**Journal of Applied Informatics and Computing (JAIC)**
Vol.9, No.4, August 2025, pp. 1587~1592
e-ISSN: 2548-6861


# SMOTE and Weighted Random Forest for Classification of Areas Based on Health Problems in Java

**Erwan Setiawan [1]*, Bagus Sartono [2]**, Khairil Anwar Notodiputro [3]****
* Mathematics Education of Suryakancana University
**School of Data Science, Mathematics, and Informatics IPB University
erwan@unsur.ac.id [1], bagusco@apps.ipb.ac.id [2], khairil@apps.ipb.ac.id [3]

**Article Info**

***Article history:***

Received 2025-06-23
Revised 2025-07-14
Accepted 2025-07-19

***Keyword:***

*Random Forest,*
*SMOTE,*
*Weighted Random Forest.*

**ABSTRACT**

Random Forest (RF) is a popular Machine Learning (ML) approach extensively employed for addressing classification issues. Nevertheless, the RF method for classification problems demonstrates suboptimal performance in cases of data imbalance. There are several approaches to enhance RF performance when coping with data imbalance issues, such as using weighting and oversampling. This research explores the intervention of RF in addressing data imbalances, focusing on case studies of health problem classification in Java This study aims to develop models to analyze the health status of regions using RF, WRF, SMOTE-RF, and SMOTE-WRF methods. The objective is to compare the performance of these models and identify the best model for classifying DBK and Non-DBK categories in Java. The research results show that SMOTE-WRF is the most effective model in classifying DBK, achieving an accuracy level of 93.62%, sensitivity of 85.71%, precision of 75.00%, F-score of 80.00%, and AUC of 93.57%. The three key variables in the SMOTE-WRF model entail access to adequate sanitation, egg and milk consumption, and the number of doctors.



## I. INTRODUCTION

Random Forest (RF) is a popular Machine Learning (ML) approach extensively employed for addressing classification issues. Instances of its application include the identification of underweight infants [1], the categorization of Palembang songket motifs [2], the assessment of creditworthiness [3], and the classification of human development index [4]. RF's functionality involves constructing multiple decision trees and amalgamating their predictions to ascertain the predominant class. Notably, RF excels in minimizing variance and mitigating overfitting problems associated with single decision trees, rendering it more dependable and precise for predictive purposes.

The issue of imbalanced data is frequently encountered in classification problems. Data imbalance arises when the distribution of classes in a dataset is skewed, with one class having significantly more or fewer samples than another. For instance, in the context of credit card fraud detection, the number of legitimate transactions (majority class) may far exceed the number of fraudulent transactions (minority class). In such scenarios, accurately predicting fraudulent transactions becomes more crucial than predicting legitimate ones. Nevertheless, the RF method for classification problems demonstrates suboptimal performance in cases of data imbalance. This is due to the RF method's tendency to favor predicting the majority class over the minority class, resulting in a biased model that performs poorly in predicting the minority class [5].

There are several approaches to enhance RF performance when coping with data imbalance issues, such as using weighting and oversampling. Within RF, the weighting technique, known as Weighted Random Forest (WRF), involves assigning higher weights to the minority classes during the bootstrapping process, thus making the model more attuned to the minority classes [6]. Compared to XGBoost or SVM, the WRF model provides a balance between performance and interpretability, particularly in imbalanced settings where adjusting class weights can directly improve minority class detection without the need for

complex optimization procedures [7][8]. Oversampling is implemented using the Synthetic Minority Oversampling Technique (SMOTE) method, which aims to increase the number of minority class data points to match the majority class by generating synthetic data based on the nearest k-neighbors [9]. Both of these methods have proven to be effective in enhancing RF performance when dealing with imbalanced data [10][6][11]. SMOTE is often preferred over ADASYN and Random Oversampling because it reduces overfitting by generating synthetic minority samples through interpolation, providing more stable and generalizable models for imbalanced data classification [12].

This research explores the intervention of RF in addressing data imbalances, focusing on case studies of health problem classification in Java. Assessing the health status of an area is crucial for gaining insight into the overall well-being of its population. This information is valuable for planning health programs, allocating health resources, and formulating more impactful policies. As part of the effort to assess the health status of an area, Indonesia's Ministry of Health's Health Research and Development Agency (*Badan Penelitian dan Pengembangan Kesehatan* - Balitbangkes) has compiled the Public Health Development Index (*Indeks Pembangunan Kesehatan Masyarakat* - IPKM). An area is considered to have health issues (*Daerah Bermasalah Kesehatan* - DBK) when its IPKM value falls below the average [13]. DBK refer to regions that face significant challenges in delivering health service due to issues such as limited healthcare personel, inadequate health infrastructure, geographic isolation, or poor public health status [13]. The classification of DBK and Non-DBK areas has real-world implications, as it informs public health funding, prioritization of government interventions, and equitable distribution of local health budgets. Ensuring precision in this classification is therefore essential for evidence-based decision-making and targeted policy design.

According to the 2018 IPKM data, 18 out of 119 regencies/cities in Java were found to have DBK status [14]. This reveals an imbalance in the data, with the DBK class (minority class) representing 15% of the samples and the Non-DBK class (majority class) representing 85% of the samples. As a result, this study aims to develop models to analyze the health status of regions using RF, WRF, SMOTE-RF, and SMOTE-WRF methods. The objective is to compare the performance of these models and identify the best model for classifying DBK and Non-DBK categories in Java.

## II. METHOD

The data used in this research is the 2018 Community Health Development Index (IPKM) data, which was obtained from various sources. These sources include the Basic Health Research (Riskesdas – *Riset Kesehatan Dasar*) 2018, Village Potential (Podes – *Potensi Desa*) 2018, and Susenas March 2018 integrated Riskesdas 2018 surveys conducted by the Litbangkes Agency and BPS. The sample used in this research includes IPKM value data from 119 districts/cities located on the Java island. The IPKM value is used as a response variable and is classified into two categories, namely Health Problem Areas (DBK) if the IPKM value is less than 0.6087 and Non-DBK if the IPKM value is greater than or equal to 0.6087. The average IPKM value in 2018 is 0.6087. The predictor variables that were used in this research can be found in Table 1.

We have chosen 11 predictor variables that represent different areas affecting the health status of a region, including environment, nutrition, health access/facilities, economy, education, and demographics. The data we have collected is from the year 2018 and we obtained it from the official website of the National Central Statistics Agency (BPS), BPS-DKI Jakarta, BPS-Banten, BPS-West Java, BPS-Central Java, BPS- DI Yogyakarta, and BPS-East Java.

TABLE 1.
THE PREDICTOR VARIABLES

| Variable | Description |
|---|---|
| X1 | Percentage of households with access to clean water |
| X2 | Percentage of access to proper sanitation |
| X3 | Weekly egg and milk consumption per capita in thousands of rupiah |
| X4 | Average weekly per capita consumption of beef in kilograms |
| X5 | The ratio of the number of health centers per 1000 residents |
| X6 | The ratio of the number of doctors per 1000 residents |
| X7 | Percentage of poor people |
| X8 | District/city minimum wages in millions of rupiah |
| X9 | Percentage of literacy rate for ages over 15 years |
| X10 | The average length of schooling in years |
| X11 | Male/female sex ratio *100 |

The study uses the Random Forest (RF) method for modeling. This method is employed to classify areas as either DBK or Non-DBK. As the data used is unbalanced, the RF method is developed through class weighting and balancing the number of samples in each class. Class weighting is done by assigning a higher weight to the minority class, which is called Weighted Random Forest (WRF). Balancing the number of samples in each class is done by generating synthetic samples to add to the minority class, which is called Synthetic Minority Oversampling Technique (SMOTE).

Experiments will be conducted in this research on four models, namely RF, WRF, SMOTE-RF, and SMOTE-WRF, using a data sample divided into training and testing data in a 60:40 ratio. Based on the results of hyperparameter tuning using the Random Search method, it was found that the optimal parameters to be used for modeling are ntree = 822, mtry = 3, and nodesize = 10. Each model will use 822 trees (ntree = 822) since classification error tends to be constant when using ntree $\geq$ 100 [15]. The number of separating variables for each model is 3 variables (mtry = 3), which is in line with Breiman and Cutler's suggestion that the number of separating features is $\sqrt{p}$, where $p$ is the total number of predictor variables [16]. In the WRF method, weighting is

performed by grid search method on the training data, and the optimal weight value is selected based on accuracy, sensitivity, and precision. The analysis procedure carried out can be seen in Figure 1.

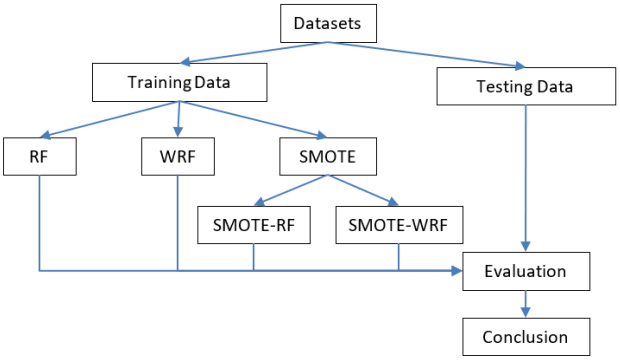
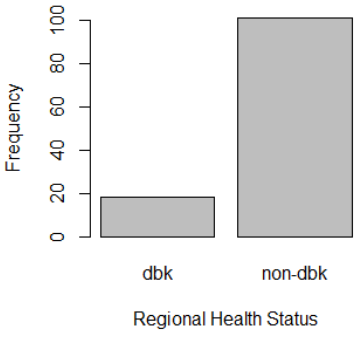



Figure 1. The analysis procedure

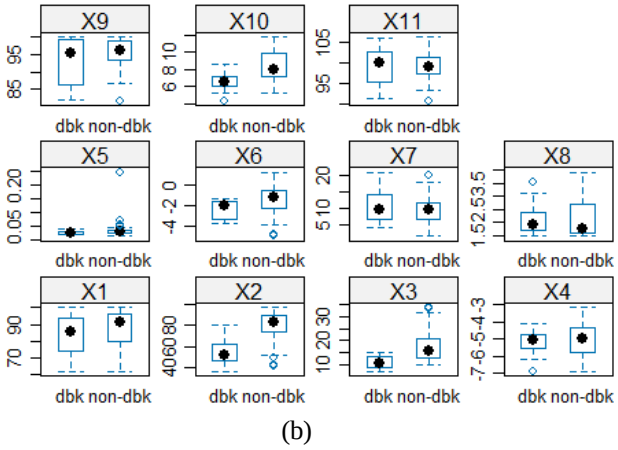
The dataset used to create the model is split into two subsets - training data and testing data. The former is utilized to train four models - RF, WRF, SMOTE-RF, and SMOTE-WRF. On the other hand, the latter is used to assess the performance of the models and compare their results. To create the SMOTE-RF and SMOTE-WRF models, the training data is first balanced for the number of samples in each class using the SMOTE method. Then, the RF and WRF methods are applied.

## III. Result And Discussion

A statistical description of the research data used is presented in Figure 2. Figure 2(a) shows that the frequency of the response variables is unbalanced, with the frequency of the Non-DBK class being higher than that of the DBK class. To classify areas with health problems, SMOTE and WRF are used in IPKM modeling to address this imbalance. In Figure 2(b), it can be seen that variables X4 and X6 have a lot of outliers. These outliers negatively impact the RF method as they can affect prediction accuracy, model stability and overfitting. Therefore, it is necessary to handle variables X4 and X6 by transforming them using the logarithmic function.




(a)

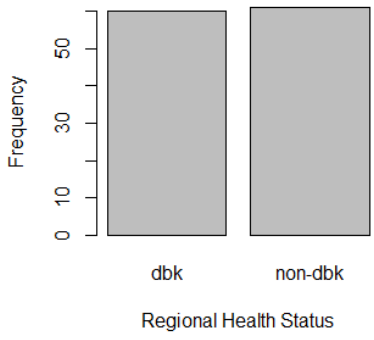



(b)

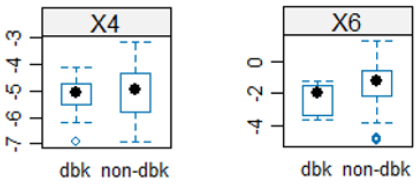
**Figure 2**. (a) the response variable distribution, (b) the prediktor variables distribution

Figure 3 presents two important preprocessing steps taken to improve model performance: class balancing using SMOTE and outlier reduction through logarithmic transformation. Subfigure (a) shows the distribution of the response variable (DBK vs. Non-DBK) after applying the Synthetic Minority Over-sampling Technique (SMOTE). Initially, the dataset exhibited a severe class imbalance, with only 15% of observations belonging to the DBK class. Through SMOTE, synthetic samples were generated for the minority class by interpolating between nearest neighbors, resulting in a balanced class distribution. This step is crucial in preventing model bias toward the majority class and has been shown to significantly enhance classifier performance in imbalanced datasets [12].

Subfigure (b) displays the boxplots of predictor variables X4 (beef consumption) and X6 (number of doctors) after applying a logarithmic transformation. The original distributions of these variables contained extreme outliers, which could distort model learning and increase the risk of overfitting, especially in tree-based algorithms like Random Forest. By applying a log transformation, the distribution is normalized and the influence of extreme values is reduced, as evidenced by the more symmetric and compact boxplots. This preprocessing strategy aligns with standard best practices in machine learning for stabilizing variance and improving model robustness [17].




(a)




(b)

Figure 3. (a) Balanced response variable distribution resulting from SMOTE, (b) Distribution of variables X4 and X6 after logarithmic transformation to address outliers

The corrected data is split into two parts: training data and testing data, with a 60:40 ratio. This ratio is commonly used when the number of samples is not too large. The training data is then used to develop four models: RF, WRF, SMOTE-RF, and SMOTE-WRF. The parameters used for these models are ntree = 822 for the number of trees, 3 variables for the separation variable, and 10 for the node size. Additionally, the weights for the minority class in the WRF model need to be adjusted to achieve a good level of sensitivity, precision, and accuracy.

Figure 4 shows how adjusting the minority class weight in the WRF model affects sensitivity, precision, and accuracy. Increasing the weight improves sensitivity, peaking at $10^6$, but leads to a decline in precision and a slight decrease in accuracy when the weight becomes too high. A weight of $10^6$ is selected as optimal because it offers a balanced trade-off, enhancing the model's ability to detect minority cases (DBK) without significantly compromising overall performance. This finding aligns with prior studies which emphasize that assigning class weights is an effective strategy for handling imbalanced data, as it helps the model account for asymmetric misclassification costs and improves minority class recall [18][19][20].

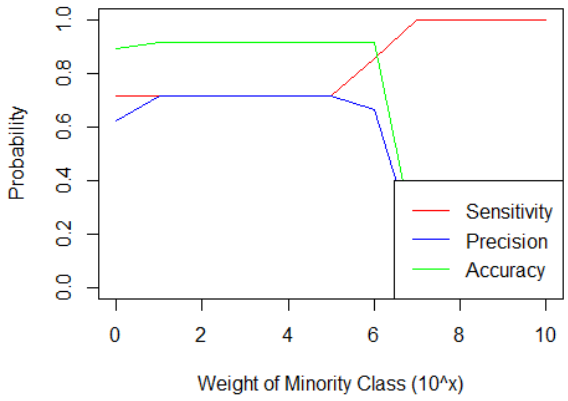



Figure 4. Experiment of several weight values for minority class

The model algorithm was developed using the R programming language. The results of the modeling are presented in Table 2 and Figure 5, which show that the four models developed have a good level of accuracy, above 90%. However, the RF model has poor performance when it comes to sensitivity, as it is only 57%. This confirms that conventional RF models are not effective in handling imbalanced data, especially when it comes to identifying minority classes.

TABLE 2.
PERFORMANCE COMPARISON IN TRAINING AND TESTING DATA

| | RF | | WRF | | SMOTE-RF | | SMOTE-WRF | |
|---|---|---|---|---|---|---|---|---|
| | **Training** | **Testing** | **Training** | **Testing** | **Training** | **Testing** | **Training** | **Testing** |
| **Accuracy** | 97.22 | 91.49 | 94.44 | 93.62 | 98.61 | 91.49 | 95.83 | 93.62 |
| **Sensitivity** | 81.82 | 57.14 | 100.00 | 85.71 | 100.00 | 71.43 | 100.00 | 85.71 |
| **Specificity** | 100.00 | 97.50 | 93.44 | 95.00 | 98.36 | 95.00 | 95.08 | 95.00 |
| **Minority Precision** | 100.00 | 80.00 | 73.33 | 75.00 | 91.67 | 71.43 | 78.57 | 75.00 |
| **Mayority Precision** | 96.83 | 92.86 | 100.00 | 97.44 | 100.00 | 95.00 | 100.00 | 97.44 |
| **F-Score** | 90.00 | 66.67 | 84.62 | 80.00 | 95.65 | 71.43 | 88.00 | 80.00 |
| **AUC** | 99.70 | 89.82 | 99.85 | 93.21 | 99.85 | 93.57 | 99.18 | 93.57 |

On the other hand, the WRF and SMOTE-WRF models have a sensitivity level above 85%, making them good models in terms of sensitivity. In alternative measurement criteria, both WRF and SMOTE-WRF demonstrate equivalent performance, achieving a specificity of 95.00%, minority precision of 75.00%, majority precision of 97.44%, and an 80.00% f-score. Nevertheless, there is one specific metric where SMOTE-WRF outperforms WRF: the AUC, with SMOTE-WRF exhibiting a 0.36% higher value than WRF. Additionally, if we look at the ROC-AUC curve in Figure 5, we can see that the WRF, SMOTE-RF, and SMOTE-WRF models have the largest area under the curve, indicating their superior performance.

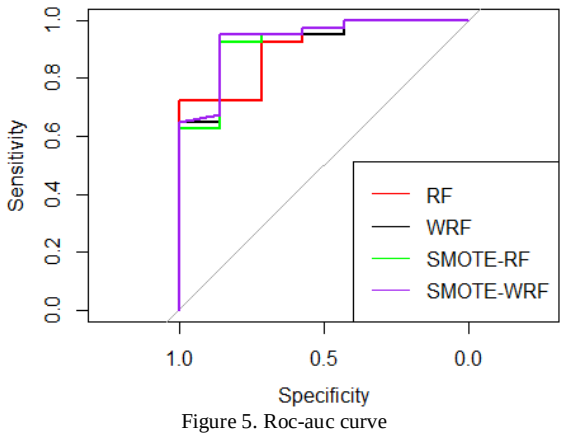



Figure 5. Roc-auc curve

After comparing several models, it has been concluded that the SMOTE-WRF model is the most effective for classifying health problem areas in Java. Figure 6 indicates that three variables - X2 (access to proper sanitation), X3 (consumption of eggs and milk), and X6 (number of doctors) - have a significant contribution to the SMOTE-WRF model. These variables have the highest values for both Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG), meaning they are frequently used to divide nodes in trees and are effective at separating target classes. Removing any of these variables from the model will result in a decrease in accuracy, demonstrating their importance to the model's performance.

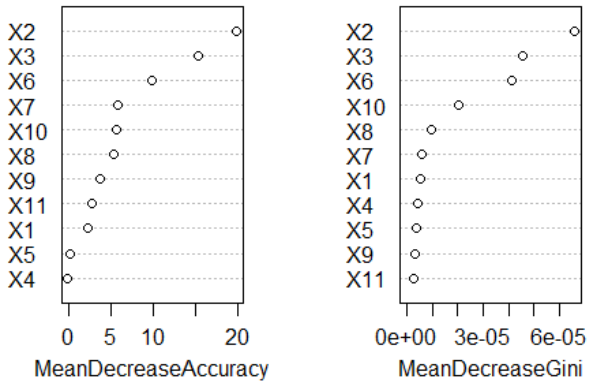



Figure 6. Variable importance

The summary statistics of the three most important variables, namely X2, X3, and X6 can be seen in Table 3.

TABEL 3.
SUMMARY STATISTICS OF THE THREE MOST IMPORTANT VARIABLES

| Variable | Class | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| X2 | DBK | 35,67 | 80,37 | 53,48 | 11,87 |
| | Non-DBK | 41,30 | 90,19 | 80,48 | 12,15 |
| X3 | DBK | 7,19 | 15,15 | 11,18 | 2,40 |
| | Non-DBK | 10,30 | 33,65 | 17,49 | 5,78 |
| X6 | DBK | 0,03 | 0,29 | 0,14 | 0,10 |
| | Non-DBK | 0,01 | 3,45 | 0,54 | 0,68 |

Table 3 presents the summary statistics of the three most influential variables identified by the SMOTE-WRF model, namely X2 (access to proper sanitation), X3 (egg and milk consumption), and X6 (number of doctors per 1,000 residents). These variables exhibit substantial differences between the DBK and Non-DBK classes, indicating their strong discriminative power.

For X2, the average access to proper sanitation in DBK regions is 53.48%, compared to 80.48% in Non-DBK areas, suggesting that inadequate sanitation is a key characteristic of health problem regions. Similarly, X3 shows that DBK regions have much lower per capita weekly expenditures on eggs and milk (IDR 11.18k) than Non-DBK areas (IDR 17.49k), reflecting disparities in nutritional access. For X6, the mean number of doctors per 1,000 residents in DBK regions is 0.14, whereas in Non-DBK regions it reaches 0.54, highlighting significant gaps in health service availability.

These findings confirm the relevance of these variables in distinguishing health-deprived areas and are consistent with previous research that links poor sanitation, inadequate nutrition, and limited access to healthcare professionals with low public health outcomes [21][22]. The ability of the model to prioritize these factors underscores its potential utility in guiding targeted policy interventions.

Figure 7 illustrates the relationship between access to proper sanitation (X2) and the classification of regions as either DBK or Non-DBK. The histogram displays the frequency and proportion of each class across five intervals of X2 values. It is evident that regions with low sanitation coverage—particularly those in the [50,60) and [60,70) intervals—are more likely to be categorized as DBK, with DBK proportions reaching 44% and 36%, respectively. In contrast, in regions where sanitation access exceeds 70% (i.e., intervals [70,80) and [80,100]), nearly all areas are classified as Non-DBK, with Non-DBK proportions at 100% and 98%, respectively.

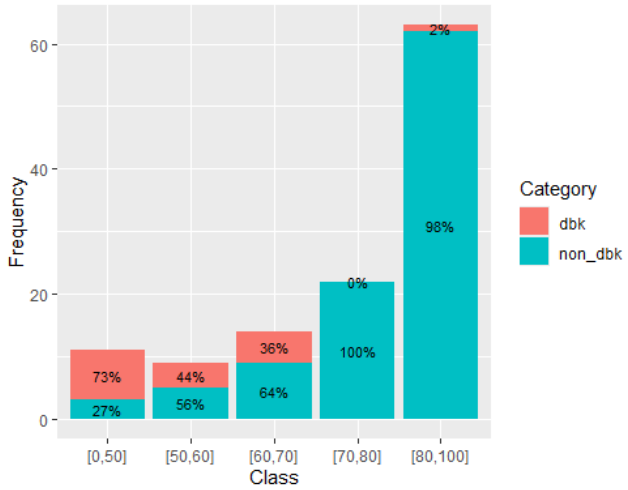



Figure 7. Histogram of variable X2 with category proportion

This pattern highlights X2 (sanitation access) as a strong discriminator between health-deprived and non-deprived areas. The stark contrast in class proportions across intervals supports the model's variable importance results, where X2 was identified as one of the top predictors. This finding aligns with the global health literature, which emphasizes that inadequate access to sanitation is strongly associated with higher risks of infectious diseases, child mortality, and overall poor health outcomes—especially in low- and middle-income countries [23] (UNICEF & WHO, 2021; Prüss-Ustün et al., 2019). Therefore, improving sanitation access is not only a health infrastructure concern but also a critical policy lever for addressing regional health disparities.

## IV. CONCLUSION

This research concludes that the SMOTE and Weighted Random Forest methods are effective in dealing with data imbalance. The RF model, when combined with weighting and SMOTE, performs better than the conventional RF model. Out of the four models formed - RF, WRF, SMOTE-RF, and SMOTE-WRF - the best model in classifying the status of health problem areas is the SMOTE-WRF. With an accuracy rate of 93.62%, sensitivity of 85.71%, precision of 75.00%, F-score of 80.00%, and AUC of 93.57%, it aligns with the theoretical properties of SMOTE-WRF, which tends to have better performance. SMOTE increases the number of minority class samples, while class weighting in RF helps the model focus on minority classes in the learning process. Access to adequate sanitation, egg and milk consumption, and the number of doctors are the three critical variables in the SMOTE-WRF model. Policymakers can use these variables as a reference to improve health status.

## ACKNOWLEDGMENT

The author would like to thank BPI Kemenristekdikti for providing a scholarship so that the author was able to continue his studies and write this article.

## REFERENCES

[1] H. R. Yarah, "Perbandingan Random Forest Dan SMOTE Random Forest Pada Klasifikasi Berat Badan Lahir Rendah (BBLR)," IPB University, 2023.

[2] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 2, pp. 310–320, 2020.

[3] O. Pahlevi, A. Amrin, and Y. Handrianto, "Implementasi Algoritma Klasifikasi Random Forest Untuk Penilaian Kelayakan Kredit," *J. Infortech*, vol. 5, no. 1, 2023, doi: https://doi.org/10.31294/infortech.v5i1.15829.

[4] T. Posangi, L. Yahya, and D. Wungguli, "Implementasi Algoritma Random Forest Dengan Forward Selection Untuk Klasifikasi Indeks Pembangunan Manusia," *JAMBURA J. Probab. Stat.*, vol. 4, no. 2, 2023, doi: https://doi.org/10.37905/jjps.v4i2.18460.

[5] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," 2004, [Online]. Available: https://api.semanticscholar.org/CorpusID:7308660

[6] L. Budianti and Suliadi, "Metode Weighted Random Forest Dalam Klasifikasi Prediksi Kelangsungan Hidup Pasien Gagal Jantung," in *Bandung Conference Series: Statistics*, 2022, pp. 103–110.

[7] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[8] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[9] J. Prasetya and Abdurakhman, "Comparison Of SMOTE Random Forest And SMOTE K-Nearest Neighbors Classification Analysis On Imbalanced Data," *Media Stat.*, vol. 15, no. 2, pp. 198–208, 2022.

[10] C. M. Lauw, J. X. Guterres, M. M. Huda, I. Saifudin, H. Hairani, and M. Mayadi, "Combination of SMOTE and Random Forest Methods for Lung Cancer Classification," *Int. J. Eng. Comput. Sci. Appl.*, vol. 2, no. 2, pp. 59–64, 2023, doi: 10.30812/IJECSA.v2i2.3333.

[11] A. Tesfahun and D. L. Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction," 2013.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, 2002, doi: https://doi.org/10.1613/jair.953.

[13] Kementerian Kesehatan RI, *Buku Saku Penanggulangan Daerah Bermasalah Kesehatan*. Jakarta: Badan Penelitian dan Pengembangan Kesehatan, 2011. [Online]. Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/3011/

[14] D. H. Tjandrarini, I. Dharmayanti, S. Suparmi, and O. Nainggolan, *Indeks Pembangunan Kesehatan Masyarakat 2018*. Jakarta: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan (LPB), 2019.

[15] C. D. Sutton, "Classification and Regression Trees, Bagging, and Boosting," *Handb. Stat.*, vol. 24, pp. 303–329, 2005, doi: https://doi.org/10.1016/S0169-7161(04)24011-1.

[16] L. Breiman and A. Cutler, "Manual--Setting Up, Using, And Understanding Random Forests V4.0," 2003. [Online]. Available: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2016.

[18] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[19] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal. An Int. J.*, vol. 6, no. 5, 2002, doi: 10.3233/IDA-2002-6504.

[20] A. Fernandez, S. Garcia, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham, Switzerland: Springer, 2018.

[21] World Health Organization (WHO), "Primary health care on the road to universal health coverage: 2019 monitoring report," 2019. [Online]. Available: https://www.who.int/publications/i/item/9789240029040

[22] C. G. Victora, A. Wagstaff, J. A. Schellenberg, D. Gwatkin, M. Claeson, and J.-P. Habicht, "Applying an equity lens to child health and mortality: more of the same is not enough," *Lancet*, vol. 362, no. 9379, pp. 233–241, 2003, doi: 10.1016/S0140-6736(03)13917-7.

[23] A. Prüss-Ustün *et al.*, "Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on lowand middle-income countries," *Int. J. Hyg. Environ. Health*, vol. 222, no. 5, pp. 765–777, 2019, doi: 10.1016/j.ijheh.2019.05.004.