MSI Journal of
AI and Technology

Logistics Discriminant Analysis Using SMOTE for Anemia Classification in Women of Reproductive Age

Muhammad Nur Aidi[1*], Amamlia Nailul Husna[2], Rahma Anisa[3], Elisa Diana Juanti[4]

[1,3] Lectures at School of Data Science, Mathematics, and Informatics, IPB University Indonesia

[2] Graduated Master at School of Data Science, Mathematics, and Informatics, IPB University, Indonesia

[4] National Research and Innovation Agency, Indonesia

*** Correspondence:** Muhammad Nur Aidi

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 25-April-2025

Accepted: 02-May-2025

Published: 08-May-2025

This article is published by **MSI Publishers** in **MSI Journal of AI and Technology**

ISSN: xxxx-xxxx (Online)

Volume: 1, Issue: 1

(April-Jun) (2025)

ABSTRACT: Anemia is a condition when the hemoglobin (Hb) level is less than normal, which is less than 12 g/dL for women of reproductive age. Analysis to detect the dependence of the risk factor of anemia is important to distinguish the status of anemia and non-anemia women using the classification method. The data in this study are numerical and categorical types so that the classification method used is logistic discriminant. The data is imbalanced on the dependent variable, where the number of non-anemia observations is much more than the anemia observations, so that the data imbalance is handled using SMOTE for modeling. The logistic discriminant discriminates the observations based on the dependent variable and obtains a model where the affected dependent variable can be identified from the significant model coefficients. The results showed that the logistic discriminant classification model in this study had a quite good classification with 73.68% accuracy. The variables that affect anemia status in this study are pneumonia, tuberculosis, hepatitis, diabetes mellitus, malaria, gestational age, and age groups.

Keywords: *anemia, logistic discriminant, SMOTE*

INTRODUCTION

Anemia is a health condition characterized by hemoglobin (Hb) levels that fall below the normal threshold—specifically, less than 12 g/dL in women of reproductive age (WHO, 2017). According to the World Health Organization's 2016 report, anemia affects approximately 33% of women in this demographic worldwide. In Indonesia, the 2018 Basic Health Research (Riskesdas) revealed that 27.2% of women of reproductive age suffer from anemia, highlighting it as a major public health concern in the country.

Several factors are known to contribute to the increased risk of anemia. White (2018) identified malaria as a primary cause, while Garrido et al. (2018) noted that respiratory conditions such as pneumonia elevate the likelihood of developing anemia. Similarly, de Mendonça et al. (2021) found that tuberculosis patients frequently experience anemia. Chang et al. (2002) observed a strong association between hepatitis and anemia risk, and Brière et al. (2021) reported that individuals with diabetes mellitus are also vulnerable. As Vexler et al. (2015) emphasized, analyzing the relationship between disease status and related variables is essential for identifying key risk factors, making it important to characterize both anemic and non-anemic women based on relevant predictors.

This research aims to classify women as anemic or non-anemic by utilizing logistic discriminant analysis—a type of supervised learning technique. Discriminant analysis is a multivariate approach used to separate observations into predefined groups and assign new data points to these groups (Johnson & Wichern, 2007). In this context, the classes (anemic and non-anemic) are predetermined, and logistic discriminant models are built using labeled data to help identify group characteristics and predict membership for new cases.

Everitt & Dunn (2001) noted that Fisher’s linear discriminant function is optimal when data are normally distributed with equal covariance. However, since this study involves both numerical and categorical variables, the assumption of multivariate normality is not satisfied. To address this, logistic discriminant analysis is employed. This method uses a logistic function to directly estimate the probability of class membership for each observation. According to Webb & Cospey (2011), logistic discriminant analysis is well-suited for mixed data types, offers clear interpretability, and accommodates various data distributions. For example, Abdolmaleki et al. (2004) successfully applied this technique to predict breast cancer malignancy with a classification accuracy of 93%.

The data set used in this study exhibits class imbalance, with significantly more observations in one group compared to the other. This imbalance can skew classification performance, often favoring the majority class. To counteract this issue, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates new synthetic examples of the minority class using the k-nearest neighbor method, thereby balancing the class distribution (Chawla et al., 2002).

The purpose of this study is to develop a classification model for distinguishing anemic from non-anemic women based on risk factors, using logistic discriminant analysis enhanced by SMOTE. Additionally, this study aims to identify key variables that significantly influence anemia status.

METHODS

Study Design, Location, and Period

This research employed a cross-sectional design utilizing secondary data from the 2013 *Riset Kesehatan Dasar* (Riskesdas). The dataset included information from 33 provinces and 497 districts across Indonesia, focusing on women of reproductive age (15–45 years). Data were collected through surveys conducted by the Health Research and Development Agency (*Litbangkes*) of the Indonesian Ministry of Health (*Kementerian Kesehatan RI*), involving structured interviews and blood sample collection.

Data Collection Procedures

The primary outcome variable in this study was the anemia status of women aged 15–45, categorized into anemic and non-anemic groups. Anemia was defined following Riskesdas criteria—specifically, hemoglobin (Hb) levels below 12 g/dL for non-pregnant women of reproductive age and below 11 g/dL for pregnant women.

The explanatory (independent) variables included in the analysis were pneumonia, tuberculosis, hepatitis, diabetes mellitus, malaria, gestational age (in weeks), and age group classifications. Each health condition (pneumonia, tuberculosis, hepatitis, diabetes mellitus, and malaria) was treated as a binary variable (yes/no) based on medical diagnoses. Pneumonia status referred to diagnoses within the previous month, while tuberculosis status was based on diagnoses within the past year. Hepatitis, diabetes mellitus, and malaria status were derived from current or recent doctor-confirmed diagnoses.

Gestational age was considered a continuous variable representing pregnancy duration in weeks. Age groups were classified using the Ministry of Health guidelines (2009) as follows: early adolescence (12–16 years), late adolescence (17–25 years), early adulthood (26–35 years), and late adulthood (36–45 years). Categorical variables were encoded using dummy variables for statistical modeling, as detailed in Table 1.

Table 1. Coding of Dummy Variables for Categorical Predictors

Variable	Code	Category
Pneumonia	X_1	0 = No pneumonia
		1 = Pneumonia
Tuberculosis	X_2	0 = No tuberculosis
		1 = Tuberculosis
Hepatitis	X_3	0 = No hepatitis
		1 = Hepatitis
Diabetes Mellitus	X_4	0 = No diabetes
		1 = Diabetes
Malaria	X_5	0 = No malaria
		1 = Malaria
Age Group	X_7	1 = Early adolescent (12–16 yrs)
		2 = Late adolescent (17–25 yrs)
		3 = Early adult (26–35 yrs)
		4 = Late adult (36–45 yrs)

Data analysis

Data in this study were analyzed using logistic discriminant analysis with SMOTE for handling imbalances of data. Logistic discriminant can be used to classificate the observations based on mixed categorical and numerical independent variables. For example, a classification will be carried out based on the independent variables X_1,

X_2, …, X_p with the first class $Y = 1$ and the second class $Y = 0$. The conditional probability $Y = 1$ for X denoted by $P(Y = 1|X) = \pi(x)$, the logit of the logistic model is given by the following equation (Hosmer *et al.* 2013):

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where p is the number of independent variables, for categorical independent variables using dummy variables in the model formation. If the categorical independent variables has k categories, then there are $k - 1$ dummy variables that are included in the model. For example, the jth explanatory variable has k_j levels, D_{ju} represents the $k_j - 1$ dummy variable and β_{ju} is dummy variable coefficient with $u = 1, 2, \ldots, k_j - 1$, so the logit model is given by the following equation (Hosmer *et al*. 2013):

$$g(x) = \beta_0 + \beta_1 X_1 + \cdots + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p X_p$$

Parameter testing was conducted to examine the role of independent variables in the model. The significance parameter testing of the model are likelihood ratio test and Wald test. likelihood ratio test aims to test the significance of the independent variables in the overall model (Hosmer *et al.* 2013). The tested hypotheses are:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_P = 0$

H_1: there is at least one $\beta_i \neq 0$, with $i = 1, 2, \ldots, p$

likelihood ratio test statistics:

$$G = -2\, ln\left[\frac{L_0}{L_p}\right] = -2\, ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}_i^{\,y_i}(1-\hat{\pi}_i)^{1-y_i}}\right]$$

where:

L_0 : *likelihood* without the variable

L_p : *likelihood* with the variable

n_1 : $\sum y_i$

n_0 : $\sum(1-y_i)$

n : the sum of n_0 and n_1

conclusions reject H_0 if $G > \chi^2_{(p,\alpha)}$.

The Wald test is for assessing the significance of parameter β_i individually. The hypotheses of the tested are:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$, with $i = 1, 2, \dots, p$

Wald test statistics:

$$Wi_i = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right)^2$$

information:

W_i : Wald test statistic of ith variable

$\hat{\beta}_i$: estimator β_i

$SE(\hat{\beta}_i)$: standard error of estimator β_i

conclusions reject H_0 if $W_i > \chi^2_{(1,\alpha)}$ (Agresti 2007).

The provision of allocation into classes is by comparing the conditional probabilities of $P(Y = 1|X)$ dan $P(Y = 0|X)$. Based on the logistic model, $P(Y = 1|X)$ dan $P(Y = 0|X)$ are formulated as follows:

$$P(Y = 1|X) = \frac{\exp\left(g(x)\right)}{1 + \exp\left(g(x)\right)}\ dan\ P(Y = 0|X) = \frac{1}{1 + \exp\left(g(x)\right)}$$

according to Everitt and Dunn (2001), the allocation rules to assign the observations into class $Y = 1$, If $P(Y = 1|X)$ is greater than $P(Y = 0|X)$. $P(Y = 1|X)$ will be greater than $P(Y = 0|X)$ if $exp\left(g(x)\right) > 1$, so that obtained if $g(x) > 0$ then the observations assign into class $Y = 1$, and if $g(x) < 0$ then the observations assign into class $Y = 0$.

After the observations are classified, then we need to identify the accuracy of the model classification to determine the goodness of the model in predicting the observations correctly. The confusion matrix is a classification table obtained from the predicted results of observations and actual data.

Table 2. Confusion matrix

Actual	Predicted	
	0	1
0	True negative	False positive
1	False negative	True positive

Confusion matrix contains *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), and *False Negative* (FN). In this study, the positive class is anemia and the negative class is non-anemia. Galar *et al.* (2011) recommends a measure of classification exactness with values of accuracy, sensitivity, and specificity. The calculation of the value of accuracy, sensitivity, and specificity is as follows:

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

$$Sensitivity = \frac{TP}{FN + TP}$$

$$Specificity = \frac{TN}{TN + FP}$$

RESULT AND DISCUSSION

Anemia situation of women at reproductive age in Indonesia. The total observations in this study were 9890 women of reproductive age 15-45 years old. Based on the data, there are 2036 women of reproductive age with anemia and 7854 women of reproductive age with non-anemia. Figure 1 shows that 21% women of reproductive age with anemia and 79% non-anemia women. The distribution of anemia status shows that at least 1 in 5 women of reproductive age in Indonesia is anemic. The large difference between percentages of anemia and non-anemia category indicates an imbalanced data.

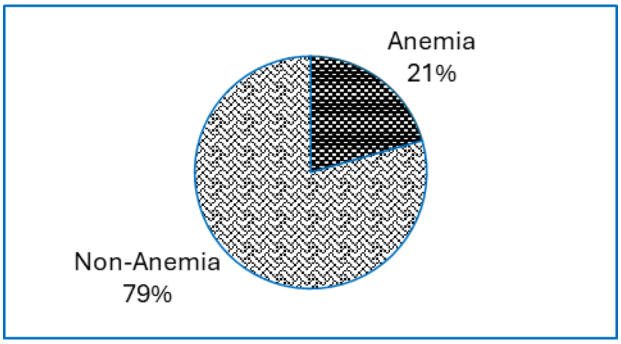

Figure 1. Percentage of anemic and non-anemic women of reproductive age

Anemia risk factor situation. According to the research of Keokenchanh *et al.* (2021) pregnant women are more susceptible to anemia because pregnant women need more blood to transport oxygen to the body's organs and also the fetus. Based on the data, there were 360 observations from the total observations which were pregnant women and about 36,38% of pregnant women were affected by anemia, while the percentage of non-pregnant women who were affected by anemia was 19,98%. In addition to pregnancy, gestational age also affects the risk of anemia. According to Lebso *et al.* (2017), pregnant women in the second and third trimesters or with a gestational age of more than 14 weeks have a high risk of anemia, it means that pregnant women with a gestational age of more than 14 weeks are vulnerable to anemia than pregnant women with a gestational age of less than 14 weeks. Characteristics of numerical variables gestational age can be seen in Table 3.

Table 3. Five number summary of gestational age variable in weeks

Status	Minimum	Q1	Q2	Q3	Maximum	Average
Anemia	4	20	26	31	38	24.83
Non-anemia	2	13	21	30	40	21.69

Table 3 provides information on the Five number summary of gestational age variables in weeks. It can be seen that the mean of gestational age for anemia status is higher than for non-anemia status. This is because more pregnant women with an older gestational age are affected by anemia. The characteristics of anemia women on the categorical variables of pneumonia, tuberculosis, hepatitis, diabetes mellitus, and malaria can be seen in Figure 2.

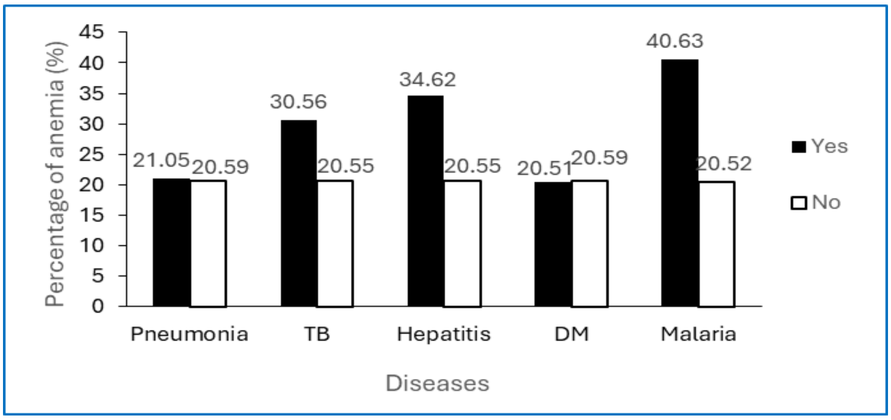

Figure 2. Percentage of women with anemia on the disease variables

Based on Figure 2, it can be seen that women who are affected tuberculosis, hepatitis, and malaria have a higher percentage of anemia than women who do not have these diseases, while women with pneumonia and diabetes mellitus have almost the same percentage of anemia as women who do not have pneumonia and diabetes mellitus. The percentage of women affected anemia by age group variables can be seen in Figure 3.

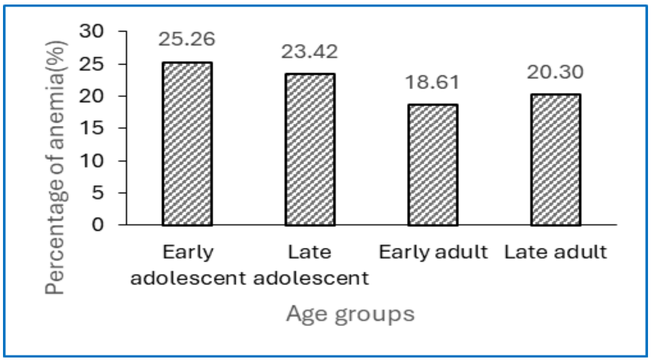

Figure 3 Percentage of women with anemia by age groups

Based on the age group variables in Figure 3, women in the early adolescent age group have a higher percentage of anemia than other age groups, then followed by the age group of late adolescent, late adult, and the lowest percentage of anemia was in the early adult age group.

Application of SMOTE. Based on the data exploration, there is an imbalance in data on anemia status, which non-anemia category is much more than anemia category, in this case non-anemia is major class and anemia is minor class. The data imbalance was handled using the *Synthetic Minority Oversampling Technique* (SMOTE). The application of SMOTE to balance the data using *unbalanced package* in *software* R Studio. There are three parameters needed for this function namely *perc.over,* *perc.under,* and *k-nearest neighbors*. The *perc.over* parameter is an oversampling parameter to determine the amount of synthetic data in the minor class. parameter *perc.over* used in this study is 200%, that means the synthetic data is generated 2 times from the amount of initial minor data, so $2 \times 2036 = 4072$. Then the synthetic data is added to the initial minor data, so that $4072 + 2036 = 6108$ total new minor data is obtained. *perc.under* is a parameter to determine the amount of major class data based on the generated synthetic data. parameter *perc.under* used in this study is 200% or the data in the new major class will be 2 times the total of synthetic data previously, which is $2 \times 4072 = 8144$. total amount of new major class. *K-nearest neighbors* is a parameter that determines the number of k nearest neighbors needed to create synthetic data, which in this study used $k = 5$. That is, the synthetic data comes from 5 data in the minor class which are located close. Determination of these parameters are the result of repeated trials so that the data obtained are quite balanced with the number of anemia observations being 6108 or 43%, while non-anemia observations are 8144 or 57%, and the number of all observations being 14252. The comparison of initial data and data after SMOTE can be seen in Figure 4. It can be seen that the percentage of initial data and data with SMOTE shows a difference with a more balanced proportion in data with SMOTE.

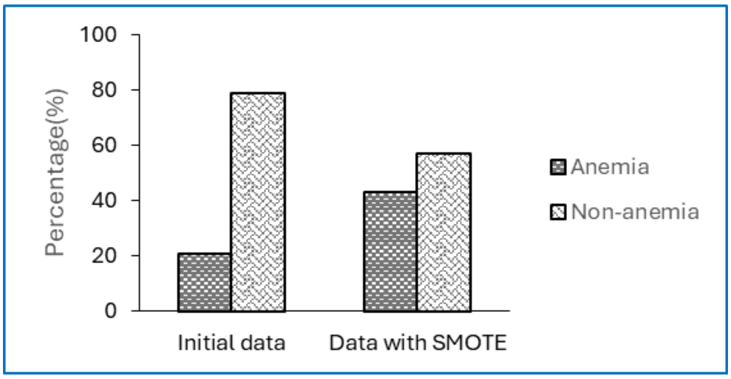

Figure 4. Comparison of initial data and data with SMOTE

Model classification using logistic discriminant. Estimation of the discriminant function is carried out using a logistic model with data after SMOTE. Before estimating the model, the data is divided into two sections, training data and testing data using cross-validation to avoid the *overfitting* or conditions when the model is appropriate with the sample data but not good at predicting data which is not part of the sample data. The data will be divided into five parts randomly with the same amount and each part will become training data and testing data consecutively. The results of the model evaluation on the five parts of testing data will be combined and presented in a confusion matrix.

The estimation of the logistic model was carried out by including all the numerical independent variables and the dummy variables of the categorical independent variables. Anemia status is assigned a value of $Y = 1$ and non-anemia status is assigned a value of $Y = 0$. The estimates logistic model obtained are:

$$\widehat{g(x)} = -0{,}704 + 4{,}047X_{1(1)} + 3{,}120X_{2(1)} + 2{,}790X_{3(1)} + 2{,}076X_{4(1)} + 4{,}416X_{5(1)} + 0{,}029X_6 + 0{,}399X_{7(2)} - 0.410X_{7(3)} - 0.218X_{7(4)}$$

The values in brackets written after the independent variables are dummy variable values for each independent variable. Parameter testing of the logistic model in likelihood ratio test resulted in test statistics 3388.95 and 0.000 p-value. It means that at least one independent variable affects the model at a significant level of 5%. In parameter testing using the Wald test, it shows that all independent variables used are pneumonia, tuberculosis, hepatitis, diabetes mellitus, malaria, gestational age, and age groups have a significant effect on anemia status. The results of Wald test can be seen in Table 4.

Table 4. The result of Wald test

Peubah	Kode	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	W_i	*p-value*
Intercept		-0.704	0.102	47.032	0.000*
Pneumonia					
Pneumonia	$X_{1(1)}$	4.047	0.324	155.950	0.000*
Tuberculosis					
Tuberculosis	$X_{2(1)}$	3.120	0.186	278.756	0.000*
Hepatitis					
Hepatitis	$X_{3(1)}$	2.790	0.253	120.978	0.000*
Diabetes mellitus					
Diabetes mellitus	$X_{4(1)}$	2.076	0.138	226.171	0.000*
Malaria					
Malaria	$X_{5(1)}$	4.416	0.233	356.756	0.000*
Gestational age	X_6	0.029	0.004	46.226	0.000*
Age groups					
Late adolescent	$X_{7(2)}$	0.399	0.111	12.880	0.000*
Early adult	$X_{7(3)}$	-0.410	0.110	13.793	0.000*
Late adult	$X_{7(4)}$	-0.218	0.018	4.044	0.044*

description: *significant at 5% significance level

Evaluation of the model. Model evaluation needs to be done to find out how well the model's ability to classify observations by looking at the confusion matrix. Based on the estimated logistic model obtained, there were 7763 observations classified into the anemia class and 24813 observations into non-anemia class. The results of the classification of these observations can be seen in Table 5.

Tabel 5. Confusion matrix

Actual	Predicted		Total
	Non-anemia	Anemia	
Non-anemia	7974	170	8144
Anemia	3581	2527	6108
Total	11555	2697	14252

Based on the confusion matrix in Table 5, the accuracy is 73.68%, it means that the model can classifate 73.68% of observations correctly, this shows that the predictive ability of the logistic discriminant model in predicting anemia and non-anemia groups is quite good. The sensitivity value obtained was 41.37%, which means that the model can classify 41.37% of the actual anemic observations. It means that the model can't predict well the observations of anemic women. The specificity value obtained was 97.91%, which means that the model can classify 97.91% of the actual non-anemic observations. This shows that the model can predict the observations of non-anemic women very well.

CONCLUSION

Based on the results of modeling using logistic discriminants with SMOTE on anemia status in women of reproductive age, the variables that affect anemia status in women of reproductive age with a significant level of 5% are pneumonia, tuberculosis, hepatitis, diabetes mellitus, malaria, gestational age, and age group. The logistic discriminant model obtained the sensitivity value is 41.37% it has not good ability to predict the anemia observations. However, it has a very good ability to predict the non-anemia observations, which produces a specificity value 97.91%. and obtained classification accuracy or overall model accuracy is 73.68%.

This study uses seven independent variables with six of them are categorical variables and one numerical variable. Therefore, it is hoped that further research can add other independent variables of numeric type.

ACKNOWLEDGEMENT

The authors would like to thank the National Research and Innovation Agency for the data in this study.

DECLARATION OF INTEREST

The author has no conflict of interest to declare.

REFERENCES

1. ABD ALMALEKI, P., MOKHTARI, D. M., Vahead, M. R., & GITI, M. (2004). Logistic discriminant analysis of breast cancer using ultrasound measurements.
2. Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.
3. Brière, M., Diedisheim, M., Dehghani, L., Dubois-Laforgue, D., & Larger, E. (2021). Anaemia and its risk factors and association with treatments in patients with diabetes: A cross-sectional study. *Diabetes & metabolism*, *47*(1), 101164.
4. Chang, C. H., Chen, K. Y., Lai, M. Y., & Chan, K. A. (2002). Meta-analysis: ribavirin-induced haemolytic anaemia in patients with chronic hepatitis C. *Alimentary pharmacology & therapeutics*, *16*(9), 1623-1632.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
6. de Mendonca, E. B., Schmaltz, C. A., Sant'Anna, F. M., Vizzoni, A. G., Mendes-de-Almeida, D. P., de Oliveira, R. D. V. C., & Rolla, V. C. (2021). Anemia in tuberculosis cases: A biomarker of severity?. *Plos one*, *16*(2), e0245458.
7. Everitt BS, Dunn G. 2001. Applied Multivariate Data Analysis. 2nd ed. New York (US): John Wiley and Sons
8. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463-484.

9. Salazar, D. I. G., Fuseau, M., Garrido, S. M., Vivas, G., & Gutiérrez, M. (2018). Prevalence of anaemia in children diagnosed with pneumonia in a Tertiary Hospital in Quito, Ecuador. *Journal of Nepal Paediatric Society*, *38*(2), 102-109.
10. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
11. Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis.
12. [MoH RI] Ministry of Health Republic of Indonesia. 2009. Age Classification by Category. Jakarta (ID): Director General of Health and Safety
13. [MoH RI] Ministry of Health of the Republic of Indonesia. 2013. Results of Basic Health Research (Riskesdas) 2013. Jakarta (ID): The Indonesian Ministry of Health Research and Development Agency.
14. Riskesdas, K. (2018). Main results of basic health research (RISKESDAS). *Journal of Physics A: Mathematical and Theoretical*, *44*(8), 1-200.
15. Keokenchanh, S., Kounnavong, S., Tokinobu, A., Midorikawa, K., Ikeda, W., Morita, A., ... & Sokejima, S. (2021). Prevalence of anemia and its associate factors among women of reproductive age in Lao PDR: evidence from a nationally representative survey. *Anemia*, *2021*(1), 8823030.
16. Lebso, M., Anato, A., & Loha, E. (2017). Prevalence of anemia and associated factors among pregnant women in Southern Ethiopia: A community based cross-sectional study. *PloS one*, *12*(12), e0188783.
17. Vexler, A., Chen, X., & Hutson, A. D. (2017). Dependence and independence: Structure and inference. *Statistical Methods in Medical Research*, *26*(5), 2114-2132.
18. Webb, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.
19. White NJ. 2018. Anemia and malaria. Malaria Journal. 17(1):371-388.
20. Anaemias, W. N. (2017). Tools for effective prevention and control. *World Health Organization: Geneva, Switzerland*, 1-83.