Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status

V M Putri, M Masjkur* and C Suhaeni

3Department of Statistics, IPB University, Bogor, Indonesia

*E-mail: masjkur@apps.ipb.ac.

Abstract. The alternative approach to predict the classification of HB vaccination status is using a machine learning approach such as random forest and naive Bayes classifier. However, for imbalanced classification, the algorithms are biased towards the majority class. To increase the accurate prediction of the classifier, we consider the Synthetic Minority Oversampling Technique (SMOTE) to have more balanced data. The purpose of this study was to compare the performance of SMOTE in the random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination data. The study used the National Socio-Economic Survey data in 2017 for Aceh province with 2264 cases and 14 variables. The results show that the application of SMOTE in the random forest and naive Bayes classifier improves the accuracy of identification of Hepatitis-B non-vaccination status by 30.08% and 26.09%, respectively, compared to non-SMOTE. Random forest with SMOTE is the best model for classification HB vaccination status. The most important factors that influence the Hepatitis-B vaccination status of Aceh province are the mother's last education, mother's occupation, father's occupation, father's previous education, and the number of health facilities.

1. Introduction
The government of Indonesia has established an immunization program that is mandatory for children and applied according to the vaccine type and the specified schedule in the immunization guidelines. Immunization programs for infants include the Hepatitis-B (HB) vaccination given at birth and a maximum of 7 days after the birth [1]. The HB vaccination is essential because the plausibility of mother-to-child transmission of hepatitis is about 90% to 95%. The HB vaccination, amongst others, is intended to prevent this transmission.

Based on the National Socio-Economic Survey in 2017, the percentage of children aged 0-59 months who received HB vaccination in Aceh Province was 59.92% on average, the lowest than other provinces [2]. It is essential to predict the HB vaccination status accurately in this province to protect children from contracting Hepatitis-B and performing the immunization program effectively. Based on [3], the influencing factors of completeness of primary immunization of infants consist of the mother's education level, mother's knowledge level, mother's employment status, and family support.

The alternative approach to accurately predicting the classification of HB vaccination status is using a machine learning approach such as random forest and naive Bayes classifier. The random forest method combines a decision tree that forms each decision tree using several random samples from the data [4]. The random forest performs better than the single tree classification method with a higher AUC (Area Under Curve) value criteria [5,6]. On the other hand, the naive Bayes method is a probabilistic classification approach that calculates the probability by adding the frequency and value...
combinations. The naive Bayes method has similar capabilities to decision trees and neural networks [7].

The HB vaccination data are imbalanced, with the majority as a vaccination-status and minority as non-vaccination status. In such cases, standard classifiers misclassify the minority class and are biased towards the majority class by predicting the overall accuracy [8]. To increase the accurate prediction of a classifier, we consider the Synthetic Minority Oversampling Technique (SMOTE) to have more balanced data [9,10].

This study compared random forest and naive Bayes classifier performance on an imbalanced HB vaccination dataset. We then investigated the impact of SMOTE on the performance of these classifiers.

2. Methodology

2.1. Data

The data used in this study based on the National Socio-Economic Survey in 2017 by the Central Statistics Agency [2]. The observation unit was children aged 0-59 months of Aceh Province with 2264 observations. The response variable was Hepatitis B vaccination status and 14 explanatory variables, including household conditions and availability of health facilities.

2.2. Data Analysis Procedure

The procedure of data analysis in this study are as follows:

- Conducting descriptive statistics to determine the characteristics of the explanatory variables and determine the distribution of the percentage of children who received the Hepatitis B vaccination in every district and city in Aceh Province.
- Handling imbalanced data on vaccination status variables with SMOTE as follows:
  - The number of k-nearest neighbors used is 5. The generating data comes from the five nearest minority data.
  - The oversampling value is 400%, meaning that minority data extended four times.
  - Calculating the distance between minority class data with VDM values. The formula for the Value Difference Metric (VDM) distance is as follows [11]:

\[
\Delta(V, W) = \sum_{r=1}^{p} \delta(V_r, W_r)
\]

where:

- \(\Delta(V, W)\) : observation distance V and W
- \(p\): the number of explanatory variables
- \(\delta(V_r, W_r)\) : The distance between observations V and W for each explanatory variable r, i.e.,

\[
\delta(V_r, W_r) = \sum_{j=1}^{q} \left| \frac{C_{v_{rj}}}{C_v} - \frac{C_{w_{rj}}}{C_w} \right|^b
\]

where

- \(C_v\) : the total occurrences of the V category
- \(C_w\) : the total occurrences of the W category
- \(C_{v_{rj}}\) : the number of V categories on the r explanatory variable in the j response variable
- \(C_{w_{rj}}\) : the number of W categories on the r explanatory variable in the j response variable
- \(b\) : constant, usually 1
- \(q\) : the number of type in the r explanatory variable
Randomly selects one minority class data and determines its five nearest neighbors based on the VDM distance.

The latest data formed is minority data with synthetic data.

Classifying the HB vaccination status using the random forest method as follows:
- Setting parameter tuning with 10-fold cross-validation to determine the optimum combination of the number of variables in each tree (m) and the number of trees (t). We consider the number of trees (t) as 50, 100, 200, 500, and 1000. The number of variables in each tree (m) is \( \frac{1}{2}\sqrt{p} = 2, \sqrt{p} = 4, \text{ and } 2\sqrt{p}=8 \) where \( p \) is the number of explanatory variables.
- It is choosing the optimum combination of \( m \) and \( k \) that produces the highest average AUC.
- We are performing modeling with selected \( m \) and \( k \) with SMOTE and without SMOTE.
- We evaluate the results of the random forest method with SMOTE and without SMOTE using 10-fold cross-validation.

Classifying the HB vaccination status with the naive Bayes method as follows:
- Setting prior, posterior, and evidence values for each condition of the data group, with a Laplace correction of 1 to predict the status of Hepatitis B vaccination on SMOTE data and without SMOTE. We use the Laplace correction at Bayes odds as follows [12]:

\[
P(H|X) = \frac{P(H)P(X|H) + 1}{P(X) + |c|}
\]

where
- \( X \): a data cluster whose class is unknown
- \( H \): hypothesized class opportunities
- \( P(H|X) \): probability \( H \) based on the condition of data set \( X \)
- \( P(X|H) \): probability \( X \) based on condition \( H \)
- \( P(H) \): probability \( H \)
- \( P(X) \): probability \( X \)
- \( |c| \): many possible categories in variable \( X \)
- We are evaluating the results of the naive Bayes SMOTE method and without SMOTE with 10-fold cross-validation.
- We compare the four classification methods, i.e., the random forest and naive Bayes classifier with SMOTE and without SMOTE using a confusion matrix. The criteria are accuracy, sensitivity, specificity, AUC, and G-mean values.
- Identifying the critical factors with the Mean Decrease Gini (MDG). The MDG value with \( p \) explanatory variables with \( r = 1, 2, ..., p \) is expressed as follows [13]:

\[
MDG_r = \frac{1}{t} \sum_{z} [d(r, z)I(r, z)]
\]

where
- \( d(r, z) \): decrease in the heterogeneity index of the \( x_r \) variable at node \( z \)
- \( I(r, z) \): the indicator function is equal to 1 if the variable \( x_r \) is used at node \( z \) and is 0 otherwise
- \( t \): the number of trees in the random forest
3. Result and Discussion

3.1. Descriptive Statistics
The data observation of 2264 children aged 0-59 months consists of 1889 children with Hepatitis B vaccination status and 375 non-HB vaccination status. Figure 1 shows the percentage.

![Figure 1](image1.png)

**Figure 1.** Percentage of Hepatitis B vaccination status in Aceh Province.

Figure 2 presents the percentage of children aged 0-59 months with non-HB vaccination status for each district and city in Aceh Province. The Subussalam district of Aceh Province is the most significant non-HB vaccination status with 31.68% cases. On the other hand, the Bener Meriah district has vaccinated all children.

![Figure 2](image2.png)

**Figure 2.** Distribution of percentage of children with non-HB vaccination status in Aceh province.

Figure 3 presents the distribution of Hepatitis B vaccination status based on the children's age categories. The 0-12 month age category was the most considerable non-vaccine status than other age categories requiring serious attention. Meanwhile, the minor proportion was on the 37-48 months age category.
Based on data, the children with non-vaccination status were predominantly in rural areas. Also, the children with non-HB vaccination status were mainly in households who live in a rental house, do not have a vehicle, are either two-wheeled or four-wheeled, and have monthly expenditure in the low category, i.e., below two million rupiahs, as much as 21.73%. Additionally, the higher percentage of children of non-HB vaccination status observed on mothers with elementary school education (SD), who cannot access the internet, mother's age is less than twenty years, and who works as a laborer as well as fathers with 21-30 years old. The number of health facilities less than five show a similar pattern.

3.2. Imbalanced Data Handling
Using SMOTE technique, as much as 400% oversampling of the total minority classes, the number of data for the minority category increased from 338 to 1690 observations. Table 1 and figure 4 present the Hepatitis-B vaccination status category’s observations and proportion before and after SMOTE. The data observation after balancing with the SMOTE method shows a quite balanced classification.

| Category             | Original Data | SMOTE Data |
|----------------------|---------------|------------|
| Vaccination Hepatitis-B | 1700 (0.83)   | 1700 (0.50) |
| Non Vaccination Hepatitis-B | 338 (0.17)   | 1690 (0.50) |
| Total                | 2038          | 3012       |

Figure 3. Percentage of hepatitis B vaccination status by age group.

Figure 4. Percentage of data before and after SMOTE.
3.3. Classification of Hepatitis B Vaccination Status by Random Forest

Selection of the optimum combination of m and t is performed based on the average AUC value using 10-fold cross-validation. A high average AUC value indicates the best performance of the random forest method in classification predictions. Figure 5 presents the average AUC value for each change in m for each number of t trees.

![Figure 5](image)

**Figure 5.** Average AUC of random forest performance for each m change.

Based on figure 5, the average AUC for m = 2 with all combinations of t number of trees shows the lowest average AUC compared to other m indicate that m = 2 is not optimal as a parameter in this study. The AUC value at m = 8 for all t trees yields the highest average AUC value than m = 2 and m = 4. Figure 6 presents the average AUC for each change in t. The number of trees t = 100 had the highest mean AUC compared to other t values. Therefore, the number of explanatory variables m = 8 and the number of trees t = 100 is the optimum combination for the random forest method.

![Figure 6](image)

**Figure 6.** Average AUC of random forest performance for each t change.

3.4. Comparison of the Random Forest and Naive Bayes Classification Methods

Table 2 present the values of accuracy, sensitivity, specificity, G-mean, and AUC for random forest and naive Bayes methods for imbalanced and SMOTE balanced data. For imbalanced data, the accuracy of the classification of hepatitis B vaccination status using the random forest method is higher than the Naive Bayes approach. The accuracies are 82.19% and 81.92%, respectively. However, both algorithms show a low sensitivity value which implies that the estimation for the minority classes (non-HB vaccination status) is inaccurate. The sensitivity value for random forest and naive Bayes methods are 16.19% and 4.22%, respectively. The G-mean value for both procedures is below 40%. Additionally, the AUC value of the random forest is more significant than naive Bayes. It indicates that the random forest method performs better than the Naive Bayes method.
Table 2. Comparison of random forest and naive Bayes without SMOTE and with SMOTE.

| Criteria    | Random forest | Random Forest SMOTE | Naive Bayes | Naive Bayes SMOTE |
|-------------|---------------|---------------------|-------------|------------------|
| Accuracy    | 82.19%        | 73.22%              | 81.92%      | 62.87%           |
| Sensitivity | 16.19%        | 46.27%              | 4.22%       | 30.31%           |
| Specificity | 95.28%        | 78.59%              | 97.39%      | 69.36%           |
| G mean      | 38.71%        | 60.19%              | 17.62%      | 45.65%           |
| AUC         | 0.556         | 0.614               | 0.508       | 0.498            |

For SMOTE balanced data, the G-mean value for random forest and naive Bayes methods increase to 60.19% and 45.65%, respectively, compared to without SMOTE. The accuracy of both algorithms slightly decreases. However, the application of SMOTE increases the sensitivity value of both methods. The random forest and naïve Bayes methods' sensitivity value increases by 30.08% and 26.09%, respectively.

The best classification method is chosen by looking at the ability of the model to predict both classes, particularly the minor status. Thus, the SMOTE random forest is more satisfying in estimating the position of Hepatitis B vaccination in Aceh Province. The SMOTE random forest has a G-mean value of 60.19%, 46.27% sensitivity, 78.59% specificity, 73.22% accuracy, and an Area Under Curve (AUC) value of 0.614. We will use the SMOTE random forest method to estimate the most important factors influencing the Hepatitis B vaccination status.

3.5. The Important Factors in Random Forest

Figure 7 shows the Mean Decrease Gini (MDG) value of all explanatory variables in a random forest. We observed that the most critical factors influencing the Hepatitis B vaccination status are the mother's latest education, mother's occupation, father's occupation, father's previous education, and the number of health facilities around the household's residence.

![Figure 7. Mean decrease Gini in SMOTE random forest.](image)

4. Conclusion

For imbalanced Hepatitis B vaccination data, the random forest classification method has better accuracy, sensitivity, G-mean, and AUC values than the Naive Bayes approach. However, the sensitivity for the accurate estimation for the minority classes (non-Hepatitis B vaccination) is minimal. The handling of imbalanced data with the Synthetic Minority Over Sampling Technique (SMOTE) could improve the sensitivity, G-mean, and AUC values. The SMOTE technique increases the sensitivity in a
random forest and naive Bayes methods by 30.08% and 26.09%, respectively. The random forest method with SMOTE is the best classification method for predicting Hepatitis B vaccination status in Aceh province. The most critical factors are the mother's education, mother's occupation, father's job, father's education, and the number of health facilities.

References
[1] Kemenkes 2017 Peraturan Menteri Kesehatan Republik Indonesia Nomor 12 Tahun 2017 tentang Penyelenggaraan Imunisasi Kemenkes Kementrian Jakarta (ID)
[2] BPS 2017 Statistik Kesejahteraan Rakyat Badan Pusat Statistik Jakarta (ID)
[3] Isriyati E 2011 Faktor-Faktor yang Berhubungan dengan Kelengkapan Imunisasi Dasar pada Bayi di Desa Kumpulrejo Kecamatan Argomulyo Kota Salatiga [skripsi] Semarang (ID): Universitas Negeri Semarang
[4] Breiman L 2001 Random Forests Machine Learning 45(1):5-32
[5] Iain B and Christophe M 2012 An experimental comparison of classification algorithms for imbalanced credit scoring data sets Expert Systems with Applications 39 3446–3453
[6] Wang Y, Zhang Y, Lu Y, and Yu X 2020 A Comparative Assessment of Credit Risk Model Based on Machine Learning, A case study of bank loan data Procedia Computer Science 174 141–149
[7] Daniela X, Christopher J H, and Roger G S. 2009 Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages IJCSI International Journal of Computer Science Issues 4(1) 16-23
[8] Anis M and Ali M 2017 Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets European Scientific Journal, ESJ, 13(33), 340.
[9] Chawla N V, Bowyer K W, Hall L O, and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique Journal of Artificial Intelligence Research 16:321–357
[10] Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, and Khalili D 2016 The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes Med Decis Making 36(1) 137-144
[11] Cost S and Salzberg S 1993 A weighted nearest neighbor algorithm for learning with symbolic features Machine Learning 10: 57-58
[12] Bindu M S and Cherian V 2017 Heart disease prediction using naive Bayes algorithm and Laplace smoothing technique Int. J. Comput. Sci. 5(2) 68–73
[13] Sandri M and Zuccolotto P 2006 Variable Selection Using Random Forest Dipartimento Metodi Quantitativi Università di Brescia