TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

The Utilization of Resampling Techniques and the Random Forest Method in Data Classification

Ciciana, Rahmawati, Laila Qadrini*

Fakultas Matematika dan Ilmu Pengetahuan Alam, Program Studi Matematika, Universitas Sulawesi Barat, Majene, Indonesia
Email: [1]ciciana99@email.com, [2]rahmah@unsulbar.ac.id, [3,*]laila.qadrini@unsulbar.ac.id
Correspondence Author Email: laila.qadrini@unsulbar.ac.id

Abstract−In data classification, there are various methods that can be employed, one of which is the random forest method. This method proves effective in handling non-linear data, exhibiting robustness against extreme data points and disturbances, and providing ease of use that results in high-quality classification outcomes. Data imbalance, where one class has more or fewer instances than the others, is a common issue. In situations of data imbalance, most classification models tend to favor the majority class, which can lead to overfitting and unsatisfactory classification results. To address this issue, resampling techniques can be applied. One such resampling technique is SMOTE, specifically an oversampling method that augments the minority class by generating synthetic data points. This research aims to evaluate the accuracy of data classification using the random forest method and assess the impact of resampling and random forest on classification. The data used in this study includes simulated breast cancer data and real-world patient data from LBW Puskesmas Banggae I Kabupaten Majene. The analysis results indicate an accuracy rate of 94.74%, a sensitivity of 93.33%, and an F1-Score of 95.89% for breast cancer data. Meanwhile, the accuracy for LBW data reached 73.75%, with a sensitivity of 77.63%, and an F1-Score of 84.89%.

Keywords: LBW; Resampling; Random Forest; SMOTE; Imbalance

1. INTRODUCTION

Fundamentally, data represents a collection of information or details about a subject acquired through observation or research from specific sources. Data, when acquired but not yet processed, can transform into a fact or assumption. Data classification, on its own, is the process of associating metadata characteristics with each asset in a digital domain, identifying the type of data linked to that asset.

Low Birth Weight (LBW) babies are infants born with a weight below 2,500 grams (Saifuddin, 2002). LBW includes infants born with a birth weight below 2,500 grams, regardless of gestational age. Factors that can influence the occurrence of LBW include the mother's age (Setianingrum, 2005), parity or the number of live births a woman has had (Setiati & Rahayu, 2017), abortions or miscarriage history (Lestariningsih & Duarsa, 2013), and gravidity or the number of pregnancies (Manuaba et al., 2008).

Based on these factors, we can perform classification to minimize the likelihood of mothers giving birth to low birth weight babies. For this research, the classification method used is random forest. The random forest method is a technique capable of addressing non-linear problems. Random forest offers numerous advantages, such as resilience to outliers and noise, user-friendliness, and the ability to provide robust classification outcomes with low error rates while effectively handling missing data (T. S. Lestari & Sirodj, 2021). Data imbalance refers to a scenario where class distribution is uneven, with one class having more or fewer instances than others. In imbalanced conditions, most classifications tend to favor the majority class, with machine classifiers more inclined to predict the majority class and disregard the minority class (Choirunnisa & Lianto, 2018). Moreover, as per Gong & Kim (2017), imbalance can lead to overfitting, suboptimal model creation, and play a significant role in misclassification (Mellor et al., 2015). To address this issue, data resampling or sampling techniques (Fadilah, 2018) can be employed.

Resampling is widely utilized to tackle imbalanced data issues by attempting to balance the original data using a set of sampling algorithms, adjusting the sample size in different classes, and then training the "balanced" new data using classification algorithms (Syukron & Subekti, 2018).

In prior research by WS & Nooraeni (2020) titled "Application of Resampling Methods in Addressing Imbalanced Data on Determinants of Diarrhea Cases in Toddlers in Indonesia," it was concluded that the application of the SMOTE method is highly suitable for improving the accuracy of multiple logistic regression analysis and avoiding overfitting in diarrhea data for Indonesian toddlers in 2017, which exhibited imbalanced characteristics (imbalanced ratio). Another study titled "Credit Card Transaction Fraud Classification Using the Random Forest Method" by A. Lestari et al. (2020) achieved an accuracy of 97.275%, sensitivity of 98.795%, precision of 97.976%, F-Measure of 98.384%, and an AUC value of 94.065%, categorizing it as an excellent classification due to consistently falling within the 90-100% range. Another research by Qadrini et al. (2022) titled "Oversampling, Undersampling, SMOTE SVM, and Random Forest in Classifying Bidikmisi Recipients in East Java in 2017" concluded that the application of random sampling oversampling and SMOTE yielded nearly identical AUC values and could be applied to imbalanced data cases, producing high accuracy, precision, recall, and AUC values without overfitting or underfitting. Based on the introductory information and previous research, the title of this study is "Resampling and Random Forest Methods for Data Classification".

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

2. RESEARCH METHODS

2.1 Research Methodology

The present work utilizes the resampling technique incorporating the Synthetic Minority Over-sampling Technique (SMOTE) algorithm to tackle the challenge of data imbalance. Additionally, the random forest classification approach is employed. In addition, the dataset is partitioned into training and testing subsets, following an 80% to 20% ratio. The following figure 1 outlines the steps of research that were undertaken.

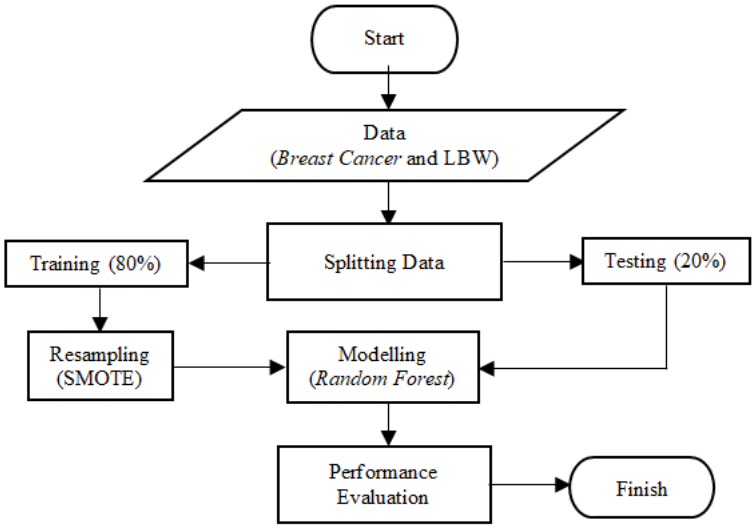

Figure 1. *Flowchart* Alur penelitian

The steps taken to classify data in this research are as follows:

a. Prepare the data, then divide it into training data (80%) and testing data (20%).
b. Perform data resampling using the SMOTE technique, with the following steps:
   1. In this stage, the data used is from the minority class that will be replicated.
   2. Calculate the Euclidean distance for each data point.
   3. Use the Euclidean distance to generate synthetic data.
c. Conduct classification using the random forest method, with the following steps:
   1. Create bootstrap samples or take samples with replacement from a size of N from the data cluster.
   2. Select m variables randomly from p variables, where $m \leq p$. m is usually chosen as an approximation of the square root of the total number of p variables, $\lfloor\sqrt{p}\rfloor$. According to Leo Breiman, the value of m can also be obtained from twice the square root of the total number of p variables $\left(m = 2\lfloor\sqrt{p}\rfloor\right)$ and half the square root of the total number of p variables $\left(m = \frac{1}{2}\lfloor\sqrt{p}\rfloor\right)$.
   3. After randomly selecting m, grow the tree without pruning.
   4. Steps 1-3 are repeated n times to form a forest (classification) with n trees.
   5. Once the forest is formed, the misclassification error is calculated to obtain the optimal mtry and achieve more stable variable importance levels.
   6. For class prediction, use majority vote.
d. Evaluate the model's performance using a confusion matrix.

2.2 Resampling

Resampling is a technique of randomly and freely taking additional samples from existing samples. This technique provides equal chances for original samples to be included in the new sample, which can have a smaller or larger size than the original sample. Resampling is the most commonly used method to address class imbalance. There are two approaches. In cases of imbalance, approaches can be made at both the data and algorithm levels. In cases of imbalance, several common issues arise (Choirunnisa, 2019):

a. Outliers, which occur when data has extreme values that differ significantly from the majority of the group.
b. Overlapping data between classes. If there is overlapping, discriminative rules become challenging to process.
c. Some data points in sub-clusters have very close distances between the two classes (small disjunction).

2.3 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (Synthetic Minority Oversampling Technique) is a derivative of oversampling (Siringoringo, 2018). SMOTE is a method for managing imbalanced data first introduced and proposed by (Chawla et al., 2002). The basic idea behind SMOTE is to increase the number of samples in the minority class to balance it with the majority class by generating

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

synthetic data based on the k-nearest neighbor. The nearest neighbors are selected based on the Euclidean distance between the data points (Chawla et al., 2002). For example, given data with p variables $x^T = [x_1, x_2, \ldots, x_p]$ and $z^T = [z_1, z_2, \ldots, z_p \quad]$, the Euclidean distance $d(x, z)$ can be calculated as follows:

$$d(x,z) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \ldots + (x_p - z_p)^2} \quad (1)$$

Synthetic data generation is performed using the following equation:

$$X_{syn} = X_i + (X_{knn} - X_i)\gamma \quad (2)$$

Where
X_{syn} : Synthetic data,
X_i : The i-th data from the minority class,
X_{knn} i : The data from the minority class with the nearest neighbor to X_i
γ : A Random number between 0 and 1.

The value X in Figure 2.4 represents a sample, and X1, X2, X3, X4, and X5 are its nearest neighbors. SMOTE generates new data (synthetic data) m1, m2, and m3 along a line between X and each of its nearest neighbors.

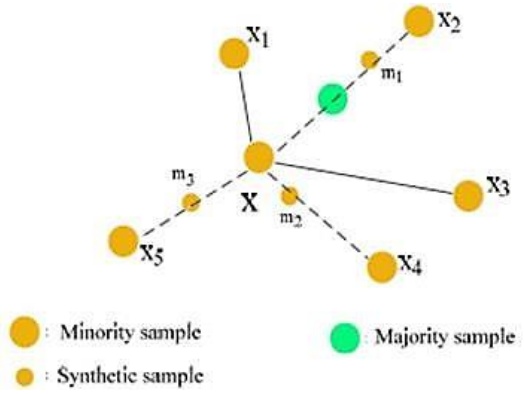

Figure 2. SMOTE Illustration

2.4 Random Forest

Random forest was first introduced by Breiman in 2001. Random forest is a modification of bagging. In random forest, there is an addition of random sub-sampling or the selection of m variables used in building a tree. The tree-building process in random forest does not involve pruning.

Random forest is a classification method that consists of a collection of classification trees. Let $\{h(x, \theta_k), k = 1, \ldots\}$ where $\{\theta_k\}$ are independent and identically distributed randoms, and each tree chooses the majority class from the data (majority vote). Given an ensemble $h_1(x), h_2(x), \ldots, h_k(x)$ with training data randomly selected from the distribution of random vectors y and x, the margin function $(mg(x,y))$ of the random forest is defined as follows:

$$mg(x,y) = \frac{\sum_1^K I(h_k(x)=y)}{K} - \max_{j \neq y} \left[\frac{\sum_1^K I(h_k(x)=j)}{K}\right] \quad (3)$$

where I is the indicator function and K is the number of trees. The margin function is used to measure the degree of the majority vote on and the average vote from the other classes.

Strength is the average size of the strength of a single tree. A larger value of indicates better prediction accuracy. The value is defined as follows:

$$s = E_{x,y} mg(x,y) \quad (4)$$

The average correlation $\bar{\rho}$ between pairs of guesses from two single trees in the random forest is defined as follows:

$$\bar{\rho} = \frac{E_{\theta,\theta'}\left(\rho(\theta,\theta')sd(\theta)sd(\theta')\right)}{E_{\theta,\theta'}(sd(\theta)sd(\theta'))} \quad (5)$$

Where $\rho(\theta,\theta')sd(\theta)sd(\theta')$ is the correlation between trees.
The upper limit of prediction error (ε_{RF}) by random forest is:

$$\varepsilon_{RF} \leq \bar{\rho}\left(\frac{1-s^2}{s^2}\right) \quad (6)$$

From this equation, it can be said that to achieve a small error, the correlation and strength must be small. Therefore, it is necessary to modify the values of m and ntree. By decreasing the value of m, the correlation and

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

strength are reduced. The same applies to the value of ntree. If ntree is large, it means that the data similarity between each tree is very high. However, if the values of m and ntree are very low, each tree will lose some important information and increase the error. Therefore, the selection of m and ntree in random forest is crucial. Here is the random forest algorithm:

a. Create a bootstrap sample or take samples with replacement Z from a size N of the data set.
b. Randomly select m variables from p variables, where $m \leq p$. Usually, the best m size is chosen by approximating the square root of the total number of p variables, which is $\lfloor\sqrt{p}\rfloor$. According to Leo Breiman, the value of m can also be obtained from twice the square root of the total number of p variables $\left(m = 2\lfloor\sqrt{p}\rfloor\right)$ and half of the square root of the total number of p variables $\left(m = \frac{1}{2}\lfloor\sqrt{p}\rfloor\right)$.
c. After randomly selecting m, the tree is grown without pruning. The best node split in a tree is done using the Gini index.
d. Steps 1-3 are repeated n times to create a forest (classification) with n trees.
e. After the forest is formed, the misclassification error (Out of Bag Error) is calculated to obtain the optimal mtry, and a more stable variable importance level is obtained.
f. For predicting a class, a majority vote is used.

2.5 Confusion Matrix

The confusion matrix is a tool used in the field of machine learning to evaluate the performance of a classification model. It is a square matrix that displays the number of correct and The utilization of the confusion matrix is a prevalent approach in the computation of accuracy and error rate. Accuracy refers to the proportion of correctly identified cases in relation to the total number of instances, whereas the error rate pertains to the cases that have been mistakenly identified relative to the total number of cases (Alber, 2021). The confusion matrix enables the determination of accuracy, error rate, precision, and recall values. In the context of performance measurement utilizing the confusion matrix, four distinct terms are employed to reflect the outcomes of the classification procedure. These terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The TP value denotes the accurate identification of positive data, whereas TN signifies the correct identification of negative data. Conversely, FP refers to the misclassification of negative data as positive, and FN represents the misclassification of positive data as negative (Saifullah, 2019). The confusion matrix for binary classes, which refers to datasets containing only two class types (Siringoringo, 2018), is presented in Table 1.

Tabel 1. *Confusion Matrix*

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

To determine the accuracy value, it can be obtained using the following equation: (Saifullah, 2019).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Sensitivity = \frac{TP}{FP+TP} \quad (8)$$

$$F1 - Score = 2 \times \frac{presisi \times recal}{presisi+recal} \quad (9)$$

To obtain the value of the F1-Score, we need the values of precision and recall. The following formula can be used to calculate precision.

$$Precision = \frac{TP}{FP+TP} \quad (10)$$

Recall, on the other hand, is the ratio of relevant items selected to the total number of relevant items available. To calculate recall, you can use the following formula:

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

where:
TP (True Positive) : The number of positive data correctly classified.
TN (True Negative) : The number of negative data correctly classified.

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

FN (False Negative) : The number of negative data incorrectly classified.
FP (False Positive) : The number of positive data incorrectly classified.

3. RESULTS AND DISCUSSION

3.1 General Data Overview

The data used in this study consists of 569 Breast Cancer data points with 31 variables and 419 LBW data points with 5 variables.

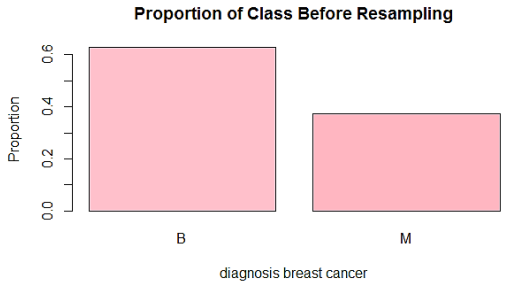

Figure 3. Distribution of Majority and Minority Classes in Breast Cancer Data

In the Breast Cancer data, the dependent variable (Y) is diagnosis with label B (benign) as the majority class with 357 data points and label M (malignant) as the minority class with 212 data points.

Figure 4. Distribution of Majority and Minority Classes in LBW Data

In the LBW data, the dependent variable (Y) is BBL with label 0 (no LBW) as the majority class with 377 data points and label 1 (LBW) as the minority class with 30 data points.

3.1.1 Data Preprocessing

Data preprocessing aims to standardize the data range. Each variable Y in each data point, i.e., the diagnosis variable which was previously of character type, was recoded and detected as a factor type. Similarly, the LBW variable, which was previously a number type, was recoded and detected as a factor type. This recoding process was performed using Rstudio.

3.1.2 Data Training and Testing Selection

The research data was divided into training and testing data sets to enable predictions. The data was divided into an 80% training set and a 20% testing set. More details can be seen in the table 2 below:

Table 2. Division of Breast Cancer Data for Training and Testing

Jenis Data	Jumlah Data		Total
	B	M	
Training	282	173	455
Testing	75	39	114

For Breast Cancer data, 455 data points were used for training and 114 data points for testing.

Table 3. Division of LBW Data for Training and Testing

Jenis Data	Jumlah Data		Total
	0	1	
Training	300	26	326
Testing	77	4	81

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

For LBW data, 326 data points were used for training and 81 data points for testing.

3.2 Data Resampling

The results of resampling using the SMOTE technique can be seen in the table 4 below:

Table 4. Resampling Results for Breast Cancer Data Using SMOTE

Kelas	Jumlah Data Sebelum Dilakukan Resampling	*SMOTE*
B	282	228
M	173	228

From Table 4, it can be observed that the M (malignant) class has an increased number of data points, initially having 173 data points and increasing to 228 after applying SMOTE. The proportion of Breast Cancer data can be seen in Figure 5 and compared to Figure 2.

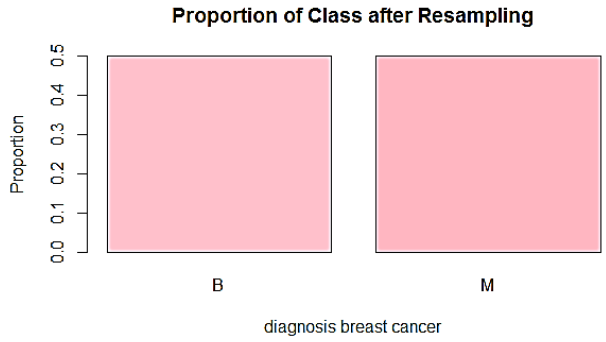

Figure 5. Proportion of Breast Cancer Data After SMOTE

Table 5 shows that the label class 1 (LBW) has an increased number of data points, initially having 26 data points, and increasing to 162 after applying SMOTE.

Table 5. Resampling Results for LBW Data Using SMOTE

Kelas	Jumlah Data Sebelum Dilakukan Resampling	*SMOTE*
0	300	163
1	26	162

The proportion of LBW data can be seen in Figure 6 and compared to Figure 3.

Figure 6. Proportion of LBW Data After SMOTE

3.3 Random Forest Classification

After observing the predictions from resampling, random forest classification is performed using the testing data. The initial step in random forest classification is to select mtry and ntree.

3.3.1 Selection of mtry and Determining ntree

In performing random forest classification, it is important to determine the number of independent variables used in classification. This selection is done randomly. The mtry values for this study are as follows:

mtry for Breast Cancer data

$$mtry = \frac{1}{2}\left|\sqrt{30}\right| = 2$$

mtry for LBW data

$$mtry = \frac{1}{2}\left|\sqrt{4}\right| = 1$$

TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

$mtry = |\sqrt{30}| = 5$ $mtry = |\sqrt{4}| = 2$

$mtry = 2|\sqrt{30}| = 10$ $mtry = 2|\sqrt{4}| = 4$

The next step is the selection of ntree. A value of ntree = 50 has provided satisfactory classification results (Breiman & Cutler, 2001). However, it is suggested that a larger ntree value above 100 can lead to constant misclassifications. Therefore, the ntree values used in this study are 25, 50, 100, 500, and 1000.

3.3.2 Out Of Bag Error (OOB Error)

Out of Bag Error (OOB Error) is the average misclassification of samples that are not included in the forest. It affects the determination of the smallest error when selecting mtry.

Table 6. OOB Error for Breast Cancer Data

mtry	*ntree*				
	25	50	100	500	1000
2	3,73%	2,63%	2,85%	3,29%	3,51%
5	4,17%	3,51%	3,95%	2,85%	2,85%
10	4,82%	4,61%	3,73%	3,07%	2,85%

From Table 6, it can be observed that the smallest OOB Error value is obtained with mtry = 2 and ntree = 50, with a value of 2.63%. This indicates that with these mtry and ntree values, optimal classification can be achieved.

Table 7. OOB Error for LBW Data

mtry	*ntree*				
	25	50	100	500	1000
1	20,62%	21,23%	24,62%	20,31%	19,69%
2	18,46%	16,92%	16,62%	17,23%	17,54%
4	14,77%	16%	17,23%	16,31%	16,31%

For LBW data, the smallest OOB Error value is obtained with mtry = 4 and ntree = 25, with a value of 14.77%. Therefore, to achieve optimal classification for LBW data, mtry = 4 and ntree = 25 are recommended.

3.3.3 Variable Importance

One of the outputs of random forest is variable importance. Variable importance indicates which independent variables have the most influence on the dependent variable.

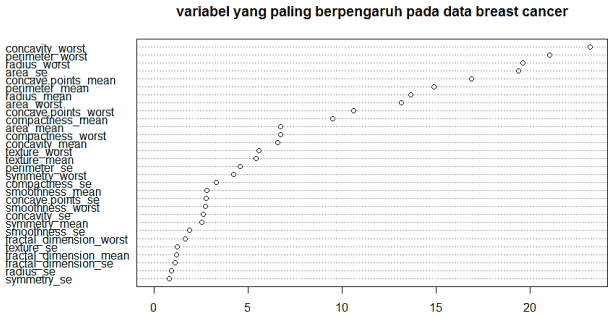

Figure 7. Most Influential Variables in Breast Cancer Data

In Breast Cancer data, the variable "concavity worst" is the most influential, indicating that "concavity worst" is the most determining factor for whether a patient has a benign or malignant breast cancer diagnosis.

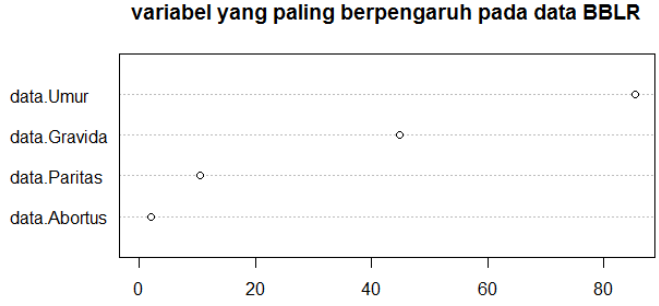

Figure 8. Most Influential Variables in LBW Data


TIN: Terapan Informatika Nusantara
Vol 4, No 4, September 2023, page 252-259
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v4i4.4342

In LBW data, the most influential variable is the mother's age, indicating that the mother's age significantly influences whether a baby

4. CONCLUSION

The accuracy results of the random forest method for Breast Cancer and BBLR data are 94.74% and 73.75%, respectively. These values are smaller than the accuracy results in the previous research. However, the obtained values are quite satisfactory and fall within the performance range of a fairly good model. From the breast cancer data, a sensitivity result of 0.9333 and an F1-Score of 0.9589 were obtained, indicating that the model's predictive ability is at 93.33%. For the BBLR data, sensitivity and F1-Score values of 0.7763 and 0.8489, respectively, were obtained.

REFERENCES

Alber, J., 2021, Klasifikasi Data Mining Untuk Menentukan Tingkat Kepuasan Pengguna Transaksi Bus Trans Metro Pekanbaru Menggunakan Metode *Naive Bayes*, *Skripsi*, Program Pasca Sarjana Teknik, Universitas Islam Riau, Pekanbaru.

Breiman, L., 2001, *Random Forest*, *Machine Learning*, 45, 5-32.

Chawla, N.V. dkk, 2002, *SMOTE* Boast: Improving Prediction Of The Minority Class In Boosting, Proc. Knowl, Discov, PP, Hal: 107-119.

Choirunnisa, S., 2019, Metode Hibrida Oversampling dan Undersampling Untuk Menangani Ketidakseimbangan Data Kegagalan Akademik Universitas XYZ, *Tesis*, Program Magister Komputer, Institut Teknologi Sepuluh Nopember, Surabaya.

Depkes RI, 2009, Pedoman Pelayanan Kesehatan Bayi berqat Lahir Rendah (LBW) Dengan perawatan Metode Kanguru Di Rumah Sakit dan Jejaringannya, Jakarta: Bakti Husada.

Fadilah, L., 2018, Klasifikasi *Random Forest* Pada Data *Imbalance*, *Skripsi*, Program Pasca Sarjana Matematika, UIN Syarif Hidayatullah, Jakarta.

Lestari, T.S. & Agustin Nuriani Sirodj, D., 2021, Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest, *Jurnal Riset Statistik*, No. 2, Volume 1, Hal: 160-167.

Lestariningsih, S. & Arta Budi Sisila, D., 2014, Hubungan Preeklasia Dalam Kehamilan Dengan Kejadian LBW di RSUD Jenderal Ahmad Yani Kota Metro Tahun 2011, *Jurnal Kesehatan Masyarakat*, No. 1, Vol. 8, Hal: 32-39.

Manuaba, 2008, Gawat Darurat Obstetri Ginekologi Dan Obsetri Ginekologi Sosial Untuk Profesi Bidan, Jakarta : EGC.

Manuaba, Ida Ayu Chandranita, dkk. *Ilmu Kebidanan, Penyakit kandungan dan KB*. Jakarta : EGC; 2010

Muqiit WS, A. dkk., 2020, Penerapan Metode Resampling Dalam Mengatasi *Imbalance* Data Pada Determinan Kasus Diare Pada Balita di Indonesia, *Jurnal Matematika dan Statistika Serta Aplikasinya*, No. 1, Vol. 8, Hal: 19-27.

Pangastuti, S.P., 2018, Perbandingan Metode *Ensemble Random Forest* Dengan *Smote-Boosting* Dan *Smote-Bagging* Pada Klasifikasi Data Mining Untuk Kelas Imbalance (Studi Kasus : Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur), *Tesis*, Program Magister sains, Institut Teknologi Sepuluh Nopember, Surabaya.

Qadrini, L. dkk., 2022, *Oversampling*, *Undersampling*, *SMOTE* SVM dan *Random Forest* Pada Klasifikasi Penerima Bidikmisi Se Jawa Timur Tahun 2017, No. 4, Vol.3, Hal: 386-391.

Qadrini, L. Seppewali, A, Aina, A. (2021). Decision Tree dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial, *Jurnal Inovasi Penelitian*. 2(7): 2722-9475.

Saifuddin, A.B., 2009, Panduan Praktis Pelayanan Kesehatan Maternal dan Neonatal, Jakarta: EGC.

Saifullah, 2019, Deteksi Kelayakan Fisik Air Untuk Konsumsi Menggunakan *Naive Bayes Clasifier*, *Skripsi*, Program Pasca Sarjana Komputer, UIN Maulana Malik Ibrahim, Malang.

Setianingrum, S., 2005, Hubungan Antara Kenaikan Berat Badan, Lingkar Lengan Atas, Kadar Hemoglobin Ibu Hamil Trimester III Dengan Berat Bayi Lahir di Puskesmas Ampel Boyolali, Jurnal Semarang.

Setiati, A.R. & Rahayu, S., 2017, Faktor Yang Mempengaruhi Kejadian LBW (Berat Badan Lahir Rendah) Di Ruang perawatan Intensif Neonatus RSUD DR Moewardi Di Surakarta, *Jurnal Keperawatan Global*, No. 1, Vol. 2, Hal: 1-61.

Siringoringo, R., 2018, Klasifikasi Data Tidak Seimbang Menggunakan Algoritma *SMOTE* dan k-Nearest Neighbor, Jurnal ISD, No. 1, Vol. 3, Hal: 44-49.

Syukron, A. & Subekti, D., 2018, Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit, *Jurnal Informatika*, No. 2, Vol. 5, Hal: 175-185.