplotlib, and seaborn. The io library was used for input and output processes, pandas was used for data manipulation and analysis, sklearn was used for classification, imblearn was used to handle data imbalance, and mathplotlib and sklearn were for data visualization.

The classification model creation was carried out using two validation methods. The first was using k-fold cross validation with k of 10 and the second was using split validation with a test size of 0.33. Both were carried out on the original data set and the data set after oversampling using SMOTE. The k value used in the SMOTE method in this study was 2 because the number of minority class data in the data set, namely malnutrition, was 3. Figure 6 is the Confusion Matrix of the Multinomial Naive Bayes classification model with 10-fold cross validation on the original data set, while Figure 7 is the Confusion Matrix of the Multinomial Naive Bayes classification model with 10-fold cross validation on the data set after being balanced with the SMOTE method. A comparison of the performance of the models that have been calculated based on the two Confusion Matrices can be seen in Table III.

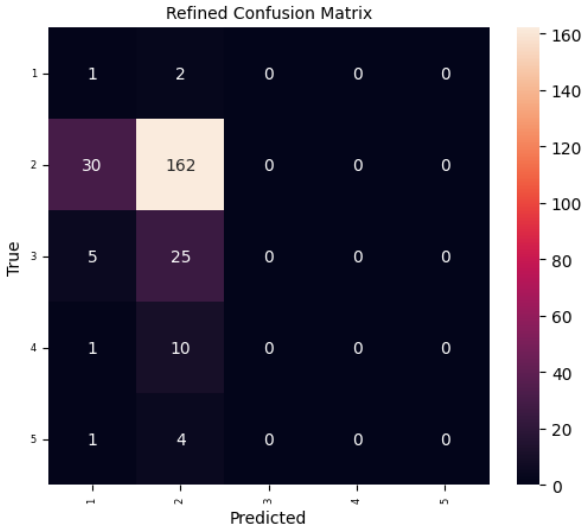

Fig. 6. Confusion Matrix Multinomial Naive Bayes with 10-Cross Validation

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

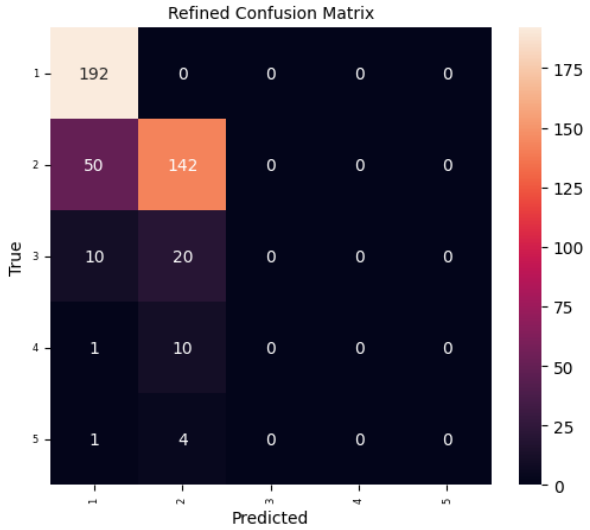

Fig. 7. Confusion Matrix Multinomial Naive Bayes-SMOTE with 10-Cross Validation

TABLE III. COMPARISON OF MODEL PERFORMANCE WITH 10-CROSS VALIDATION

Variable	MNB (%)	MNB-SMOTE (%)
Accuracy	67,65	77,67
Precision	64,00	71,61
Recall	67,65	77,67
f-measure	65,38	73,10
g-means	44,98	80,06

Based on the analysis and testing results in Table 3, overall the SMOTE method affects the performance of Multinomial Naive Bayes with 10-cross validation in classifying the toddler nutritional status data set. This can be specifically explained based on each measurement variable. The performance of the model with the original data set has an accuracy level of 67.65%, while with the balanced data set, the accuracy level increases to 77.67%. This shows that the MNB-SMOTE model is overall better at predicting classes correctly. The model performance with the original data set has a precision value of 64% while with the balanced data set, the precision value increases to 71.61%. This increasing indicates that the MNB-SMOTE model is better at avoiding false positive predictions. The model performance with the original data set has a recall value of 67.65% while with the balanced data set, the recall value increases to 77.67%. This increasing indicates that the MNB-SMOTE model is better at identifying all positive instances. The model performance with the original data set has an *f*-measure value of 65.38% while with the balanced data set, the *f*-measure value increases to 73.10%. This increasing indicates an increase in the balance between precision and recall in the MNB-SMOTE model.

In the last variable, namely *g*-means, there is also an increasing, the g-means value on the original data set is 44.98% while on the balanced data set, the *g*-means value is 80.06%, meaning that the SMOTE method effectively improves the model's ability to classify data based on class labels better, especially minority classes that may be underrepresented in the original data.

Figure 8 is the Confusion Matrix of the Multinomial Naive Bayes classification model with split validation on the original data set, while Figure 9 is the Confusion Matrix of the Multinomial Naive Bayes classification model with split validation on the data set after being balanced with the SMOTE method. A comparison of the model performance calculated based on the two Confusion Matrices can be seen in Table IV.

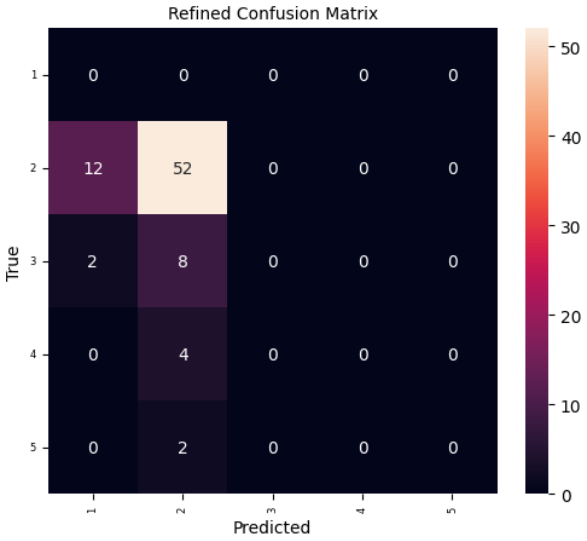

Fig. 8. Confusion Matrix Multinomial Naive Bayes with Split Validation

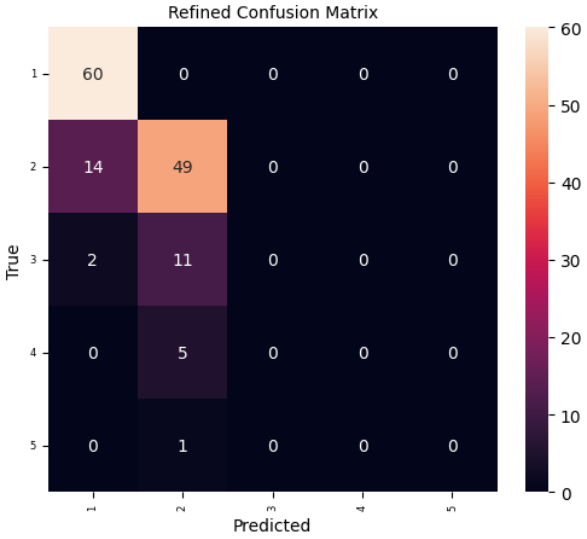

Fig. 9. Confusion Matrix Multinomial Naive Bayes-SMOTE with Split Validation

TABLE IV. COMPARISON OF MODEL PERFORMANCE WITH SPLIT VALIDATION

Variable	MNB (%)	MNB-SMOTE (%)
Accuracy	65,00	76,80
Precision	63,00	66,30
Recall	65,00	76,80
f-measure	64,00	71,00
g-means	44,20	79,40

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

Based on the analysis and testing results in Table IV, it can be stated that the classification model of the Multinomial Naive Bayes method with split validation also overall shows that the SMOTE method affects the performance in classifying the toddler nutritional status data set. This can be explained specifically based on each measurement variable. The accuracy level of the model with the original data set is 65%, while with the balanced data set, the accuracy level increases to 76.80%. This shows that overall the MNB-SMOTE model is better at predicting classes correctly. The precision value of the model with the original data set is 63% while with the balanced data set, the precision value increases to 66.30%. This shows that the MNB-SMOTE model is better at avoiding false positive predictions. The recall value of the model with the original data set is 65% while with the balanced data set, the recall value increases to 76.80%. This shows that the MNB-SMOTE model is better at identifying all positive instances. The *f*-measure value of the model with the original data set is 64% while with the balanced data set, the f-measure value increases to 71%. This shows an improvement in the MNB-SMOTE model in measuring the classification of minority class labels on imbalanced data. The *g*-means value provides an overview of the overall model performance. The *g*-means value of the model with the original data set is 44.20% while on the balanced data set it increases to 80.06%, meaning that the SMOTE method effectively improves the overall capabilities of the Multinomial Naive Bayes model.

IV. CONCLUSION

The conclusion of this study is that data resampling had been carried out by using SMOTE method, with the initial data set was from 241 data to 430 data. In classification model with 10-cross validation technique, an increase in accuracy was 10.02%, in precision was 7.61%, in recall was 10.02%, in f-measure was 7.72%, and in g-means was 35.08%. The classification model with the split validation technique also showed an increase in accuracy of 11.8%, in precision of 3.3%, in recall of 11.8%, in f-measure of 7%, and g-means of 35.2%. Based on these results, it can be stated the SMOTE method affects the performance of the Multinomial Naive Bayes classification model. Although the results of the two validation techniques used in this study did not provide a significant comparison, it can also be stated that the *k*-fold validation technique with $k = 10$ can maximize the use of information from a small dataset.

REFERENCES

[1] J. I. Kesehatan, S. Husada, and K. Rahmadhita, "Permasalahan Stunting dan Pencegahannya," *Juni*, vol. 11, no. 1, pp. 225–229, 2020, doi: 10.35816/jiskh.v10i2.253.

[2] C. R. Titaley, I. Ariawan, D. Hapsari, A. Muasyaroh, and M. J. Dibley, "Determinants of the stunting of children under two years old in Indonesia: A multilevel analysis of the 2013 Indonesia basic health survey," *Nutrients*, vol. 11, no. 5, May 2019, doi: 10.3390/nu11051106.

[3] H. S. Mediani, "Predictors of Stunting Among Children Under Five Year of Age in Indonesia: A Scoping Review," *Glob J Health Sci*, vol. 12, no. 8, p. 83, Jun. 2020, doi: 10.5539/gjhs.v12n8p83.

[4] Kementerian Kesehatan, "Stunting di Indonesia dan Determinannya."

[5] D. Sartika, I. Saluza, and M. H. Irfani, "Perbandingan Akurasi Metode Principal Component Analysis (PCA) dan Correlation-Based Feature Selection (CFS) Pada Klasifikasi Perpanjangan Kontrak Karyawan Menggunakan Metode Naïve Bayes," *Jurnal Informatika Global*, vol. 13, no. 2, pp. 82–87, 2022, doi: 10.36982/jiig.v13i2.2292.

[6] R. Wijaya, Suciati Nanik, and W. N. Khotimah, "Implementasi Nearest Neighbour pada Data Kategorik dengan Pembobotan Atribut Menggunakan Weighted Simple Matching Coefficient," *Jurnal Teknik ITS*, vol. 6, no. 2, pp. A468–A471, 2017.

[7] E. R. Arumi, Sumarno Adi Subrata, and Anisa Rahmawati, "Implementation of Naïve bayes Method for Predictor Prevalence Level for Malnutrition Toddlers in Magelang City," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 201–207, Mar. 2023, doi: 10.29207/resti.v7i2.4438.

[8] E. Faizal, "Case Based Reasoning Diagnosis Penyakit Cardiovascular Dengan Metode Simple Matching Coefficient Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 1, no. 2, pp. 83–90, 2014.

[9] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM & Naive Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," in *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, 2018. [Online]. Available: www.twitter.com

[10] E. Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indonesian Journal of Data and Science (IJODAS)*, vol. 3, no. 2, pp. 47–54, 2022.

[11] G. Gumelar and H. Al Fatta, "Kombinasi Algoritma Klasifikasi Dengan Algoritma Oversampling Untuk Menangani Ketidakseimbangan Kelas Pada Level Data," vol. 10, no. 2, pp. 29–39, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[12] Z. Abdurahman Baizal, M. Arif Bijaksana, and A. S. Sastrawan, "Analisis Pengaruh Metode Over Sampling Dalam Churn Prediction Untuk Perusahaan Telekomunikasi," *Seminar Nasional Aplikasi Teknologi Informasi*, pp. 61–66, 2009.

[13] A. Surya Firmansyah, A. Aziz, and M. Ahsan, "Optimasi K-Nearest Neighbor Menggunakan Algoritma SMOTE Untuk Mengatasi Imbalance Class Pada Klasifikasi Analisis Sentimen," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 6, pp. 3341–3347, 2023.

[14] A. Prawita Ningrum, S. Winarno, and V. Praskatama, "Klasifikasi Kualitas Biji Kedelai Menggunakan Transfer Learning Convolutional Neural Network dan SMOTE," *Journal of Applied Computer Science and Technology*, vol. 5, no. 2, pp. 155–164, Dec. 2024, doi: 10.52158/jacost.v5i2.1002.

[15] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS One*, vol. 14, no. 11, Nov. 2019, doi: 10.1371/journal.pone.0224365.

[16] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthc Inform Res*, vol. 27, no. 3, pp. 189–199, Jul. 2021, doi: 10.4258/HIR.2021.27.3.189.

[17] T. T. Wong and H. C. Tsai, "Multinomial Naive Bayesian classifier with generalized Dirichlet priors for high-dimensional imbalanced data," *Knowl Based Syst*, vol. 228, pp. 1–8, Sep. 2021, doi: 10.1016/j.knosys.2021.107288.

[18] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/ijitcs.2021.06.05.

[19] C. Muhamad Sidik Ramdani, A. N. Rachman, and R. Setiawan, "Comparison of the Multinomial Naive Bayes Algorithm and Decision Tree with the Application of AdaBoost in Sentiment Analysis Reviews PeduliLindungi Application," *International Journal of Information System & Technology Akreditasi*, vol. 6, no. 158, pp. 419–430, 2022.

[20] F. Ramadhani, A. Satria, and I. P. Sari, "Implementasi Metode Fuzzy K-Nearest Neighbor dalam Klasifikasi Penyakit Demam Berdarah," *Hello World Jurnal Ilmu Komputer*, vol. 2, no. 2, pp. 58–62, May 2023, doi: 10.56211/helloworld.v2i2.253.

[21] W. Irmayani, "Visualisasi Data pada Data Mining Menggunakan Metode Klasifikasi Naive Bayes," *Jurnal Khatulistiwa Informatika*, vol. 9, no. 1, pp. 68–72, 2021, [Online]. Available: www.bsi.ac.id

[22] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[23] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan k-Nearest Neighbor," *Jurnal ISD*, vol. 3, no. 1, pp. 44–49, 2018.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i3.2381, Copyright ©2025
Submitted : Mey 19, 2025, Revised : June 13, 2025, Accepted : June 16, 2025, Published : July 28, 2025