BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1585–1594
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1585-1594

A COMPARISON OF ARTIFICIAL NEURAL NETWORK AND NAIVE BAYES CLASSIFICATION USING UNBALANCED DATA HANDLING

Nila Lestari [1*], Indahwati[2], Erfiani[3], Elisa D. Julianti[4]

[1,2,3]Departement of Statistics IPB, FMIPA, IPB University
Dramaga Street., Bogor City, 16680, Indonesia

[4]National Research and Innovation Agency
Jakarta-Bogor Street., Bogor City, 16911, Indonesia

*Corresponding author's e-mail: * **nilalestari@apps.ipb.ac.id***

ABSTRACT

Article History:

Received: 31st March 2023
Revised: 7th August 2023
Accepted: 14th August 2023

Keywords:

Anaemia;
Artificial Neural Network;
Naïve Bayes;
SMOTE;
Unbalanced data

Classification is a supervised learning method that predicts the class of objects whose labels are unknown. Classification in machine learning will produce good performance if it has a balanced data class on the response variable. Therefore, unbalanced classification is a problem that must be taken seriously. This study will handle unbalanced data using the Synthetic Minority Over-Sampling Technique (SMOTE). The classification methods that are quite popular are the Naïve Bayes Classifier (NB) and the Resilient Backpropagation Artificial Neural Network (Rprop-ANN). The data used comes from the Health Nutrition Research and Development Agency (Balitbangkes) which consists of 2499 observations. This study examines the use of NB and ANN using the SMOTE method to classify the incidence of anemia in young women in Indonesia. Modeling is done on 80% of training data and predictions on 20% of test data. The analysis shows that SMOTE can perform better than not handling unbalanced data. Based on the results of the study, the best method for predicting the incidence of anemia is the Naïve Bayes method, with the sensitivity value of 82%.

How to cite this article:

N. Lestari, Indahwati, Erfiani, and E. D. Juliani, "A COMPARISON OF ARTIFICIAL NEURAL NETWORK AND NAÏVE BAYES CLASSIFICATION USING UNBALANCED DATA HANDLING", *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1585-1594, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

The supervised learning algorithm is a model that can predict the response variable Yi for each predictor variable Xi (i = 1, 2, …, n). Classification is one of the supervised learning methods that predict the class of objects whose labels are unknown [1]. The classification method divides the response variables into several classes, which are qualitative or categorical variables [2]. The response variable, which consists of two classes, is called binary classification; if it consists of more than two classes, it is called multiclass classification. Several classifier methods include the naive Bayes method, k-NN, classification tree, ANN, and SVM.

The Naïve Bayes (NB) method is a primary classification method that predicts future opportunities based on past data. The NB method has many advantages: it can be used for quantitative and qualitative data. Another advantage is that it does not require a large amount of data and missing data can be ignored in the calculation. Calculations on this method are fast and efficient. The NB method can be used for binary or multiclass classification, and the algorithm is simple and easy to understand [3].

In Handayani and Pribadi's research [4], it resulted in a high average accuracy in classifying 110 call center services using the Naïve Bayes Classifier. The classification of Naive Bayes in predicting red blood cell thresholds by Patgiri and Ganguly [5] resulted in an accuracy of up to 98.87%. The NB method has a weakness in predictions because of the assumption of independence. To handle this, this method can do weight optimization.

Another primary classification method is Artificial Neural Network (ANN). ANN is the leading computational technique for predicting accurate results for linear, non-linear, and even complex datasets [6]. The sensitivity of the ANN analysis is calculated based on the normalized weights associated with each variable in the model. The advantages of ANN are that it can model non-linear function relationships, can use all the features collected to get a solution, and has a high accuracy value. Despite the advantages of this method, ANN has experienced many updated ideas for solutions to existing deficiencies ranging from feed-forward, backpropagation, and resilient backpropagation ANN (Rprop-ANN) [7].

Research by Kuvvetli [8] predicts the number of daily cases and deaths caused by COVID-19 in the future to generalize the spread in different countries. This research aims to design a predictive model based on the Artificial Neural Network (ANN) model. The results of this test show that the ANN model is quite good, with an accuracy value of 86%.

The backpropagation operation adjusts the weight of misclassified cases based on the error rate. This method is based on an adaptive process that iteratively retrains a model to improve its fitness and predictive power. The backpropagation algorithm has a weakness in the learning rate [9]. If you use a puny learning rate, the computation will operate longer. When using a large learning rate, the weight value is farther from the minimum weight. This underlies the development of a new algorithm, namely resilient backpropagation [10].

The problem that often arises when doing classification is the state of unbalanced class data. Data class imbalance is a condition when the proportion of data is skewed to one class. This condition will greatly affect the formed model, especially when making predictions [11]. The problem of unbalanced data will be handled using the Synthetic Minority Over-sampling Technique (SMOTE) method. This method will carry out the data modification process by making synthetic data for the minority class so that it has a balanced proportion.

One of the world's main nutritional problems today, especially in developing countries, is a deficiency of micronutrients, particularly iron. Iron deficiency affects approximately 66% to 80% of the world's population, with more than 30% suffering from anemia [12]. The 2013 Riskesdas data showed that the prevalence of anemia in women aged 15 years and over was 22.7% [13]. Young women (Rematri) are prone to anemia because during menstruation, they lose a lot of blood [14]. If anemia in young women is not treated immediately, it will impact their intelligence and understanding. In addition, the impact is greater if the female suffers from anemia, namely when pregnant. Women who are pregnant and infected with anemia can be at risk for fetal defects and premature birth of children [15].

Previous research that examined anemia using machine learning classification included [16]. The aim is to determine which individual classifier achieves maximum accuracy in the classification of red blood cells for anemia detection. There are four classification methods used, namely Decision tree, Artificial Neural Network, Naïve Bayes, and K-Nearest Neighbor. The results show that the stacked ensemble method achieves the highest accuracy among the ansamble methods. Next is research conducted by [17] which examines the classification of types of anemia. Researchers used four methods, namely ANN, SVM, NB, and Ensemble Decision Tree. The highest accuracy was achieved using Bagged Decision Trees, followed by Boosted Trees and Artificial Neural Networks.

This study will model the incidence of anemia in young women using the NB and ANN classification methods with unbalanced data handling. Based on previous research, the advantages of this study in this study will use a basic and simple classification method so that other parties can easily analyze anemia status with precise accuracy. In addition, this study uses mixed predictor variables between categorical variables and numerical variables.

2. RESEARCH METHODS

This study will predict the response variable (Y), which is the incidence of anemia (yes or no). The occurrence of anemia is called the positive class, symbolized by 1, while non-anemia is called the negative class, symbolized by 0. The prediction was made using the Naive Bayes (NB) method and Artificial Neural Network (ANN). Before carrying out the modeling stages of the two methods, check and handle unbalanced data using the Synthetic Minority Over-sampling Technique (SMOTE) first.

The stages of research on empirical data are (1) Variable selection and data cleaning if needed. (2) Perform descriptive data analysis to see an overview of the data. (3) Split the data into two parts: 80% training data and 20% testing data. (4) Handle data imbalance using SMOTE. (5) Create a machine learning classification model using training data with default parameters on the NB and ANN models. (6) Make predictions using data testing. (7) Evaluate the best model. They tested the goodness of the model using data testing by considering the accuracy of its classification through the values of sensitivity, accuracy, and AUC-ROC. (8) Repeating steps (3) to (7) 100 times then calculating a triad of statistics for each evaluation metric used.

2.1 Data

The data used in this study were public health data for 2013 and biomedical data for 2017. This secondary data was obtained from the Health Nutrition Research and Development Agency (Balitbangkes) compiled in the 2013 Basic Health Research [13]. The Research and Development Center of the Ministry of Health of the Republic of Indonesia has conducted an analysis of biomedical data in the form of stored biological material, including measurements of ferritin, hemoglobin, CRP, and sTfR. The observations will be examined from a group of young women aged 12-25 years as many as 2499 spread across 497 districts/cities and 33 provinces in Indonesia. The following variables will be used:

Table 1. The Variables Used

Variable		Scale	Description
Y	HB	Nominal	0 = No, 1 = Yes
X1	Diarrhea	Nominal	0 = No, 1 = Yes
X2	Malaria	Nominal	0 = No, 1 = Yes
X3	Hepatitis	Nominal	0 = No, 1 = Yes
X4	Cancer	Nominal	0 = No, 1 = Yes
X5	Fruit Consumption	Ordinal	0 = Less, 1 = Enough
X6	Vegetable Consumption	Ordinal	0 = Less, 1 = Enough
X7	Nutritional Status	Ordinal	1 = Undernourished 2 = Normal 3 = Overnutrition
X8	Pregnant Status	Nominal	0 = No, 1 = Yes
X9	Ferritin Serum	Numerical	
X10	CRP	Numerical	
X11	STfR	Numerical	

2.2 Naive Bayes (NB) Model

The Naïve Bayes Classifier model used for empirical data is as follows [18]:

$$P(\gamma_j|\mathbf{x})P(\gamma_j) \ = P(x_1|\gamma_j)\cdot P(x_2|\gamma_j)\cdots P(x_k|\gamma_j)P(\gamma_j)$$

$$= \prod_{k=1}^{n} P(x_k|\gamma_j)P(\gamma_j) \qquad (1)$$

obtained

$$P(\gamma_j|\mathbf{x}) = \frac{\prod_{k=1}^{n} P(x_k|\gamma_j)P(\gamma_j)}{P(\mathbf{x})} \qquad (2)$$

$$y = argmax\ P(\gamma_j) \prod_{k=1}^{n} P(x_k|\gamma_j) \qquad (3)$$

Where $P(\gamma_j|\mathbf{x})$ is the posterior probability, x is the observed vector **x** = {x_1, x_2,…, x_n} with n attributes each describing a value *A_1*, *A_2*,…, *A_n*. At the same time, γ_j is the response variable in the form of the *j*th class (*j* = 1,2).

2.3 Artificial Neural Network (ANN) Model

The Artificial Neural Network model used in this study for empirical data is as follows:

$$Y_j = g_j\left(\sum_{k=1}^{K} f_k\left(\sum_{i=1}^{n} w_{ik}^{h} * X_i + B_{1k}\right) + B_{2j}\right)$$

Or (4)

$$Y_j = g_j\left(B_{2j} + \sum_{k=1}^{K} w_{jk}^{0} * h_j\right)$$

So obtained

$$Y_j = g_j(w_1^0 h_1 + w_2^0 h_2 + \cdots + w_2^0 h_2 + B_{2j}) \qquad (5)$$

Where g_j dan f_k each is the transfer function of the *j*th output and the *k*th hidden layer. X_i is the input variable and Y_j the output variable of each case. w_{ik}^{h} is the weight from the input (predictor) neuron *i* to the hidden neuron *k*. B_{1k} and B_{2j} are the bias values of the k hidden neurons and the output neurons *j*, respectively **[8]**. Furthermore w_{jk}^{0} is the weight from the hidden neuron to the output neuron and *h* is the output from the hidden layer.

2.4 SMOTE

Unbalanced data is a condition where the distribution of data classes has different proportions. There are more response data classes (majority class) than other data classes **[11]**. One method to overcome the problem of unbalanced response data is the Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE is a technique for balancing data classes by adding synthetic data to minority classes. Minor classes are generated based on the k-nearest neighbor. Calculation of the distance to make synthesis data using the distance formula according to the type of data in the independent variables **[19]**.

2.5 Evaluation of the Classification Model

The performance level of binary classification and classification with more than two classes is evaluated based on the relationship between the predicted class and the actual classes. If it is correct, it is labeled (1) and if it is wrong, it is labeled (0). These calculations are organized into a table consisting of four different combinations known as a confusion matrix. The following is a two-class confusion matrix table:

Table 2. ***Confusion Matrix***

Actual Class	**Prediction Class**	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Accuracy: Comparison of data that is predicted correctly to the overall prediction results.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Sensitivity: Comparison between the correct predicted value of the positive class to the amount of data that has actual positive data.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

TP = True Positive
TN= True Negative
FP = False Positive
FN= False Negative

The AUC (Area Under Curve) value measures the area under the curve, indicating the level of prediction accuracy. AUC is used as a discriminatory performance indicator by estimating the probability of randomly selected observed outputs. The greater the AUC value, the stronger the classification performance **[20]**.

3. RESULTS AND DISCUSSION

3.1 Data Exploration

The proportion of the response variable (Y) is shown in the following pie chart

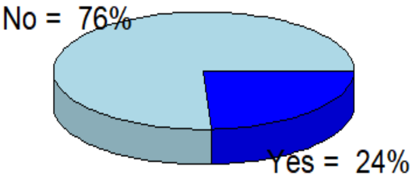


Figure 1. Class Comparison of Response Variables (Anemia Status)

There were 1,849 individuals who did not experience symptoms of anemia, with a value of 76%, and 595 with symptoms of anemia with a value of 24%. This data is considered unbalanced, so it is necessary to handle data imbalance. The observation class with anemia status is the minority class, from now on referred to as the positive class (Y = 1). At the same time, the observation class with non-anemic status is the majority class, hereinafter referred to as the negative class (Y = 0).

Next is to see a comparison of anemia status data based on predictor variables, both categorical and numerical types, as shown in Figure 2.

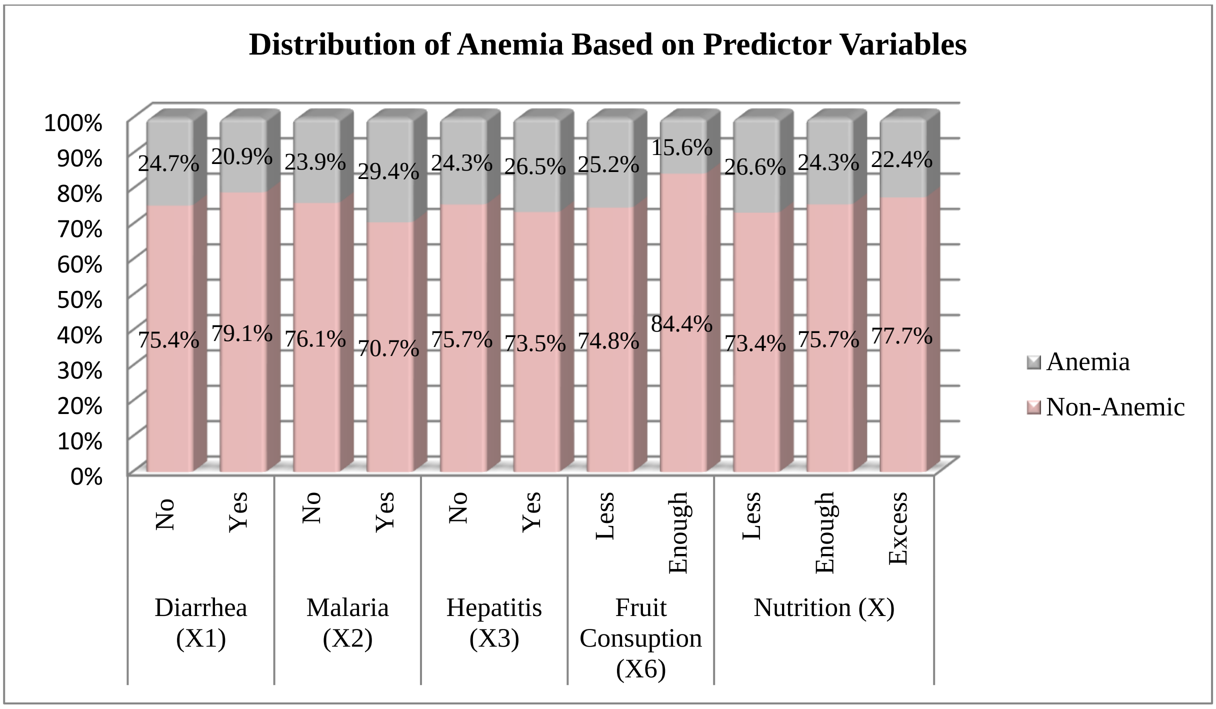
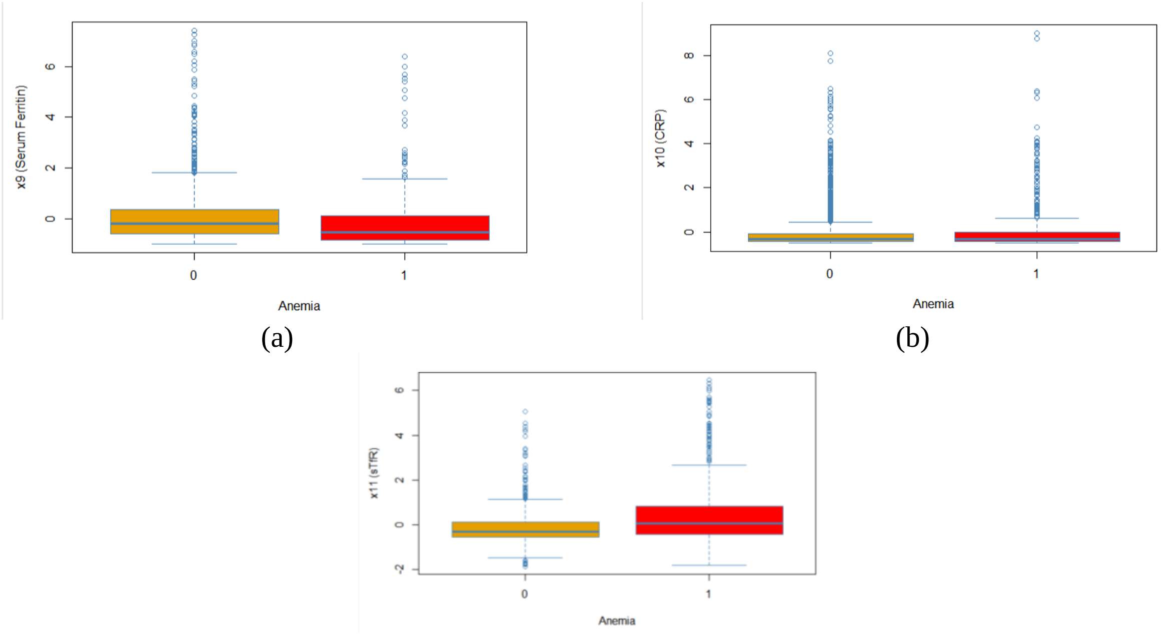


Figure 2. Comparison of Anemia Status Based on Categorical Predictor Variables

It was observed that those who suffered from diarrhea experienced fewer positive anemia than those who did not have diarrhea, namely 20.9%. The difference in observations with anemia for those with diarrhea and without diarrhea was 3.8%. Furthermore, namely malaria and hepatitis, in which sufferers experience positive anemia more than those who do not experience anemia. Positive observations of anemia for those affected by malaria and hepatitis were 29.4% and 26.5%, respectively. The difference in observations with anemia for those suffering from malaria and hepatitis compared to those without the disease was 5.5% and 2.2%, respectively. Individuals who consumed less vegetables experienced more positive anemia than individuals who consumed enough vegetables, namely 25.2%. The difference between individuals who have anemia for those who consume less vegetables and consume enough vegetables is 9.6%. Individuals with less nutritional status also experienced more positive anemia than individuals with good nutritional status and more nutrition, as much as 26.6%. The difference between individuals with anemia for those with less nutritional status compared to those with sufficient nutrition was 2.3%.



(a)

(b)

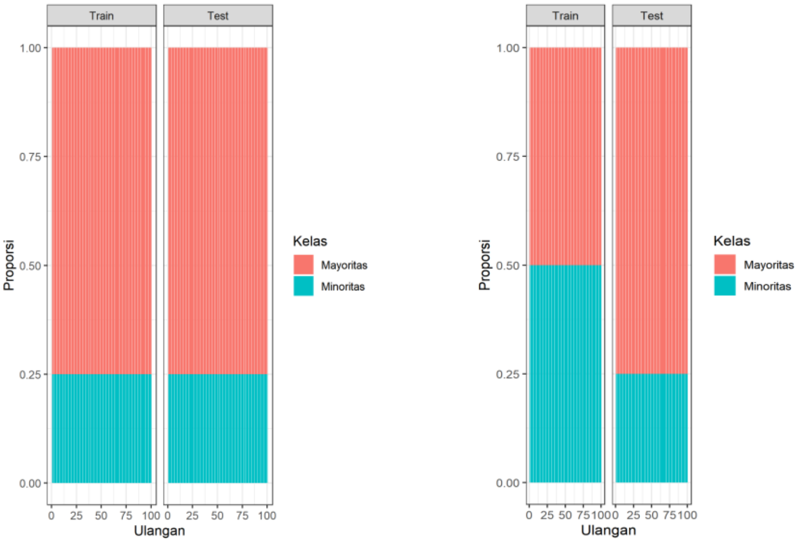
(c)

Figure 3. Comparison of Anemia Status Based on Numerical Predictor Variables
(a) Serum Ferritin, (b) CRP, (c) sTfR

The box plot above shows the relationship between the predictor variables with numeric data types and anemia status. The plot results show the relationship between each predictor variable, namely Serum Ferritin, CRP, and sTfR with the response variable (Anemia). From the data, all variables have a data distribution that tends to be unfavorable. This is because there are many outliers.

3.2 Modeling

Figure 4 shows the distribution of training and test data before and after unbalanced data handling. Before handling unbalanced data, it appears that the proportion of training and test data for the majority class is 75%, and for the minority class is 25%. Furthermore, unbalanced data handling was carried out on the training data using the SMOTE re-pilot method. To produce a balanced proportion of data on the training data that is equal to 50% for the majority and minority classes.

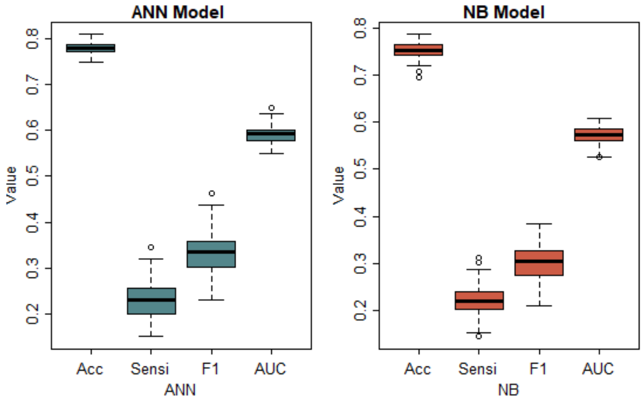


(a) **(b)**

Figure 4. Proportion of training data and test data
(a) The proportion of data sharing before data handling is unbalanced,
(b) The proportion of data sharing after data handling is unbalanced

3.2.1 Prediction Results Without SMOTE

Figure 5 shows a boxplot of the comparison of metric results from predictions of anemia repeated 100 times using test data from the ANN and NB models without unbalanced data handling. The ANN model does not provide variations in accuracy, sensitivity, F1 Score, and AUC performance. The resulting value tends to be monotonous, namely an accuracy of 0.78, the median value for sensitivity and F1 score very low and an AUC is 0.59. Whereas for the NB model, the median value for accuracy is 0.75, the median value for sensitivity and F1 score very low, and the AUC value is 0.57. So it can be concluded that unbalanced data predictions cannot perform well for the two methods used.

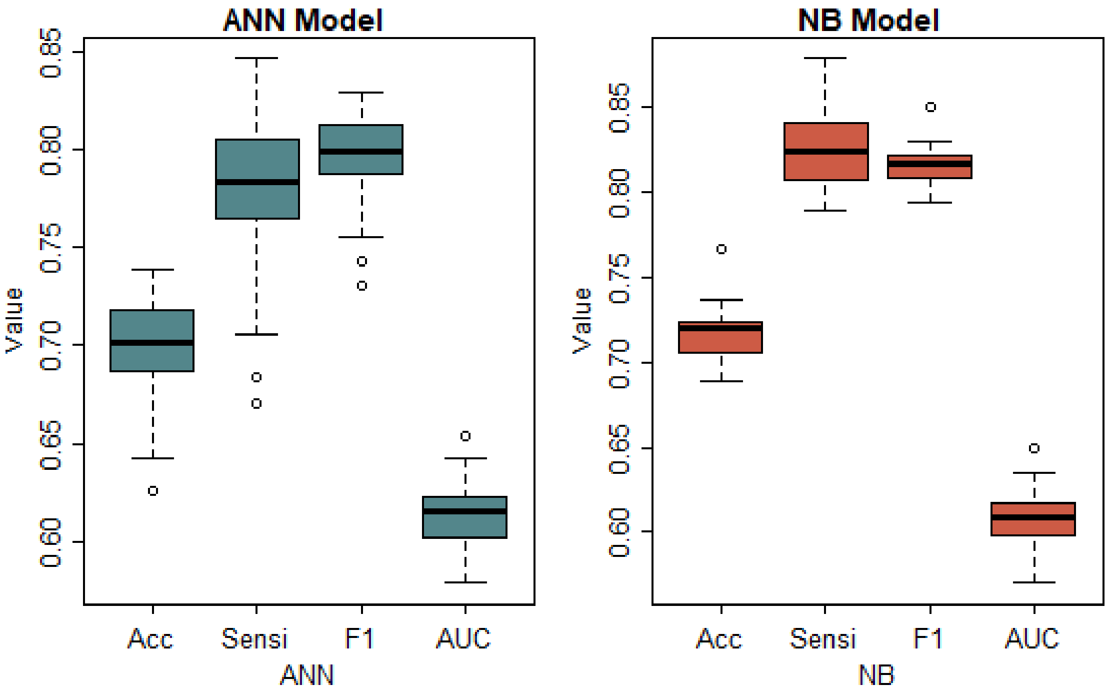


(a) **(b)**

Figure 5. Comparison of Model Evaluation Metrics Without SMOTE
(a) ANN Model Without SMOTE, (b) NB Model Without SMOTE

3.2.2 Prediction results with SMOTE

Figure 6 shows a boxplot of comparing metric results from predictions of anemia repeated 100 times using test data from the ANN and NB models with unbalanced data handling. It can be seen that the ANN model produces a median value for accuracy of 0.70, a median value for the sensitivity of 0.78, the F1 Score is 0,79 and an AUC of 0.62. As for the NB model, the median value for accuracy is 0.72, the median value for sensitivity is 0.82, the F1 Score is 0,82 and the AUC value is 6.1. Based on the four metrics used, the NB Method excels in the three resulting metrics.



(a) (b)

Figure 6. Comparison of Model Evaluation Metrics With SMOTE
(a) ANN Model With SMOTE,
(b) (b) NB Model With SMOTE

Analysis using unbalanced data handling on training data for both prediction models gives better results. When using SMOTE, the resulting accuracy values tend to be lower. Even though the accuracy value is lower, the resulting accuracy is quite good. Furthermore, the resulting sensitivity value and F1 score are much better than the analysis without unbalanced data handling. So the resulting AUC value is also higher. Based on Figure 6 it is found that the NB method has better performance than ANN.

4. CONCLUSIONS

The predictor variables used in the modeling are eight variables consisting of five categorical variables and three numerical variables. Based on the analysis results, it was found that applying SMOTE to training data provided better performance than not using unbalanced data handling. In modeling using SMOTE, ANN has not been able to increase the predictive ability of the minority class, so that the resulting sensitivity value is still meager. Based on the evaluation of the model obtained, the NB method is superior with a median value performance for a sensitivity of 0.82. Therefore, in this study it can be concluded that the best method for predicting the incidence of anemia in young women is the Naïve Bayes method.

ACKNOWLEDGMENT

This research was funded by the Education Fund Management Institute of the Republic of Indonesia (LPDP) and was supported by the National Research and Innovation Agency.

REFERENCES

[1] Bustami, “Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi,” *J. Inform. Ahmad Dahlan*, vol. 8, no. 1, p. 102632, 2014, doi: 10.26555/jifo.v8i1.a2086.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*. 2013.

[3] R. Nisbet, G. Miner, and K. Yale, *Chapter 11 Model Evaluation and Enhancement*. 2018.

[4] F. Handayani and S. Pribadi, “Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110,” *J. Tek. Elektro*, vol. 7, no. 1, pp. 19–24, 2015.

[5] C. Patgiri and A. Ganguly, “Adaptive thresholding technique based classification of red blood cell and sickle cell using Naïve Bayes Classifier and K-nearest neighbor classifier,” *Biomed. Signal Process. Control*, vol. 68, no. July 2020, p. 102745, 2021, doi: 10.1016/j.bspc.2021.102745.

[6] R. Mahadeva, M. Kumar, S. P. Patole, and G. Manik, “Employing Artificial Neural Network for Accurate Modeling, Simulation and Performance Analysis of An RO-Based Desalination Process,” *Sustain. Comput. Informatics Syst.*, p. 100735, 2022, doi: 10.1016/j.suscom.2022.100735.

[7] O. Erkaymaz, “Resilient back-propagation approach in small-world feed-forward neural network topology based on Newman–Watts algorithm,” *Neural Comput. Appl.*, vol. 32, no. 20, pp. 16279–16289, 2020, doi: 10.1007/s00521-020-05161-6.

[8] Y. Kuvvetli, M. Deveci, T. Paksoy, and H. Garg, “A predictive analytics model for COVID-19 pandemic using artificial neural networks,” *Decis. Anal. J.*, vol. 1, no. August, p. 100007, 2021, doi: 10.1016/j.dajour.2021.100007.

[9] M. Almiani, A. Abughazleh, Y. Jararweh, and A. Razaque, “Resilient Back Propagation Neural Network Security Model For Containerized Cloud Computing,” *Simul. Model. Pract. Theory*, vol. 118, no. April, p. 102544, 2022, doi: 10.1016/j.simpat.2022.102544.

[10] S. R. Andani and R. Dewi, “Model Algoritma Resilient Backpropagation Dalam,” vol. 2, no. 2, pp. 67–75, 2019.

[11] F. Sağlam and M. A. Cengiz, “A novel SMOTE-based resampling technique trough noise detection and the boosting procedure,” *Expert Syst. Appl.*, vol. 200, no. April 2020, pp. 1–12, 2022, doi: 10.1016/j.eswa.2022.117023.

[12] K. Meena, D. K. Tayal, V. Gupta, and A. Fatima, “Using classification techniques for statistical analysis of Anemia,” *Artif. Intell. Med.*, vol. 94, no. August 2018, pp. 138–152, 2019, doi: 10.1016/j.artmed.2019.02.005.

[13] Kemenkes RI, “Riset Kesehatan Dasar - Riskesdas 2013,” Jakarta, 2013. doi: 10.1517/13543784.7.5.803.

[14] A.-L. M. Heath, C. M. Skeaff, S. Williams, and R. S. Gibson, “The role of blood loss and diet in the aetiology of mild iron deficiency in premenopausal adult New Zealand women,” *Public Health Nutr.*, vol. 4, no. 2, pp. 197–206, 2001, doi: 10.1079/phn200054.

[15] N. N. Abu-Baker, A. M. Eyadat, and A. M. Khamaiseh, “The Impact of Nutrition Education on Knowledge, Attitude, and Practice Regarding Iron Deficiency Anemia Among Female Adolescent Students in Jordan,” *Heliyon*, vol. 7, no. 2, 2021, doi: 10.1016/j.heliyon.2021.e06348.

[16] P. T. Dalvi and N. Vernekar, “Anemia detection using ensemble learning techniques and statistical models,” *2016 IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2016 - Proc.*, pp. 1747–1751, 2017, doi: 10.1109/RTEICT.2016.7808133.

[17] T. K. Yıldız, N. Yurtay, and B. Öneç, “Classifying anemia types using artificial learning methods,” *Eng. Sci. Technol. an Int. J.*, vol. 24, no. 1, pp. 50–70, 2021, doi: 10.1016/j.jestch.2020.12.003.

[18] J. Lin and J. Yu, “Weighted Naive Bayes classification algorithm based on particle swarm optimization,” *2011 IEEE 3rd Int. Conf. Commun. Softw. Networks, ICCSN 2011*, pp. 444–447, 2011, doi: 10.1109/ICCSN.2011.6014307.

[19] A. Arafa, N. El-fishawy, M. Badawy, and M. Radad, “RN-SMOTE : Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5059–5074, 2022, doi: 10.1016/j.jksuci.2022.06.005.

[20] M. Artur, “Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features,” *Procedia Comput. Sci.*, vol. 190, no. 2019, pp. 564–570, 2021, doi: 10.1016/j.procs.2021.06.066.