BAREKENG: Journal of Mathematics and Its Applications
June 2023 Volume 17 Issue 2 Page 0767–0774
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss2pp0767-0774

PRE-PROCESSING DATA ON MULTICLASS CLASSIFICATION OF ANEMIA AND IRON DEFICIENCY WITH THE XGBOOST METHOD

Fathu Nurrahman[1], Hari Wijayanto[2*], Aji Hamim Wigena[3], Nunung Nurjanah[4]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University*
Meranti Street, Wing 22 lv. 4, Campus IPB, Bogor, 16680, West Jawa, Indonesia

[4]*National Research and Innovation Agency (BRIN)*
Raya Jakarta-Bogor Street, Cibinong Science Center, Pakansari, Bogor, 16915, West Jawa, Indonesia

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

Article History:

Received: 5th December 2022
Revised: 7th April 2023
Accepted: 9th April 2023

Keywords:

Anemia;
Boruta;
MissForest;
Multiclass Classification;
SMOTE;
XGBoost

ABSTRACT

Anemia and iron deficiency are health problems in Indonesia and globally. In Multiclass Classification, data problems often occur, such as missing data, too many variables, and unbalanced data. Then pre-processing data will be carried out using MissForest imputation, Boruta featuring selection, and SMOTE to help improve the performance of the classification model in predicting a particular class. After the data pre-processing process is carried out, classification modeling will be carried out using the XGBoost algorithm. It was found that when pre-processing the data could improve the performance of the model in predicting multiclass classification for cases of anemia and iron deficiency in women in Indonesia by 0.815 for the accuracy value and 0.9693 for the AUC value.

How to cite this article:

F. Nurrahman, H. Wijayanto, A. H. Wigena and N. Nurjanah, "PRE-PROCESSING DATA ON MULTICLASS CLASSIFICATION OF ANEMIA AND IRON DEFICIENCY WITH THE XGBOOST METHOD," *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 0767-0774, June, 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

In today's rapid technological developments, data is important and valuable for all sectors in this world. By using data, information, and knowledge can be obtained. Big data is data with large observations, various forms of data and high speed, and many variables that are becoming a trend in the current era. This is what causes classical analysis to be less able to classify big data properly [1]. According to [2], machine learning is a collection of computational methods that are useful for making and improving predictions of objects or observations into a particular class or group.

This machine-learning technique is commonly used for classification analysis. Classification analysis is divided into two based on the number of classes, namely binary and multi-class classifications. The problems that often occur in carrying out classification analysis are missing data, a large number of independent variables, unbalanced data, and limited solidity algorithms or learning methods. Therefore, the development of methods in classification analysis continues to be carried out to overcome these problems. In dealing with missing data, an imputation technique will be carried out. This imputation technique handles missing values, which will be carried out in the pre-processing step of classification modeling [3].

The MissForest imputation method can work on mixed data simultaneously and has a non-parametric character, so it does not depend on certain data distribution assumptions. According to [4], MissForest outperforms other imputations for mixed data, such as MICE, MissPALasso, and KNN, because it is the only method with consistent and relatively lower imputation errors.

According to [5], including irrelevant variables in the model will cause the model to produce lower accuracy values. Boruta is a wrapping feature selection that can select important variables through the addition of shadow variables with a processing algorithm similar to random forest classification [6]. In several studies, boruta showed better results than those who did not use the boruta model in terms of high accuracy, faster computation, and easier interpretation [7], [8].

Data imbalance also often occurs in the case of health data. This can complicate the classification method when performing generalization functions, namely, how well the model performs on new data or has never been trained in the machine learning process [9]. The synthetic Minority Oversampling Technique (SMOTE) is a data-level approach to handling unbalanced data, which is a modification of the oversampling approach. [10] conducted preliminary research on SMOTE and concluded that the SMOTE approach could improve the accuracy of classifiers for minority classes.

Supervised learning is a classification model based on a classification tree. The advantage of a classification tree is that it does not depend on certain assumptions, such as normally distributed data or no multicollinearity between predictors. According to [11], ensemble trees are used to overcome the instability of a single tree and the height variance of a single tree. There are many classification algorithms in the ensemble tree, one of which is XGBoost (Extreme Gradient Boosting). XGBoost works by applying the concept of the boosting method. This is done by building the model sequentially and combining all the models for prediction so that the new model learns from the mistakes of the previous model.

In a study conducted by [12], when predicting the spinal cord infiltration model in patients with malignant lymphoma, logistic regression and XGBoost were used. XGBoost is a better model with an AUC value of 0.844 compared to logistic regression. In making predictions using machine learning for fetal risk analysis using cardiotocography data, it was found that the XGBoost algorithm provides a high precision value of 96% compared to other algorithms [13].

Anemia is the most common nutritional problem in both developed and developing countries and remains a major human health challenge. Cases of anemia are much more common in women than in men. In 2019 global anemia cases will affect around a third (29.9%) of women of childbearing age, 36.5% of pregnant women, and 29.6% of non-pregnant women. According to the results of the 2018 Basic Health Research (RISKESDAS), the incidence of anemia in pregnant women in Indonesia itself is quite high. Data shows that the prevalence of anemia in pregnant women increased from 2013 (37.1%) to (48.9%) in 2018. Based on the parameters of hemoglobin and ferritin, anemia is classified into 3 groups, namely anemia, iron deficiency, anemia iron (ADI)/iron deficiency anemia (IDA). Cases of anemia in pregnant and non-pregnant women in developing countries are generally suspected to be due to iron deficiency, especially during pregnancy; the need for iron increases significantly [14].

Based on the problems above, the researchers plan to carry out an analysis of the classification of anemia and iron deficiency in women in Indonesia in the age range of 10-45 years using the XGBoost

algorithm to see which level of accuracy is best and to perform some data handling such as MissForest for missing data, boruta in selecting influential variables and SMOTE in handling unbalanced data.

2. RESEARCH METHODS

2.1 Anemia

According to [15], anemia is a condition in which the number of red blood cells (and consequently the oxygen-carrying capacity) is insufficient to meet the physiological needs of the body. These physiological needs vary according to a person's age, gender, geographic location of residence, and different stages of pregnancy for women. Determination of the diagnosis of anemia is done by laboratory examination with hemoglobin/Hb levels in the blood using the method of determining serum ferritin levels. Serum ferritin indicates iron stores in the body.

According to the Minister of Health Regulation Number 37 of 2012 concerning the Implementation of a Public Health Center Laboratory, a person is said to have anemia when the blood hemoglobin level shows a value of less than 12 g/dL. Based on the parameters of hemoglobin and ferritin, the classification of anemia in pregnant and non-pregnant women can be seen in Table 1.

Table 1. Classification of anemia in pregnant and non-pregnant women

Anemia Category	Pregnant Women	Non-Pregnant Women
Normal/Not Anemia	Hb $\geq$ 11 mg/dL; Ferritin $\geq$ 15 ug/L	Hb $\geq$ 12 mg/dL; Ferritin $\geq$ 15 ug/L
Iron Deficiency	Hb $\geq$ 11 mg/dL; Ferritin $<$ 15 ug/L	Hb $\geq$ 12 mg/dL; Ferritin $<$ 15 ug/L
Iron Deficiency Anemia	Hb $<$ 11 mg/dL; Ferritin $<$ 15 ug/L	Hb $<$ 12 mg/dL; Ferritin $<$ 15 ug/L
Anemia	Hb $<$ 11 mg/dL; Ferritin $\geq$ 15 ug/L	Hb $<$ 12 mg/dL; Ferritin $\geq$ 15 ug/L

In general, anemia is caused by inadequate production or quality of red blood cells and blood loss, both acute and chronic. Symptoms that are very often found in general in people with anemia are 5L (Lethargic, Tired, Weak, Tired, Neglect).

2.2 MissForest

According to [4], MissForest imputation is a technique for dealing with missing data with an iterative imputation scheme and also utilizes a random forest algorithm that is built from observed data. MissForest can work on categorical and numerical data, data that has a non-linear relationship between variables, data that has interactions between variables, and high-dimensional data. In addition, MissForest does not need any prior data-related information.

2.3 Boruta

Boruta is an algorithm that is based on random forests but can also be used on other tree algorithms without having to specify parameter values and estimate the values of important features. This method is able to increase the value of accuracy, stability, and runtime and avoid overfitting. The algorithm used by Boruta consists of the following steps [6], [16] :

1. Add data by making copies of all the initial variables by randomizing all variables.
2. Random variables are added to remove correlation with the response
3. Run random forest classification on the new data to get Z-scores
4. Determine the maximum Z-scores from the new data and separate each variable with a better score. Perform a two-tailed equivalence test for each variable with an undetermined level of importance.
5. Discard variables that are significantly lower than the Z-scores and delete them permanently
6. Remove all shadow variables and repeat the algorithm until all important scores are obtained for each variable.

2.4 Synthetic Minority Oversampling Technique (SMOTE)

When the number of objects in a data class is more than that of other classes, there will be a data imbalance. The data class with more objects is called the major class, and the other classes are called the minor classes. This data imbalance will cause the model to tend to be misclassified into the major class and ignore the minor class, thereby affecting estimation and accuracy **[10]**. The solution to this problem is to change the class distribution to get a more balanced sample using SMOTE (Synthetic Minority Oversampling Technique). This method handles unbalanced data with the principle of adding the amount of minor class data to be equivalent to the major class by generating artificial data or synthesis using k-nearest neighbors. Generating artificial data that is numerically different from categorical. Numerical data is calculated based on its proximity to the Euclidean distance, while categorical data is simpler, namely the mode value **[17]**.

2.5 Extreme Gradient Boosting (XGBoost)

[18] developed the ensemble technique method from gradient boosting to extreme gradient boosting. This method optimizes a weak set of methods into a more accurate model by increasing performance and speed so that it is 10 times faster than other gradient-boosting methods. The algorithms in XGBoost add prevention of overfitting and speed up the computation process. Overfitting prevention is done by adding a penalty component or optimizing the loss function value. In principle, this XGBoost builds a tree sequentially with the minimum output value in nodes. The equation is as follows.

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (1)$$

where $\left(y_i, \hat{y}_i^{(t-1)}\right)$ is a loss function to measure prediction error and $\Omega(f_t)$ is a regularization parameter that will make the model avoid overfitting.

2.6 Classification Model Evaluation

The confusion matrix is a performance measurement for classification resulting from a method that is expected to classify all data correctly, but it cannot be denied that the performance of a system cannot be 100% accurate and correct **[19]**. To see the confusion matrix can be seen in **Table 2**.

Table 2. Confusion matrix for Classifying the Four Classes

		Prediction Class			
		Normal	**ID**	**IDA**	**Anemia**
Observation Class	**Normal**	*True* Normal	*False* ID	*false* IDA	*False* Anemia
	ID	*False* Normal	*True* ID	*false* IDA	*False* Anemia
	IDA	*False* Normal	*False* ID	*True* IDA	*False* Anemia
	Anemia	*False* Normal	*False* ID	*false* IDA	*True* Anemia

The performance of a classification model is measured by calculating accuracy, sensitivity (recall), and specificity. The AUC (Area Under the Curve) curve is the goodness of the model in differentiating between classes. The AUC value is between 0 to 1. The higher the AUC, the better the model is in predicting anemia class.

2.7 Data

The data used is secondary data obtained from research by the Health Research and Development Agency at RISKESDAS 2013 and a national profile study of the nutritional status of iron (Fe) and Vitamin A (VA) in Indonesia in 2016, with a response variable of anemia status. The data is specifically for pregnant women and non-pregnant women aged 10-45 years. Related the response variables and indicators used in this study are presented in **Table 3**.

Table 3. List of Variable Names

Variable	Variable Name	Variable	Variable Name	Variable	Variable Name
Y	Anemia Status	X_8	Pneumonia	X_{16}	Gestational Age
X_1	Zone	X_9	Malaria	X_{17}	Weight (Kg)
X_2	Region	X_{10}	TB Pulmonary	X_{18}	Height (cm)
X_3	Marital Status	X_{11}	Hepatitis	X_{19}	Nutritional Status
X_4	Age	X_{12}	Cancer	X_{20}	TBI Status
X_5	WUS	X_{13}	Diabetes	X_{21}	CRP Status
X_6	ISPA	X_{14}	History of Pregnancy		
X_7	Diarrhea	X_{15}	Number of Pregnancy		

3. RESULTS AND DISCUSSION

From a total of 11,327 observations of the 21 available variables, it has been identified in **Table 4** that there are missing data on the variables ISPA, Diarrhea, Pneumonia, Malaria, Hepatitis, Diabetes, Weight, Height, Nutritional Status. This missing data can affect the performance of the classification model. This missing data occurs randomly. Lost data itself does not affect the occurrence of lost data. So that we assume this missing data is MCAR; therefore, an imputation process is carried out on empty data using the MissForest Imputation method approach.

Table 4. Number of missing data

Variable Name	Missing Data	Variable Name	Missing Data	Variable Name	Missing Data
Zone	0	Pneumonia	1	Gestational Age	0
Region	0	Malaria	1	Weight (Kg)	0
Marital Status	0	TB Pulmonary	0	Height (cm)	28
Age	0	Hepatitis	6	Nutritional Status	50
WUS	0	Cancer	0	TBI Status	1664
ISPA	2	Diabetes	1505	CRP Status	0
Diarrhea	2	History of Pregnancy	0	Gestational Age	0

At MissForest, each tree is built using samples obtained from the bootstrap process. Each bootstrap sample randomly leaves about one-third of the observations. These observations left for a given tree are referred to as Out of Bag (OOB) **[20]**. OOB observations are not included in the tree-building process. MissForest performance can be measured based on predicted OOB and assumed test data. Imputation performance on numeric data was measured by NRMSE, and categorical data by PFC. **Table 5** shows that the average MissForest imputation value is 0.0756 (NRMSE), 0.0446 (PFC) with an average time of 10 iterations of 194.3 s. This states that the performance of MissForest imputation is good if it is close to 0.

Table 5. MissForest's performance in OOB

Iteration	NRMSE	PFC	Times
1	0.0765	0.0447	128.35 s
2	0.0756	0.0443	115.8 s
3	0.0756	0.0442	173.91 s
4	0.0755	0.0443	166.48 s
5	0.0754	0.0448	114.83 s
6	0.0756	0.0444	297.93 s
7	0.0756	0.0450	230.87 s
8	0.0756	0.0450	136.94 s
9	0.0757	0.0450	298.36 s
10	0.0756	0.0449	279.58 s
Mean	0.0756	0.0446	194.3 s

In a multiclass classification analysis, it is very important to see the distribution of how many target classes are obtained as well as data exploration for each variable to facilitate the classification modeling process as shown in **Figure 1**.

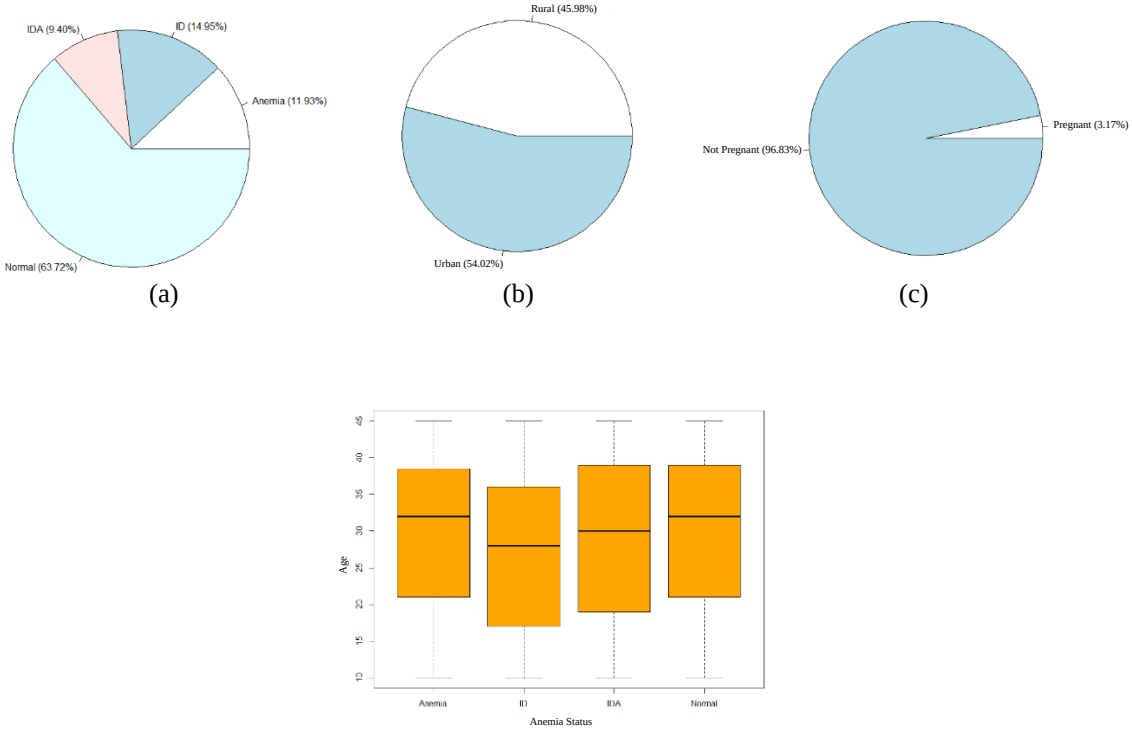

Figure 1. EDA for some data (a) anemia and iron deficiency categories, (b) regions, (c) WUS, (d) age and category boxplots

Based on the results of the descriptive statistical analysis in **Figure 1**, it can be seen that of the 11327 respondents observed in the study, 9.40% were in the IDA category, 14.95% were in the ID category, 11.93% were in the Anemia category, and 63.72% were in the Normal category. Then the results on area variables show that of the 11327 respondents observed in the study, 45.98% came from rural areas, and 54.02% came from urban areas. And based on the WUS variable, it is known that 96.87% are non-pregnant women, and 3.17% are pregnant women. Furthermore, the results of the boxplot graph for each category for age show that the average age of people affected by Anemia tends to be higher than those affected by ID and IDA. In contrast, the average age of normal people is almost the same as the average age of people affected by Anemia.

In classification modeling, including irrelevant variables in the model will cause the model to produce a lower accuracy value. Then a featuring selection is performed using BORUTA. Where the selected variables are variables that have the potential to affect anemia and iron deficiency so as to increase model accuracy and streamline modeling time. Of the 21 existing variables, when BORUTA was carried out, it was found that there were 11 potential variables, namely Status, Age, WUS, History of Pregnancy, Number of Pregnancy, Gestational Age, Weight, Height, Nutritional Status, TBI Status and CRP Status. These 11 variables will later be used in the classification modeling stage.

Before carrying out classification modeling, we know that for the target class or response variables, there is an imbalance in the data. This will trigger an error in the classification. Because the classification will tend to classify to the majority class. Then unbalanced data handling will be carried out using the SMOTE method, where balancing this data will change the amount of data that is almost balanced. We know that at the beginning, the data for each category were IDA (1065), Anemia (1351), ID (1693), and Normal (7218). After SMOTE was carried out, all categories had a more or less proportional number of objects, namely the IDA category of 6177, then the Anemia category of 6448, the ID category of 8964, and the normal category of 7544.

In the classification modeling stage using the XGBoost algorithm, we need to determine the best hyperparameter (hyperparameter tuning) for each model. The XGBoost model requires the max depth,

nrounds, eta, gamma, colsample_bytree, min_child_weight, and subsample hyperparameters. The search process uses Grid SearchCV with k fold validation, where k = 10.

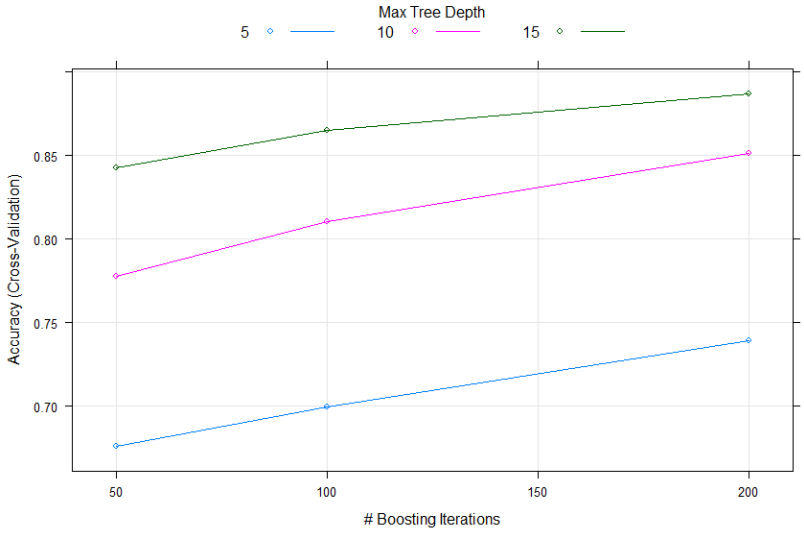

Figure 2. Best hyperparameter determination

In **Figure 2**, the best hyperparameter value is obtained. This comparison is also based on the accuracy value of each. It can be seen that as the value of the nrounds hyperparameter increases, the model's performance improves. In addition, the higher the maxdepth value, the better the model is. The best hyperparamter values obtained with the highest accuracy are nrounds=200, max_depth=15, eta=0.05, gamma=0.01, colsample_bytree=0.75, min_child_weight=0, and subsample=0.5. So the XGBoost model with the hyperparameter will be used.

After obtaining the model, then the model evaluation stage will be carried out. Where the evaluation of this model will look at the level of accuracy, AUC value, sensitivity, and specificity of data carried out by all data handlers and without data handling (Boruta and SMOTE). Then the comparison results are obtained in **Table 6**.

Table 6. Goodness of Fit model XGBoost

	Evaluation Model	Anemia	ID	IDA	Normal	Accuracy	AUC
Unhandled Data	**Specificity**	0.0296	0.0000	0.4890	0.9935	0.6829	0.6688
	Sensitivity	0.9933	1.0000	0.9792	0.1933		
Handled Data	**Specificity**	0.8025	0.8312	0.8575	0.9514	0.8615	0.9693
	Sensitivity	0.9636	0.9266	0.9782	0.9432		

From **Table 6**, it can be seen that when multiclass classification data is handled at the pre-processing stage, it will increase the model performance value. Judging from the Sensitivity and Specificity values in each category, they tend to be proportional or not too much different. This indicates that the use of the SMOTE method to address data imbalance cases is appropriate. In addition, the goodness of the model in prediction can also be seen from the high AUC value of 0.9693.

4. CONCLUSIONS

In the multiclass classification case, especially in anemia and iron deficiency classification in women in Indonesia in the age range of 10-45 years, many problems were found at the data preparation stage. Therefore, data handling was carried out to overcome missing data using MissForest, to handle the selection of variables that used Boruta a lot, and to balance data using SMOTE. After that, a classification analysis was carried out using the XGBoost algorithm. So that the specificity and sensitivity values for each category of Anemia, ID, IDA, and Normal are (0.8025, 0.8312, 0.8575, and 0.9514) and (0.9639, 0.9266, 0.9782, 0.9432) where these values tend to be proportional indicating that the use of SMOTE is quite appropriate. The accuracy and AUC values were 0.8615 and 0.9693, respectively, which indicated that the model performance was quite good in predicting cases of anemia and iron deficiency.

This prediction can estimate a person's category in cases of anemia and iron deficiency in Indonesia. It also helps the government to evaluate performance and policies to make certain decisions. However, the scope of this study is limited to predicting the categories of anemia and iron deficiency. This can also be done with more nutritional needs in a person's body as well as other variables in future research. Despite their superior performance, MissForest, Boruta, and SMOTE's handling of imputed data suffer from deficiencies in computational efficiency. For further research, choosing another type of data handling can be used to reduce computation time. It doesn't reduce accuracy significantly, but it also has to be adjusted to the size and complexity of the dataset.

ACKNOWLEDGMENT

The 2013 RISKESDAS research data is data obtained from Laboratorium Manajemen Data Badan Penelitian dan Pengembangan Kesehatan Kementerian Kesehatan RI, where this is a collaborative research between the National Research and Innovation Agency (BRIN) and lecturers IPB University of the statistics department.

REFERENCES

[1] R. Sharda, S. Voß, and S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, vol. 36. New York, 2016. doi: 10.1007/978-1-4899-7641-3.
[2] C. Molnar, “Interpretable Machine Learning A Guide for Making Black Box Models Explainable,” 2021.
[3] S. I. Khan and A. S. M. L. Hoque, “SICE: an improved missing data imputation technique,” *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00313-w.
[4] D. J. Stekhoven and P. Bühlmann, “Missforest-Non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.
[5] G. Heinze, C. Wallisch, and D. Dunkler, “Variable selection – A review and recommendations for the practicing statistician,” *Biometrical Journal*, vol. 60, no. 3. Wiley-VCH Verlag, pp. 431–449, May 01, 2018. doi: 10.1002/bimj.201700067.
[6] M. B. Kursa and W. R. Rudnicki, “Feature Selection with the Boruta Package,” 2010. [Online]. Available: http://www.jstatsoft.org/
[7] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, “A comparison of random forest variable selection methods for classification prediction modeling,” *Expert Systems with Applications*, vol. 134. Elsevier Ltd, pp. 93–101, Nov. 15, 2019. doi: 10.1016/j.eswa.2019.05.028.
[8] Y. Rimal, “BORUTA ALGORITHM IS SIGNIFICANT FOR LARGE FEATURE SELECTION OF STUDENT MARKS DATA OF POKHARA UNIVERSITY NEPAL,” vol. 1, no. 2, 2020, [Online]. Available: www.uijir.com
[9] S. Fotouhi, S. Asadi, and M. W. Kattan, “A comprehensive data level analysis for cancer diagnosis on imbalanced data,” *Journal of Biomedical Informatics*, vol. 90. Academic Press Inc., Feb. 01, 2019. doi: 10.1016/j.jbi.2018.12.003.
[10] N. v Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” 2002.
[11] B. Sartono and U. D. Syafitri, “METODE POHON GABUNGAN: SOLUSI PILIHAN UNTUK MENGATASI KELEMAHAN POHON REGRESI DAN KLASIFIKASI TUNGGAL,” *Forum Statistika dan Komputasi*, vol. 15, no. No 1, pp. 1–7, 2010.
[12] Y. Huang, C. Chen, and Y. Miao, “Prediction Model of Bone Marrow Infiltration in Patients with Malignant Lymphoma Based on Logistic Regression and XGBoost Algorithm,” *Comput Math Methods Med*, vol. 2022, pp. 1–7, Jun. 2022, doi: 10.1155/2022/9620780.
[13] Z. Hoodbhoy, M. Noman, A. Shafique, A. Nasim, D. Chowdhury, and B. Hasan, “Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data,” *Int J Appl Basic Med Res*, vol. 9, no. 4, p. 226, 2019, doi: 10.4103/ijabmr.ijabmr_370_18.
[14] N. R. van den Broek and E. A. Letsky, “Etiology of Anemia in Pregnancy in South Malawi,” 2000. [Online]. Available: https://academic.oup.com/ajcn/article/72/1/247S/4729620
[15] WHO, “Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity,” 2011.
[16] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, “Boruta - A system for feature selection,” *Fundam Inform*, vol. 101, no. 4, pp. 271–285, 2010, doi: 10.3233/FI-2010-288.
[17] R. Azmatul Barro, I. D. Sulvianti, and M. Afendi, “PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) TERHADAP DATA TIDAK SEIMBANG PADA PEMBUATAN MODEL KOMPOSISI JAMU,” 2013.
[18] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[19] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. 2013.
[20] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.