Low Birth Weight Classification With Synthetic Minority Over-Sampling Technique Random Forest

Sachnaz Desta Oktarina[1], Hari Wijayanto[2], Helena Ramadhini Yarah[3]
[1]Dept. of Statistics and Data Science, IPB University, Indonesia
sachnazdes@apps.ipb.ac.id
[2]Dept. of Statistics and Data Science, IPB University, Indonesia
hari@apps.ipb.ac.id
[3]Dept. of Statistics and Data Science, IPB University, Indonesia
helena_yarah@apps.ipb.ac.id

ARTICLE INFO

Article history:

Received October 3rd, 2022
Revised June 24th, 2023
Accepted July 31st, 2023

Keyword:

LBW, random forest, SMOTE

ABSTRACT

Low birth weight (LBW) is defined as a condition where the birth weight is less than 2500 grams. Infants born with LBW conditions are more susceptible to disease and have a higher risk of dying at an early age. LBW conditions that are prone to unbalanced data can be classified using the Synthetic Minority Oversampling Technique (SMOTE) random forest method. The analysis was processed on the 2017 Indonesian Demographic and Health Survey (IDHS) data to identify important variables in predicting the incidence of LBW. The results showed that the SMOTE random forest model provided an accuracy value of 79.84%, sensitivity of 30.99%, specificity of 83.6%, and AUC of 62%. Important variables in predicting the incidence of LBW were the number of antenatal care visits, wealth quantile, maternal age at delivery, iron supplementation, marital status, and twins' birth.

Corresponding Author:

Sachnaz Desta Oktarina
Dept. of Statistics and Data Science, IPB University, Indonesia
Email: sachnazdes@apps.ipb.ac.id

INTRODUCTION

Birth weight is an indicator of infant health and survival. Low birth weight (LBW) is defined as the condition of babies born weighing less than 2500 grams (1). Approximately 15% of babies worldwide are born LBW and most of them occur in Asia (2). Based on the 2017 Indonesian Demographic and Health Survey (IDHS), the prevalence of LBW in Indonesia is around 7%. LBW has a higher risk of early death in infants (3). Infants born with LBW conditions are also more susceptible to disease and growth and development disorders such as stunting (4). LBW can be caused by several factors derived from maternal characteristics, infant characteristics, and household characteristics (5). Prevention and treatment of mothers who have the potential to give birth to LBW babies can be done by classifying LBW. This aims to identify the condition of the baby before birth to reduce the impact of LBW.

Statistical methods that can be used for LBW classification include logistic regression, naive bayes, Support Vector Machine (SVM), Classification and Regression Tree (CART), random forest, and others. Hassan and Mirza (6) conducted a comparison between SVM, CART, naive bayes, random forest, and logistic regression methods. The results showed that random forest produced the highest accuracy value compared to other methods. Research conducted by Yuliati and Sihombing (7) also shows that the random forest model provides the best performance compared to classification and regression tree, naive bayes, and support vector machine. Therefore, research that aims to identify the most important variables in predicting the incidence of LBW will use the random forest classification method.

Data on LBW cases is unbalanced. Classification methods such as random forest are vulnerable to the problem of unbalanced data. Data imbalance occurs when the data has an unbalanced proportion between two or more data groups which are usually referred to as minority and majority classes. The ratio of data imbalance is about 1:4 to 1:100 (8). The problem of data imbalance can cause bias in parameter estimation and can lead to errors in decision making. Handling the data imbalance problem can be done by balancing the distribution of minority and majority classes through under sampling, oversampling, or a combination of both methods. The oversampling approach is more often used than under sampling because the under sampling method will eliminate data in the majority class so that it can cause the loss of important information from the data (9). One oversampling method that can be used is the Synthetic Minority Oversampling Technique (SMOTE). According to Chawla et al. (10), SMOTE can improve classifier accuracy for minority classes. Therefore, this study will use the method of handling unbalanced data with SMOTE random forest in identifying risk factors for LBW incidence.

METHOD

The data used is sourced from the Indonesian Demographic and Health Survey (IDHS). The IDHS is conducted every 5 years, therefore the latest data available and retrievable is the 2017 IDHS data. The samples taken for this study were infants who had a birth weight report. The response variable in this study was birth weight with the categories of LBW (birth weight < 2500 grams) and Non LBW (birth weight ≥ 2500 grams) (1). The explanatory variables used in this study refer to the research of Chhea et al. (5), Yuliati and Sihombing (7), and Oktriyanto et al. (11). The description of the variables can be seen in Table 1.

Analysis Procedure

The stages of data analysis in this study are as follows (see Figure 1):

1. Pre-processing the data, namely labelling the response variables where birth weight < 2500 grams is labelled 1 (LBW) and birth weight ≥ 2500 grams is labelled 0 (Non-LBW), and checking for missing data.
2. Exploring the data to find out the proportion between LBW and non-LBW, as well as to find out the general description of each variable.
3. Selecting explanatory variables that will be used in random forest modelling with the Chi-Square test.
4. Perform SMOTE random forest modelling using k-fold cross validation. Based on the recommendations of James et al. (12), k that will be used in this study is k = 10. The stages of modelling using 10-fold cross validation are as follows:
   a. Divide the data into 10-fold where 9-fold as training data and 1-fold as test data. Each fold will alternate as test data.
   b. Perform random forest modelling with SMOTE on the training data for each combination of m and n-tree. Based on the recommendation of Breiman and Cutler (13), the number of sorting variables (m) to be used is $\sqrt{p}/2$, $\sqrt{p}$, and $2\sqrt{p}$, where p is the number of explanatory variables used for modelling. The number of trees (n-tree) to be used are 50, 100, 200, 500, and 1000.

c. Evaluate the model on the test data and store the accuracy, sensitivity, specificity, and AUC values.
d. Repeating steps 4.b and 4.c until each fold serves as test data.
e. Calculate the average accuracy, sensitivity, specificity, and AUC values.

Table 1. Explanation of the variables used

Code	Variable	Category
Y	Birth weight	0 = Non LBW; 1 = LBW
X_*twin*	twin birth	0 = no; 1 = yes
X_urut	order of birth	0 = first born 1 = 2nd or 3rd born 2 = 4th or more born
X_*gap*	birth gap	0 = < 2 years; 1 = ≥ 2 years
X_*gender*	baby gender	1 = male; 2 = female
X_*age*	age of mother when give birth	0 = < 20 years old 1 = 20 – 35 years old 2 = > 35 years
X_Fe	iron supplement	0 = not consume; 1 = consume
X_*visit*	number of maternal check-up	1 = < 4 times; 2 = ≥ 4 times
X_kompl	pregnancy complication	0 = No; 1 = Yes
X_asur	insurance ownership	0 = No; 1 = Yes
X_*want*	pregnancy intention	1 = wanted pregnancy 2 = untimely pregnancy 3 = unwanted pregnancy
X_rokok	smoking habit	0 = No; 1 = Yes
X_*edu*	mother education	0 = no school 1 = Elementary school 2 = Secondary High school 3 = University
X_*job*	mother job	0 = Not working; 1 = Working
X_*marry*	marital status	0 = without status; 1 = Marriage status
X_*rurban*	domicile	1 = Urban; 2 = Rural
X_kaya	wealth quantile	1 = lowest 2 = medium to low 3 = medium 4 = medium to high 5 = highest
X_minum	Source of water	0 = not proper; 1 = proper
X_sani	Sanitation	0 = not proper; 1 = proper
X_*cook*	fuel for cooking	0 = otherwise; 1 = electricity/ gas
X_*hhsize*	household size	0 = ≤ 4 1 = 5 – 7 2 = > 7
X_*blexam*	blood-check during pregnancy	0 = no; 1 = yes

5. Identify important variables in predicting LBW incidence based on Mean Decrease Gini (MDG) and Mean Decrease Accuracy (MDA) values.

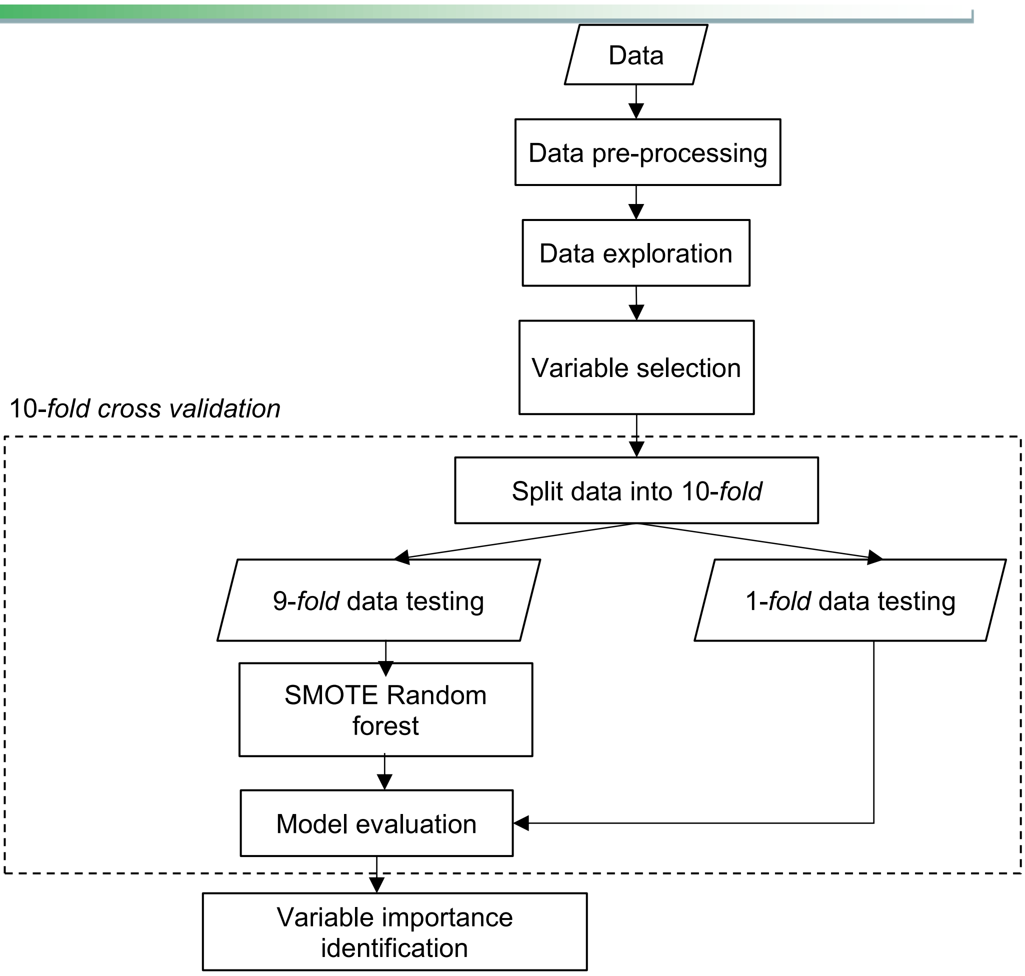

Figure 1. Flowchart of Data Analysis

RESULTS

The proportion of infants born with LBW condition was 7.1% (987) and infants born with non-BLW condition was 92.9% (12829) as shown in Figure 2. The difference in proportion between LBW and non-BLW indicates that the data is not balanced, where the minority class is LBW and the majority class is non-BLW.

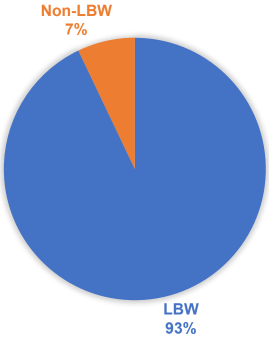

Figure 2. Pie chart of birth weight percentage
Source: IDHS 2017

Among babies without LBW, most of them do not have twins. According to Figure 3a, The percentage of twin births that experienced LBW was greater than the percentage of non-twin births that experienced LBW. Babies born as twins with LBW had a percentage of 63.89%, while babies born without twins with LBW had a percentage of 6.7%. LBW condition is considered also to be associated with iron consumption during maternity. Based on figure 3b, Percentage of mother who received iron supplement intake during prenatal care are dominated by those who gave birth to non-LBW babies. The percentage of mothers who did not take iron supplements and gave birth to LBW babies was 9.17%, while the percentage of mothers who took iron supplements and gave birth to LBW babies was 6.91%.

The percentage of birth weight based on the number of prenatal check-up visits can be seen in Figure 3c. Mothers who had less than 4 prenatal check-up visits and gave birth to LBW babies had a higher percentage than mothers who had 4 or more prenatal check-up visits. The percentage of mothers who had less than 4 check-up visits and gave birth to LBW babies was 11.08%, while the percentage of mothers who had 4 or more than 4 check-up visits and gave birth to LBW babies was 6.83%.

Furthermore, the percentage of LBW babies born based on marital status can be seen in Figure 3d. LBW babies born from the marital status of both parents who were not married had a higher percentage compared to LBW babies born from the status of both parents who were married. The percentage of LBW babies and the marital status of both parents who were not married was 11.39%, while the percentage of LBW babies and the marital status of both parents who were married was 6.95%.

The percentage of birth weight based on maternal age at delivery can be seen in Figure 3e. Mother’s age group less than 20 years at delivery indicated the highest percentage of LBW compared to other age groups. The age of mothers who were less than 20 years old when they gave birth to babies with LBW conditions had a percentage of 9.76%. Meanwhile, for those who gave birth to a LBW babies, the maternal age group between 20 to 35 years had a percentage of 6.92%. Subsequently, the maternal age of more than 35 years had a percentage of 7.25% of LBW babies.

LBW babies tend to be associated with wealth quintile group. The Figure 3f explained that the bottom wealth quintile group shows the highest percentage of LBW compared to the other wealth quintile groups. The lowest percentage of LBW is in the highest wealth quintile group. The percentage of LBW in the lowest wealth quintile group was 9.67% followed by the medium to low wealth quintile group at 7.3%, the medium group at 6.59%, the medium to high group at 6.31%, and the highest group at 5.35%.

For other indicators, the percentage of LBW was highest at birth spacing of less than 2 years, baby born with female sex, complications during pregnancy, mothers who smoked, mothers who did not work, rural residence, inadequate drinking water source, inadequate sanitation, cooking fuel other than electricity/gas, mothers who did not have blood tests during pregnancy, number of household members more than 5, order or the birth is 4th child or more, and mothers who did not go to school.

This group of explanatory variables was further screened before further analysis. This was done to avoid overfitting the model. Overfitting occurs when a model provides excellent predictions for training data but poorly predicts test data or new data (14). The explanatory variables selected for modeling are explanatory variables that seems to be related to the response variable. The tendency to relate between each explanatory variable and the response variable can be seen by conducting the Chi-Square test. Table 2 shows the tendency of association between each explanatory variable and the response variable based on the Chi-Square test. If the p-value is less than 5% then the null hypothesis is rejected, meaning that there is a tendency to relate between the explanatory variables and the response variable.

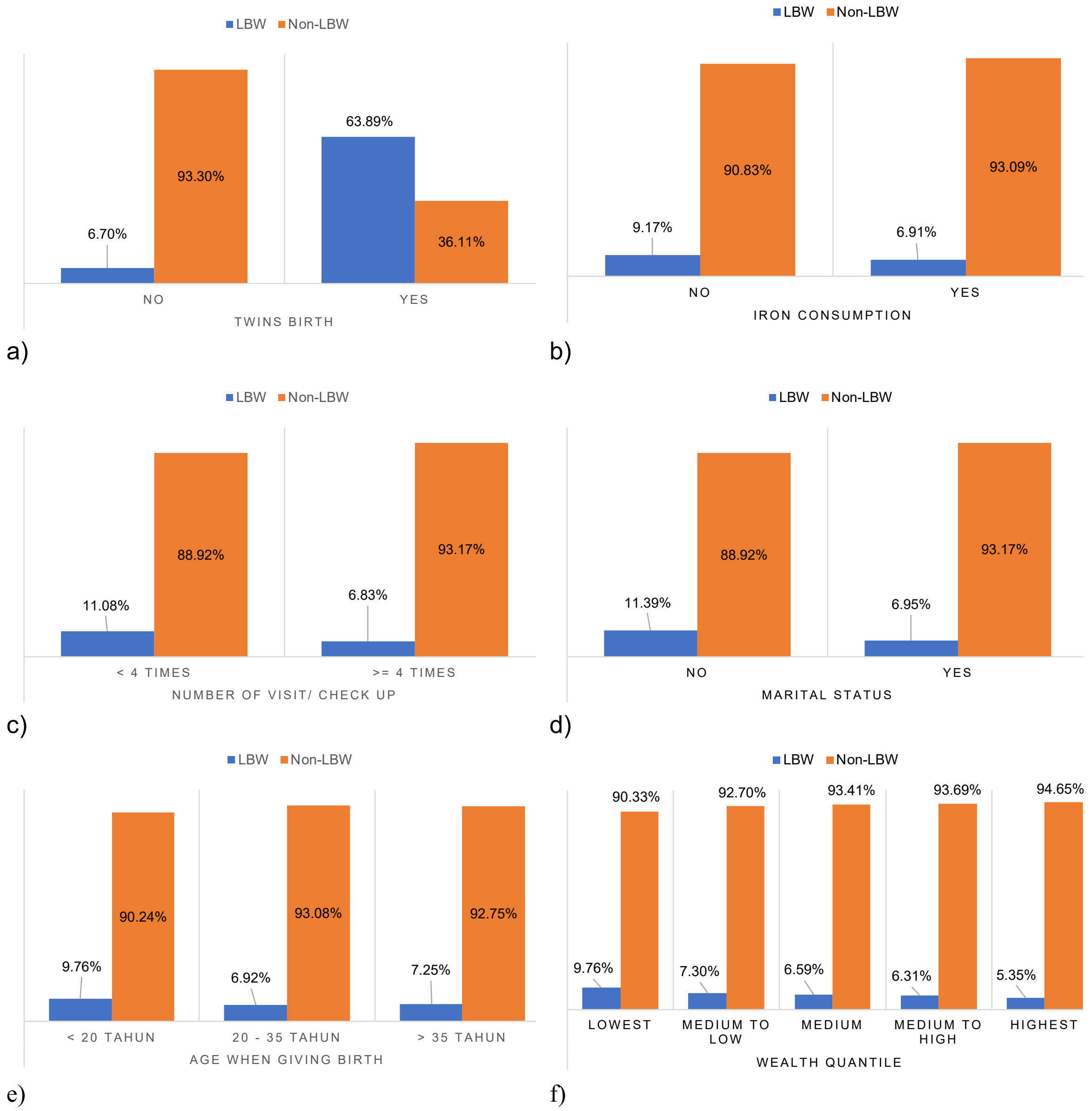

Figure 3 a)-f) Explorative analysis of LBW incidence with corresponding covariates.

There are 13 explanatory variables that have a tendency to be associated with the response variable including twin birth (X_twin), birth order (X_urut), birth spacing (X_gap), maternal age at delivery (X_age), iron supplementation (X_Fe), number of antenatal check-up visits (X_visit), complications during pregnancy (X_kompl), mothers education (X_edu), marital status (X_marry), wealth quintile (X_kaya), drinking water source (X_minum), sanitation (X_sani), and cooking fuel (X_cook). Eight other explanatory variables had p-values of more than 5%, so the null hypothesis was accepted, meaning that these variables had no tendency to be related to the response variable. Next, modeling was conducted using 13 explanatory variables that tended to be related to the response variable.

Table 2. Chi-square Analysis

Variable	p-value	Variable	p-value
X_*twin*	0,00*	X_*edu*	0,00*
X_urut	0,00*	X_*job*	0,14
X_*gap*	0,00*	X_*marry*	0,00*
X_*gender*	0,12	X_*rurban*	0,17
X_*age*	0,01*	X_kaya	0,00*
X_Fe	0,00*	X_minum	0,00*
X_*visit*	0,00*	X_sani	0,00*
X_kompl	0,00*	X_*cook*	0,00*
X_asur	0,14	X_*hhsize*	0,06
X_*want*	0,34	X_*blexam*	0,25
X_rokok	0,96		

* Tends to be related at the 5% significant level

SMOTE random forest classification begins with generating synthetic data on the training data and determining the parameters of the number of variables (m) and the number of trees (n-tree) to be used. The m values used are √p/2 = 2, √p = 4, and 2√p = 7, where p is the number of explanatory variables of 13. The ntree values used are 50, 100, 200, 500, 1000. Modeling was conducted using 10-fold cross validation. The selection of the optimal combination of m and n-tree was seen based on the average AUC value. The average value of AUC with various combinations of m and ntree is presented in Figure 4. The greater the value of m makes the AUC value decrease. The greater the ntree value used, the greater the resulting AUC value. The value of m = 2 and ntree = 1000 is the optimal parameter for the SMOTE random forest classification model because it has the largest average AUC value, which is 0.62.

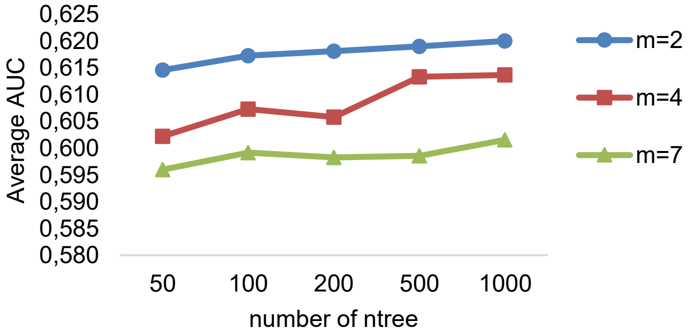

Figure 4 Average AUC against many trees (ntree) based on the explanatory variable of the parser (m) using SMOTE random forest.

Furthermore, the SMOTE random forest model is good enough to predict the incidence of LBW. The SMOTE Random Forest model showed an accuracy value of 79.84%, sensitivity of 30.99%, specificity of 83.6%, and AUC of 62%. The SMOTE Random Forest can explain the level of importance of variables in classification.

The importance of variables in the classification can be seen through the Mean Decrease Gini (MDG) value. MDG shows the influence of the explanatory variables based on how much the Gini index decreases in the explanatory variables during the sorting process of the decision tree model formation. The greater the MDG value, the higher the importance of the explanatory variable in determining the condition of birth weight in infants.

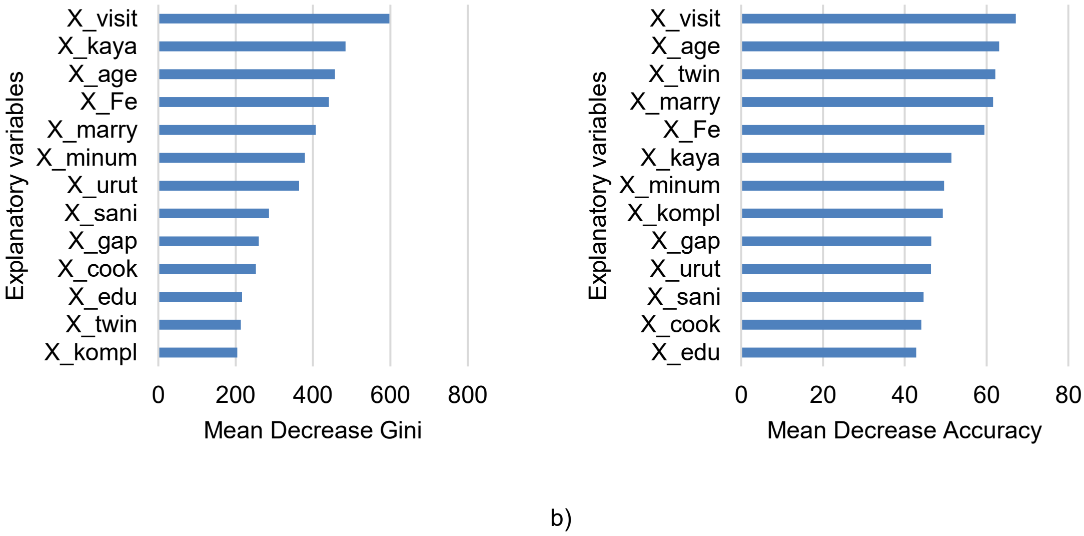

Figure 5a) Mean Decrease Gini and 5b) Mean Decrease Accuracy of SMOTE Random Forest

Based on Figure 5a, the variable number of antenatal care visits (X_visit) has the largest MDG value. It is followed by the variables of wealth quintile (X_kaya), maternal age at delivery (X_age), iron supplement (X_Fe), and marital status (X_marry).

The importance of variables in classification can also be seen through the Mean Decrease Accuracy (MDA) value. MDA shows the influence of explanatory variables based on how much the accuracy of the model decreases if the explanatory variables are not included in the decision tree model. The greater the MDA value, the higher the importance of the explanatory variable in determining the condition of birth weight in infants. Based on Figure 5b, the variable number of antenatal check-up visits (X_visit) has the largest MDA value. This is followed by the variables of maternal age at delivery (X_age), multiple births (X_twin), marital status (X_marry), and iron supplements (X_Fe).

The SMOTE random forest indicated risk factors in LBW incident in Indonesia. Among the disadvantages of this analysis is that the result is difficult to be interpreted and requires proper model tuning for the data. However, once the importance variable has been discovered, the interpretation can be supported by looking back to the explorative analysis and Chi square analysis.

DISCUSSION

The results of the Chi square test showed that multiple births tended to be associated with LBW. This is in line with the results of research conducted by Tonasih and Kumalasary (15), and Hartiningrum et al. (16). Mothers who are pregnant with twins often experience uterine distension, which is a condition of increasing the size of the uterus that is not in accordance with gestational age, causing premature birth and babies born with small sizes. In addition, the variables of maternal age at delivery and pregnancy spacing also have a relationship with the incidence of LBW. The age of the mother who is too young or too old can increase the risk of LBW. The close spacing of pregnancies will affect the process of calcium loss in the bones, which can increase the risk of LBW (17).

Other variables that tended to be associated with LBW were birth order, iron supplementation, number of antenatal check-up visits, and wealth quintile. The greater the birth order, the higher the probability of LBW. Mothers who never took iron supplements during pregnancy had a greater risk of giving birth to LBW babies. Mothers who made at least 4 antenatal check-up visits were less likely to give birth to LBW babies. The higher the level of wealth, the less likely LBW will occur. Mothers with low levels of wealth experience greater stress that affects their pregnancy conditions (18).

Maternal education and sanitation tend to be associated with the incidence of LBW which is in line with the results of research by Sohibien and Yuhan (19). The higher the mother's education, the easier it will be for the mother to obtain and digest information about pregnancy, so that the mother can take better care of her pregnancy so that the baby is born in good health. Inadequate sanitation has a greater risk of viral and bacterial infections so that nutritional intake is hampered and causes low birth weight.

Mothers who experience complications during pregnancy can increase the risk of giving birth to LBW babies (11). Babies born to parents without marriage ties will increase the risk of LBW (20). In addition, the variables of drinking water source and cooking fuel also tend to be associated with the incidence of LBW based on the results of the Chi square test. Mothers who live in areas with inadequate drinking water sources are more vulnerable to bacterial infections or harmful substances that can inhibit the absorption of nutrients and risk giving birth to LBW babies (19). The use of hygienic cooking fuels such as electricity/gas can reduce the likelihood of delivering LBW babies, while unhygienic cooking fuels such as biomass burning can increase the likelihood of delivering LBW babies (21).

The results indicated that low birthweight incidence was associated with number of antenatal visits, class of wealth, age of the mother, iron supplement consumption, and marital status. The strength of this study contributed to the knowledge on how machine learning techniques can be used to decipher relationship between LBW and its determining factors. The random forest analysis, as a part of machine learning techniques can deal with data that is not normally distributed (non-parametric). Furthermore, this technique also equipped with SMOTE analysis to balance the occurrence of minority event (incident of LBW) so that the outcome is more representative to the population. However, this result is still far from perfection because it is not discussing other LBW determining factors such as placenta factors, mothers' level of activity, infertility, race, prior LBW incidence, and so on.

CONCLUSION

The SMOTE random forest model showed an accuracy value of 79.84%, sensitivity of 30.99%, specificity of 83.6%, and AUC of 62%. Important variables in predicting the incidence of LBW were the number of antenatal care visits, wealth quintile, maternal age at delivery, iron supplementation, marital status, and multiple births. In future research, classification modeling can be done by adding other explanatory variables that are relevant in influencing the incidence of LBW and using other classification methods such as XGboost, Adaboost, K-Nearest Neighbor, Discriminant Analysis, Neural Network, and so on. Suggestions for the government are expected to create programs in efforts to prevent and treat LBW based on the level of importance of variables that affect the incidence of LBW, such as providing counseling on the importance of making pregnancy check-up visits to health facilities.

AUTHOR CREDIT STATEMENT

The writing of this article certainly cannot be separated from the support and assistance of various parties. The author would like to thank Dr. Ir. I Made Sumertajaya, M.Si who has provided advice on the writing of this scientific work. All lecturers and administrative staff of the Department of Statistics IPB University who have provided knowledge and other supporting facilities. All parties who have helped the author in preparing this scientific work that the author cannot mention one by one. Hopefully this scientific work will be useful for those who need it and for the advancement of science.

DECLARATION OF COMPETING INTEREST

The authors declared that this research has no competing interest.

REFERENCES

1. KC, Anil; BASEL, Prem Lal; SINGH, Sarswoti. Low birth weight and its associated risk factors: Health facility-based case-control study. PloS one. 2020; 15.6: e0234907.
2. UNICEF-WHO; United Nations Children’s Fund, World Health Organization. Low birthweight estimates: Levels and trends 2000–2015. Geneva: World Health Organization; 2019.
3. [BKKBN]. Badan Kependudukan dan Keluarga Berencana Nasional. Jakarta: Survei Demografi dan Kesehatan Indonesia 2017. 2018. Available from: www.sdki.bkkbn.go.id
4. [Kemenkes RI]. Kementerian Kesehatan Republik Indonesia. Jakarta: Situasi Balita Pendek (Stunting) di Indonesia. 2018.
5. Chhea C, Ir P, Sopheab H. Low birth weight of institutional births in Cambonia: Analysis of the demographic and health surveys 2010-2014. PLoS One. 2018; 13(11):1–16.
6. Hassan MM, Mirza T. Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. International Journal of Computer Applications. 2020; 175(17):42-53.
7. Yuliati IF, Sihombing PR. Penerapan Metode *Machine Learning* dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. *MATRIK J Manajemen, Tek Inform dan Rekayasa Komput*. 2021; 20(2):417–426.
8. Johnson,Justin M.; Khoshgoftaar, Taghi M. Survey on deep learning with class imbalance. Journal of Big Data. 2019: 6.1: 1-54.
9. Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. IEEE Computational Intelligence Magazine. 2018; 13(4):59-76.
10. Parsa, Amir Bahador, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 2020, 136: 105405..
11. Oktriyanto, Rahardja MB, FN DN, Amrullah H, Pujihasvuty R, PN MM. Determinants of Low Birth Weight in Indonesia. Jurnal Kesehatan Masyarakat. 2022; 17(4):583-593.
12. James G, Witten D, Hastie T, Tibshirani R, Taylor J. Springer Texts in Statistics An Introduction to Statistical Learning with Applications in Python. New York: Springer. 2023.
13. Breiman L, Cutler A, Liaw A, & Matthew Wiener. Package ‘randomForest’: Breiman and Cutler's Random Forests for Classification and Regression. 2022; 4.7-1.1. Available from: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
14. Hidayatulloh NGT. Perbandingan Kinerja *Random Forest* dan *Double Random Forest* untuk Klasifikasi Status Kemiskinan di Level Kabupaten/Kota [undergraduate thesis]. Bogor: Departemen Statistika, Fakultas MIPA, Institut Pertanian Bogor. 2022.
15. Tonasih, Kumalasary D. Faktor-Faktor Yang Mempengaruhi Kejadian Berat Bayi Lahir Rendah (BBLR) Di Puskesmas Wilayah Kecamatan Harjamukti Kota Cirebon Tahun 2016. Jurnal Riset Kebidanan Indonesia. 2018; 2(1):21-27.

16. Hartiningrum I, Fitriyah N. Bayi berat lahir rendah (BBLR) di Provinsi Jawa Timur tahun 2012-2016. Jurnal Biometrika Dan Kependudukan. 2018 Dec;7(2):97-104.
17. Agustin S, Setiawan BD, Fauzi MA. Klasifikasi Berat Badan Lahir Rendah (BBLR) pada Bayi dengan Metode Learning Vector Quantization (LVQ). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2019; 3(3):2929-2936.
18. Zaveri A, Paul P, Saha J, Bikash B, Pradip C. Maternal determinants of low birth weight among Indian children Evidence from the National Family Health Survey-4, 2015-16. PLoS ONE. 2020; 15(12).
19. Sohibien GPD, Yuhan RJ. Determinan Kejadian Berat Badan Lahir Rendah (BBLR) di Indonesia. Jurnal Aplikasi Statistika dan Komputasi Statistik. 2019; 11(1):1-14.
20. Barr JJ, Marugg L. Impact of Marriage on Birth Outcomes: Pregnancy Risk Assessment Monitoring System, 2012-2014. Linacre Q. 2019; May;86(2-3):225-230.
21. Patel R, Chauhan S. Risk of low birth weight and exposure to type of cooking fuel in India. International Journal of Pregnancy & Childbirth. 2020; 6(1):8-11