Performance Analysis of Tree-Based Models for Classifying Complete Basic Childhood Immunization in West Java

Windi Pangesti[1], Mega Maulina [1], Hazelita Dwi Rahmasari [1*], Bagus Sartono [1], Budi Susetyo[1], Aulia Rizki Firdawanti [1], and Gerry Alfa Dito [1]
[1]School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia
*Corresponding author: h042002elitadwi@apps.ipb.ac.id

Received: 30 January 2026
Revised: 30 May 2026
Accepted: 2 June 2026

ABSTRACT **Complete basic immunization is a key public health indicator, and disparities in coverage remain a major concern in West Java. The 2024 Universal Child Immunization (UCI) rate in West Java reached only 77.47%, declining from 2023 and reflecting persistent disparities in access to immunization services, particularly between urban and rural areas. This study aims to identify the determinants of complete basic immunization among children in West Java using the SUSENAS 2024 survey data (n = 4,672). Three tree-based classification algorithms, namely CART, Random Forest, and LightGBM, were applied, with class imbalance addressed using SMOTE, Tomek Links, and the SMOTE-Tomek Links hybrid method. Model performance was evaluated using balanced accuracy, and threshold adjustment using Youden's Index. Random Forest combined with SMOTE was retained as the main model, achieving an accuracy of 60.9%, sensitivity of 71.2%, specificity of 50.2%, and balanced accuracy of 60.7% after threshold adjustment. Global feature importance results indicate that household spending category, KIA book ownership, maternal age at first birth, and maternal education were the strongest predictors of complete basic immunization. SHAP analysis reveals contrasting patterns between urban and rural areas. In urban areas, KIA book ownership, maternal education, household head education, and household spending category contributed positively to complete immunization, whereas in rural areas, rural residence and low maternal education contributed negatively. These findings underscore the importance of geographically targeted immunization strategies to support equitable access across urban and rural communities in West Java.**

Keywords– **Basic Immunization, Class Imbalance, Random Forest, SHAP**

I. INTRODUCTION

Immunization is an important preventive intervention for building active immunity against preventable diseases and is a key indicator in efforts to improve public health. In Indonesia, basic immunization is the right of every child, as stipulated in Law Number 17 of 2023 concerning Health, Article 44, Paragraph (2), which states that every infant and child has the right to receive basic immunization in accordance with applicable regulations to prevent diseases that can be prevented through immunization[1]. In addition to providing direct protection against various diseases, immunization plays an important role in national health development and contributes to the attainment of the *Sustainable Development Goals* (SDGs), particularly the third goal of ensuring healthy lives and promoting well-being for all. [2].

The Complete Basic Immunization Program aims to provide continuous immunization services for all children in Indonesia, including initial immunizations such as BCG, Hepatitis B, Polio, DPT-HB-Hib, and Measles–Rubella (MR), as well as follow-up immunizations at 18–24 months of age [3]. Although this program has been implemented nationally, immunization coverage still varies across regions. One of the performance indicators is Universal Child Immunization (UCI) coverage, defined as the percentage of villages or subdistricts with a minimum basic immunization coverage of 80%. In 2024, West Java's UCI coverage was 77.47%, a decline of 4.58 percentage points from 2023 and lower than the 2022 achievement of 82.05%. [4]. This decline indicates a gap in access to immunization services, particularly between urban and rural areas. Therefore, an analysis of the factors affecting complete basic immunization in West Java is needed to assess the effectiveness of the program and to formulate more targeted interventions.

In analyzing the factors that influence this gap, this study employs three tree-based classifiers, namely Classification and Regression Tree (CART), Random Forest (RF), and Light Gradient Boosting Machine (LightGBM). CART was selected as the base model because of its ability to represent nonlinear relationships in the form of easily interpretable decision trees. Random Forest was used as a bagging-based ensemble method to improve stability and reduce prediction variance through the construction of multiple random decision trees [5]. Meanwhile, LightGBM, as a *gradient boosting* method, was developed to accelerate the model training process and reduce memory usage through histogram-based tree learning techniques [6]. The simultaneous use of these three methods enables the evaluation and comparison of model performance across a single-tree approach, a bagging ensemble, and a boosting ensemble, thereby identifying the most suitable approach for handling the complex characteristics of immunization data while maintaining adequate interpretability.

The use of these *tree-based* methods is also supported by various previous studies. Perdana et al. [7] conducted research on predicting stunting in children using the random forest algorithm and achieved an accuracy of 97%. Khairunnisa et al. [8] compared the RF and Double Random Forest (DRF) algorithms in terms of model interpretability using the association rule approach. Meanwhile, Ramadanti et al. [9] examined diabetes classification and showed that

LightGBM with *hyperparameter* optimization through GridSearchCV and SMOTE application produced the best performance with an accuracy of 84%.

Immunization data often exhibit class imbalance, in which the number of children with complete immunization is substantially greater than those with incomplete immunization. This condition can reduce model accuracy and lead to prediction results that are biased toward the majority class [10]. To address this issue, this study applied three data balancing techniques, namely SMOTE, Tomek Links, and the hybrid SMOTE–Tomek Links method. SMOTE is used to improve the representation of the minority class by generating synthetic samples through interpolation between nearest neighbors, while Tomek Links is applied to remove pairs of observations from different classes that lie on the decision boundary in order to reduce classification ambiguity [11] The combination of these two techniques is expected to produce a more proportional data distribution, thereby improving the reliability of the model in identifying children at risk of not receiving complete basic immunization. These three techniques were applied because each employs a different mechanism for handling class imbalance and decision boundaries. This comparison was conducted to empirically evaluate which approach was most effective—adding synthetic samples, removing ambiguous observations, or a combination of both—in order to determine which provided the most optimal model performance on immunization data.

In addition to requiring models with strong predictive performance, model interpretability is crucial in the context of public health so that policymakers can understand the basis for predictions. However, predictions generated by machine learning models often make it difficult to explain the specific relationships between changes in input variables and the resulting outputs [12]. To improve interpretability, this study employs Shapley Additive Explanations (SHAP) to measure the contribution of each predictor to the prediction results in a consistent and transparent manner [12]. This approach is similar to feature importance plots but provides additional insight by indicating whether each predictor has a positive or negative relationship with the response variable [13], thereby bridging the complexity of the model with the need for interpretability in the context of public health policy.

Therefore, this study aims to compare the performance of several decision tree–based classification algorithms, namely Classification and Regression Tree (CART), Random Forest, and Light Gradient Boosting Machine (LightGBM), in predicting the complete basic immunization among children status in West Java Province. In addition to comparing the predictive capabilities of each model, this study also seeks to identify the determinants of basic immunization completion through an analysis of the contributions of independent variables using a model interpretability approach, thereby enabling a clearer understanding of the role of each variable. The evaluation was conducted using both balanced and unbalanced data to assess the effectiveness of class imbalance handling techniques. This approach is expected to generate contextual insights that support the formulation of more targeted immunization interventions.

II. LITERATURE REVIEW

A. Classification and Regression Trees (CART)

CART is a decision tree method with a relatively simple concept, yet it demonstrates strong performance for both classification and regression tasks. This method operates by recursively constructing a binary tree structure, in which each node in the decision tree is split into two sub-nodes at each stage of the partitioning process [14]. The splitting process is performed by selecting the split that minimizes impurity, using measures such as the Gini index, information gain, the Twoing criterion, and entropy at each node. The decision tree construction process is stopped when all observations within a node exhibit homogeneous characteristics or when the maximum tree depth reaches a specified limit. CART applies pruning to reduce overfitting by balancing tree complexity. This method is highly sensitive to data variations, as changes in the training or test data can lead to different decision tree structures [15].

B. Random Forest

Random Forest is an ensemble-based machine learning method that uses decision trees as its base learners. This method was developed from the CART approach by employing bootstrap aggregating (bagging) techniques and random feature selection during the construction of each tree [16]. Unlike CART, Random Forest does not apply pruning procedures to the individual trees [17]. This method demonstrates improved performance, as indicated by its lower generalization error and reduced correlations among individual decision trees. Random Forest is an effective predictive method and can prevent overfitting even when the number of trees is large, as it leverages the Law of Large Numbers. In addition, the application of random feature selection during tree construction enables Random Forest to achieve high accuracy and robust performance in both classification and regression tasks [18].

C. Light Gradient Boosting Machine (LightGBM)

LightGBM is a Gradient Boosting Decision Tree (GBDT) framework that optimizes the decision tree algorithm by applying two main approaches, namely gradient-based one-side sampling (GOSS), which improves sample selection efficiency, and exclusive feature bundling (EFB), which combines mutually exclusive features to reduce computational complexity. This method utilizes a histogram-based decision tree approach, accompanied by a node growth strategy controlled through depth limitation (leaf-wise with depth limitation) to manage model complexity and prevent overfitting, as well as multithreaded computational optimization. This approach effectively reduces memory usage compared to other boosting methods while enabling large-scale data processing and delivering faster performance with lower error rates and more accurate predictions [19].

D. Shapley Additive Explanations (SHAP)

SHAP is an interpretability technique based on a game theory framework (Shapley values) that describes the contribution of each feature to classification predictions, both at the overall model level (global) and for individual observations (local). This method provides explanations for model predictions through an additive approach by comparing the model output for a given observation with a baseline value, defined as the average prediction when feature information is unknown. The Shapley value of a feature represents its contribution to shifting the prediction from the baseline to the final value produced by the model [20]. The following presents the calculation of the Shapley value, as shown in Equation (1):

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \big(v(S \cup \{i\}) - v(S)\big) \tag{1}$$

Explanation:

ϕ_i	=	Shapley value for the i-th variable
M	=	The set of all features used in the model (total number of predictors)
S	=	A subset of predictors that does not include the i-th feature
$v(S)$	=	The model prediction value when only the variables in subset S are used
$v(S \cup \{i\})$	=	The model prediction value when the i-th variable is added to subset S
$\frac{\lvert S\rvert!(M-\lvert S\rvert-1)!}{M!}$	=	Weighting factor that reflects the probability of subset S being selected in the Shapley value calculation

There are several types of SHAP methods. In this study, TreeSHAP was employed, which is an algorithm specifically designed to efficiently and rapidly compute SHAP values for decision tree–based ensemble models [21].

E. Resampling Methods

The resampling method is an approach used to address class imbalance by modifying the composition of the training data so that the model learns more evenly from the majority and minority classes [22]. There are three main categories of resampling: oversampling, undersampling, and hybrid sampling. Oversampling operates by generating synthetic data for the minority class, undersampling reduces the number of observations in the majority class, and hybrid sampling combines oversampling and undersampling to achieve a more balanced class distribution [23]. In this study, the oversampling method applied was SMOTE, the undersampling method was Tomek Links, and the hybrid sampling method was SMOTE combined with Tomek Links.

F. Performance Evaluation Metrics

The classification model was evaluated to assess its ability to accurately predict the target class. One metric used to evaluate classification performance is the confusion matrix, which is a table that compares the model's predicted results with the actual values, thereby providing an overview of the number of correct predictions and prediction errors.

Table 1 Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

From the confusion matrix, various evaluation metrics can be derived to assess the model's performance more comprehensively, particularly for the evaluation of imbalanced data:

Table 2 Evaluation Metrics

Metrics	**Formula**	**Description**
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	The proportion of correct predictions out of all predictions.
Sensitivity	$\frac{TP}{TP + FN}$	The ability of the model to predict all actual positive cases.
Specificity	$\frac{TN}{TN + FP}$	The ability of the model to predict all actual negative cases.

Balanced Accuracy	$\frac{Sensitivity + Specificity}{2}$	The average of sensitivity and specificity, making it fairer for unbalanced data.

III. METHODOLOGY

A. Data and Research Variables

This study adopts a quantitative approach using SUSENAS 2024 data collected by BPS for West Java Province. The survey data focus on the basic immunization status of children, measured based on whether children received all required basic vaccines in each household in West Java Province. The target population in this study consists of children aged 9 to 59 months. The basic immunizations used as research indicators comprise five vaccines: Hepatitis B, BCG, Polio, DPT, and Measles–Rubella(MR). Children who received all five vaccines were classified as having complete immunization, while those who did not receive one or more vaccines were classified as having incomplete immunization. The explanatory variables were grouped into four categories: maternal characteristics, household head characteristics, residential area, and household conditions. The units and variables used in this study are presented in the following Table 3.

Table 3 Predictor variables

Factor	Variable	Type
Mother's Characteristics	Mother's Age	Numeric
	Mother's Age at First Birth	Numeric
	KIA Book Ownership	Nominal
	Mother's Health Insurance	Nominal
	Mother's Education Level	Ordinal
	Mother's Employment Status	Nominal
	Mother's Access to Technology	Nominal
	Place of Birth	Nominal
	Birth Attendant	Nominal
Household Head's Characteristics	Age of Head of Household	Numeric
	Household Head's Education	Ordinal
	Household Head's Employment Status	Nominal
	Household Head's Health Insurance	Nominal
Household Residence Area	Residential Area Type (urban/rural)	Nominal
Household Conditions	Home Ownership Status	Nominal
	Household Size (person)	Numeric
	Poverty Status	Nominal
	Presence of Both Parents	Nominal
	Social Assistance Status	Nominal
	PKH Recipient Status	Numeric
	Household Spending Category	Ordinal

B. Data Analysis Procedure

The data analysis procedure in this study was carried out through several interrelated stages, which are described as follows.

1. Data Pre-processing
   The initial stage focused on preparing the data to ensure its suitability for the modeling process. This stage included data cleaning and checking for missing values to maintain data quality.
2. Data Splitting
   The dataset was divided into two subsets, namely training data and test data. This division aimed to prevent data leakage and to ensure that model performance was evaluated objectively using data that were not used during the training process.
3. Imbalanced Data Handling
   The class proportions in the response variable were balanced due to the presence of class imbalance in the data. Three approaches were applied: oversampling using SMOTE, undersampling using Tomek Links, and a hybrid approach combining SMOTE and Tomek Links. All resampling techniques were applied only to the training data within the cross-validation process to preserve the validity of the evaluation metrics on the test data.
4. Classification Modeling
   Classification modeling was performed using three algorithms, namely CART, Random Forest, and LightGBM. Each model was built and evaluated under four training data scenarios: (1) training data without class balancing, (2) SMOTE-balanced data, (3) Tomek Links–processed data, and (4) hybrid SMOTE–Tomek Links–processed data. The optimal hyperparameter settings for each model were determined through a hyperparameter tuning process using a Bayesian Optimization approach with a 5-fold cross-validation scheme. The hyperparameter search space was defined based on the specified value ranges presented in Table 4.

Table 4 Hyperparameter Domain

Hyperparameter	Domain
trees	100 - 2000
mtry	1 - 20
min_n	2 - 40
tree_depth	2 - 20
learning_rate	0.0001 – 0.5

5. Model Evaluation
   Model performance was evaluated using independent test data. The evaluation metrics included accuracy, sensitivity (recall for the positive class), specificity (recall for the negative class), and balanced accuracy, which is more appropriate for imbalanced data. This combination of metrics provides a more comprehensive assessment of the model's ability to accurately classify both classes.
6. Determining the Best Model
   The best model was selected based on the evaluation metric most relevant to the research objectives. In this study, the model with the highest balanced accuracy was identified as the best-performing model, as this metric is more robust to class imbalance and reflects performance equally across both classes.
7. SHAP Analysis of the Best Model
   The final stage involved interpretability analysis using the SHAP method. This approach was applied to identify the most influential features in the best-performing model. SHAP provides transparent interpretations and enables a deeper understanding of each variable's contribution to the model's predictions.

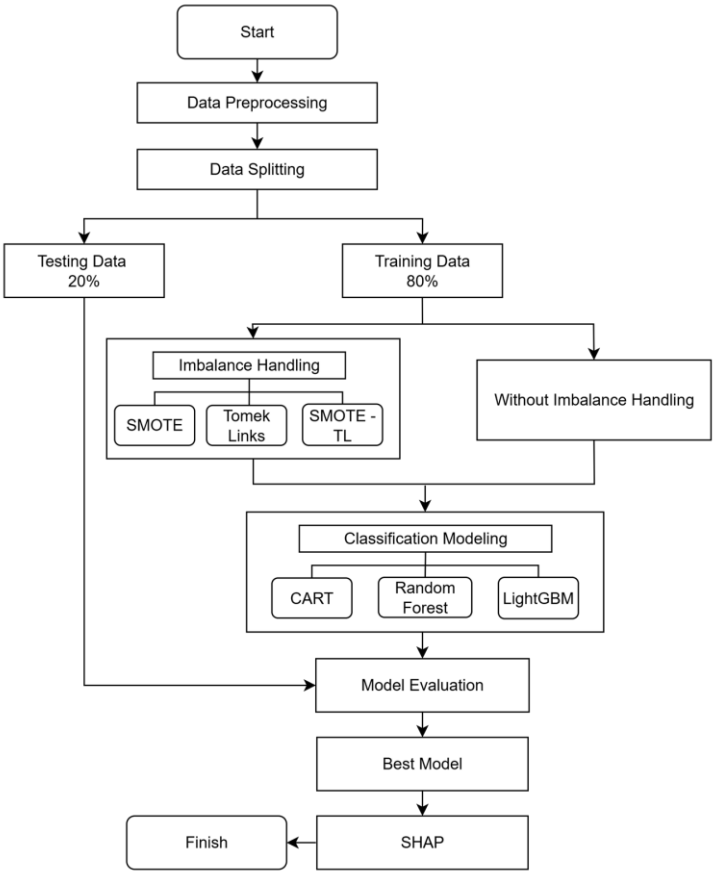

Figure 1 Data Analysis Procedure

IV. RESULTS AND DISCUSSIONS

A. *Data Exploration*

As an initial stage of the analysis, the distribution of basic immunization status among children in West Java Province was explored using SUSENAS 2024 data to obtain an overview of patterns of complete basic immunization within the study population. Presenting this distribution is important for identifying the proportions of children with complete and incomplete immunization, thereby providing an initial context for subsequent analyses of the determinants that may influence immunization coverage.

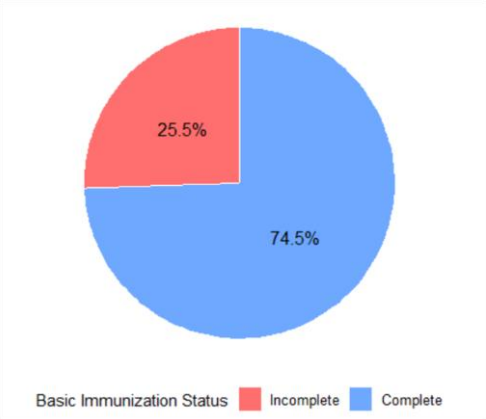

Figure 2 Distribution of basic immunization status

Figure 2 shows the distribution of complete basic immunization status among 4,672 children. Most children were classified in the complete immunization category, accounting for 74.5% (3,546 children), while the remaining 25.5% (1,216 children) were classified as having incomplete immunization. These results indicate that the proportion of children with incomplete immunization in West Java remains substantial, potentially increasing vulnerability to diseases that can be prevented through immunization. This imbalance in class proportions is an important consideration in the classification modeling process, particularly with respect to handling class imbalance in the subsequent stage of analysis. The results of the exploratory analysis of categorical variables for each factor are presented in Table 5.

Table 5 Immunization Status Distribution by Maternal and Household Characteristics

Variable	Complete n (%)	Incomplete n (%)	Variable	Complete n (%)	Incomplete n (%)
KIA Book Ownership			**Household Head's Health Insurance**		
Yes	3.382 (77.8)	966 (22.2)	Yes	2.678 (76.6)	818 (23.4)
No	164 (39.6)	250 (60.4)	No	868 (68.6)	398 (31.4)
Mother's Health Insurance			**Household Head's Education**		
Yes	2.760 (76.5)	850 (23.5)	No Schooling	54 (52.9)	48 (47.1)
No	786 (68.2)	366 (31.8)	Primary School	1.029 (69.8)	445 (30.2)
Mother's Education			Junior High School	793 (72)	308 (28)
No Schooling	35 (57.4)	26 (42.6)	Senior High School	1.238 (78.7)	335 (21.3)
Primary School	802 (68)	377 (32)	Higher Education	432 (84.4)	80 (15.6)
Junior High School	1.085 (71.6)	431 (28.4)	**Household Head's Employment Status**		
Senior High School	1.151 (79.5)	297 (20.5)	Working	3.410 (74.6)	1.162 (25.4)
Higher Education	473 (84.8)	85 (15.2)	Not Working	136 (71.6)	54 (28.4)
Mother's Employment Status			**Residential Area Type**		
Working	1.306 (76.3)	405 (23.7)	Urban	2.426 (77.3)	714 (22.7)
Not Working	2.240 (73.4)	811 (26.6)	Rural	1.120 (69.1)	502 (30.9)
Mother's Access to Technology			**Home Ownership Status**		
Yes	3.352 (75.1)	1.110 (24.9)	Yes	2.607 (73.6)	936 (26.4)
No	194 (64.7)	106 (35.3)	No	939 (77)	280 (23)
Place of Giving Birth			**Poverty Status**		
Health Facility	1.019 (73)	376 (27)	Yes	302 (62.5)	181 (37.5)
Non-Health Facility	123 (55.7)	98 (44.3)	No	3.244 (75.8)	1.035 (24.2)
Unknown	2.404 (76.4)	742 (23.6)	**Presence of Both Parents**		
Birth Attendant			Complete	3.463 (74.6)	1.179 (25.4)
Health Personnel	1.092 (72)	425 (28)	Incomplete	83 (69.2)	37 (30.8)
Non-Health Personnel	50 (50.5)	49 (49.5)	**Social Assistance Status**		
Unknown	2.404 (76.4)	742 (23.6)	Yes	1.485 (72.4)	566 (27.6)
Household Spending Category			No	2.061 (76)	650 (24)
Low	1.285 (67.4)	622 (32.6)	**PKH Recipient Status**		
Middle	1.456 (76.4)	449 (23.6)	Yes	465 (73.9)	164 (26.1)
High	805 (84.7)	145 (15.3)	No	3.081 (74.5)	1.052 (25.5)

The results presented in Table 5 indicate variations in the proportion of basic immunization coverage among children based on KIA Book Ownership, Mother's Education, Household Head's Education, Residential Area Type, and Household Spending Category. Children who had a KIA book showed a substantially higher proportion of complete basic immunization coverage (77.8%) compared to those without a KIA book (39.6%). With respect to Mother's Education, mothers with Higher Education exhibited a higher proportion of complete basic immunization coverage (84.8%) than mothers with No Schooling (57.4%). Regarding Household Head's Education, households headed by individuals with Higher Education were associated with a higher proportion of complete basic immunization coverage (84.4%) compared to those headed by individuals with No Schooling (52.9%). In terms of Residential Area Type, children residing in Urban areas showed a higher proportion of complete basic immunization coverage (77.3%) compared to those

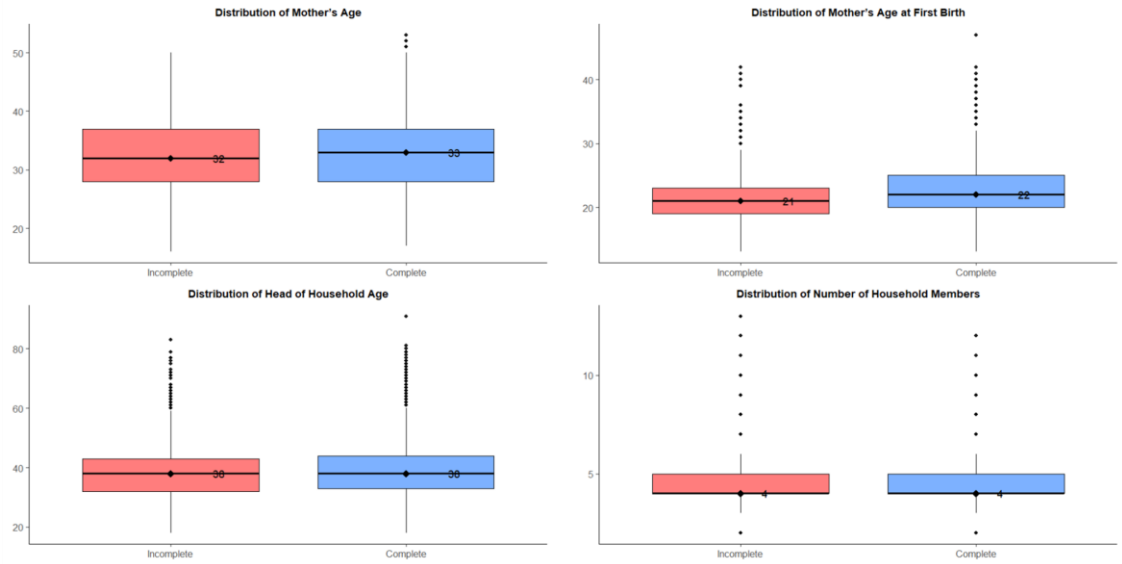

Figure 3 Numeric Distribution of basic immunization status

living in Rural areas (69.1%). Household economic conditions, as reflected by Household Spending Category, also influenced immunization coverage, with children in the High spending group exhibiting a higher proportion of complete basic immunization coverage (84.7%) than those in the Low spending group (67.4%).

The age characteristics of mothers and household heads did not show substantial differences between the complete and incomplete immunization groups. The average age of mothers was only slightly higher in the complete immunization group (33 years) than in the incomplete group (32 years). A similar pattern was observed for age at first birth, with mothers of completely immunized children having an average age of 22 years, compared to 21 years among those with incomplete immunization. The age of the household head was also relatively similar across both groups, with an average age of 38 years. Likewise, the number of household members did not differ markedly between the two groups, with an average of four members in each.

B. Data Preparation and Preprocessing

Data preparation and preprocessing were conducted to ensure data quality prior to the modeling stage. In the initial step, data cleaning was performed by checking for missing values. Missing values were identified for children who lacked information about their mothers in the household member structure. As this information constitutes an important attribute for the study, all children records without this information were removed to prevent potential bias in the analysis. After this process, a total of 4,672 children observations remained and were ready for analysis, consisting of 3,546 children (74.5%) in the complete immunization category and 1,216 children (25.5%) in the incomplete immunization category.

The dataset was subsequently divided into training and testing data using a stratified random splitting approach. This approach was applied to preserve the proportion of complete and incomplete immunization classes in both subsets. The training data were used for model development, hyperparameter tuning, and imbalance handling, while the testing data were kept separate and were used only for final model evaluation. The distribution of the response variable in the training and testing data is presented in Table 6.

Table 6 Distribution of response variable classes in training and testing data

Category	Data Training	Data Testing
Complete	2837	709
Incomplete	973	243
Total	3810	952

Based on Table 6, the training data consisted of 3,810 observations, comprising 2,837 children with complete immunization and 973 children with incomplete immunization. Meanwhile, the testing data consisted of 952 observations, with 709 children in the complete immunization category and 243 children in the incomplete immunization category. This distribution shows that the proportion of complete and incomplete immunization cases was relatively maintained in both subsets through the stratified random splitting procedure. However, the response variable still exhibited class imbalance, as the number of children with complete immunization was higher than those with incomplete immunization. To address this imbalance, three imbalance-handling techniques were incorporated into the modeling pipeline, namely SMOTE, Tomek Links, and the hybrid SMOTE–Tomek Links approach. During hyperparameter tuning, the training data were evaluated using a 5-fold cross-validation scheme. In each cross-validation iteration, the resampling procedure was applied only to the training fold, while the validation fold was kept in its original class distribution. After the best hyperparameter combination was obtained, the final model was trained according to each imbalance-handling scenario and subsequently evaluated using the independent testing data.

C. *Model Evaluation Results*

The data preparation process resulted in four variations of training data: without imbalance handling, oversampling using SMOTE, undersampling using Tomek Links, and the SMOTE–Tomek Links hybrid method. These four variations were used to evaluate the performance of the CART, Random Forest, and LightGBM models. The selection of the best model for each data variation was based on the balanced accuracy metric, as it is more representative under conditions of class imbalance. Table 7 presents the evaluation results of the three models across all data variations.

Table 7 Performance Evaluation of CART, Random Forest, and LightGBM

Model	Handling Method	Evaluation Metrics			
		Accuracy	Sensitivity	Specificity	Balanced Accuracy
CART	Without handling	0.763	0.962	0.181	0.571
	SMOTE	**0.592**	**0.595**	**0.584**	**0.590**
	Tomek Links	0.755	0.952	0.181	0.567
	SMOTE – Tomek	0.592	0.595	0.584	0.588
Random Forest	Without handling	0.769	0.968	0.189	0.578
	SMOTE	**0.703**	**0.807**	**0.399**	**0.603**
	Tomek Links	0.768	0.972	0.173	0.572
	SMOTE – Tomek	0.756	0.937	0.230	0.583
LightGBM	Without handling	0.764	0.975	0.148	0.561
	SMOTE	**0.739**	**0.891**	**0.296**	**0.594**
	Tomek Links	0.757	0.962	0.160	0.561
	SMOTE – Tomek	0.767	0.961	0.202	0.581

Before evaluating model performance across the four data variations, all models were first optimized through hyperparameter tuning using Bayesian Optimization with 5-fold cross-validation. In this approach, the training data were divided into five folds, the model was trained on four folds and validated on the remaining fold iteratively until each fold had been used once as the validation set. This procedure ensured that each model was evaluated across different data partitions, thereby reducing bias in hyperparameter selection and minimizing the risk of overfitting. Based on the tuning results, the CART model achieved an optimal configuration with a tree depth of 16, a minimum split of 22, and a complexity parameter of 0.03. The Random Forest model attained its best performance with an mtry value of 1, 667 trees, and a terminal node size of 18. Meanwhile, the LightGBM model showed an optimal configuration with 1,166 trees, a tree depth of 11, and a learning rate of 0.0001. These selected hyperparameter settings were then used for model performance evaluation.

The performance evaluation results in Table 7 show that Random Forest performed better than CART and LightGBM. The best model in the main evaluation was selected based on balanced accuracy, as this metric accounts for the model's ability to identify both classes more evenly. The highest performance was achieved by Random Forest with SMOTE, with a balanced accuracy of 0.603, accuracy of 0.703, sensitivity of 0.807, and specificity of 0.399. The high sensitivity indicates that the model was able to identify children with complete immunization reasonably well. However, the relatively low specificity suggests that the model was still limited in detecting children with incomplete immunization as the minority class. Among the imbalance-handling methods, SMOTE yielded the highest balanced accuracy within each model, although the performance differences across scenarios were relatively small. Therefore, the threshold adjustment analysis was focused on the SMOTE-based models to further examine whether modifying the classification threshold could improve the detection of children with incomplete immunization.

In binary classification, the default threshold of 0.50 may not always provide the best classification performance, particularly for imbalanced data where predicted probabilities can be influenced by the majority class [24]. Therefore, the optimal threshold was determined using Youden's Index, which balances sensitivity and specificity [25]. A higher Youden's Index indicates a better threshold for distinguishing both classes in a balanced manner. The threshold adjustment results for the SMOTE-based models are presented in Table 8.

Table 8 Threshold Adjustment Results for SMOTE-Based Models Using Youden's Index

Model	Threshold	Evaluation Metrics				Best YI
		Accuracy	Sensitivity	Specificity	Balanced Accuracy	
CART	0.37	0.592	0.595	0.584	0.590	0.180
Random Forest	**0.48**	**0.609**	**0.712**	**0.502**	**0.607**	**0.214**
LightGBM	0.49	0.686	0.772	0.436	0.604	0.208

Based on Table 8, Random Forest with SMOTE achieved the best performance after threshold adjustment compared with the other SMOTE-based models, with an optimal threshold of 0.48, a balanced accuracy of 0.607, and a Youden's Index of 0.214. Compared with the default threshold of 0.50, the threshold adjustment in Random Forest with SMOTE increased specificity from 0.399 to 0.502 and balanced accuracy from 0.603 to 0.607. The increase in specificity indicates that the model improved its ability to detect children with incomplete immunization.

However, this improvement was accompanied by a decrease in accuracy from 0.703 to 0.609 and sensitivity from 0.807 to 0.712. This finding indicates a trade-off, as lowering the threshold makes the model more responsive to the minority class but may also increase the likelihood that some children with complete immunization are classified as incomplete. Therefore, Random Forest with SMOTE was retained as the main model in this study.

D. Feature Importance Analysis of the Best Model

After obtaining the best model, the analysis proceeded by examining the global contribution of each variable in the Random Forest model built using the full dataset. This evaluation illustrates how variables are utilized by the model in forming classification decisions, while also identifying those that play a dominant role and those with more marginal contributions. These findings provide an overview of the key factors associated with complete basic immunization. A summary of the variables with the greatest influence is presented in Figure 4.

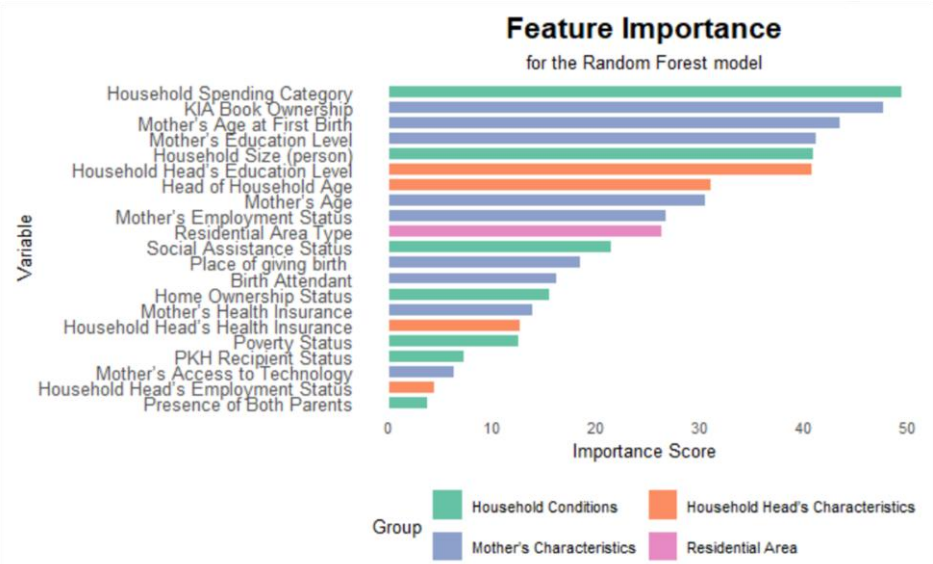

Figure 4 Feature importance of the Random Forest model

Figure 4 illustrates the contribution of each variable to the performance of the Random Forest model in classifying the basic immunization status of children in West Java Province. Household Spending Category emerged as the most influential factor, indicating that household socio-economic conditions play a crucial role in access to health services and adherence to the immunization schedule. The next most influential variable was KIA Book Ownership (Maternal and Child Health Record Book), which functions as a tool for recording maternal and child health information from pregnancy until the child reaches five years of age. Ownership of this book is often associated with parents' knowledge and awareness of health services, including compliance with basic immunization requirements.

Furthermore, Mother's Age at First Birth, Mother's Education Level, and Mother's Age also showed substantial contributions, suggesting that maternal knowledge, reproductive health readiness, and caregiving capacity are important determinants of complete immunization. Other variables, such as Household Head's Education, Number of Household Members, and Health Insurance Coverage, remained influential but contributed to a lesser extent. Overall, these findings confirm that variables closely related to maternal characteristics and household conditions are the dominant factors shaping the model's predictions of basic immunization completeness.

E. Model Interpretability Analysis Using SHAP

To gain a deeper understanding of the decision-making mechanism of the Random Forest model, an interpretability analysis was conducted using SHAP. This approach makes it possible to identify features that increase the model's tendency to predict complete basic immunization among children, as well as features that contribute negatively by shifting predictions toward incomplete immunization. After obtaining a global overview of feature contributions through Random Forest feature importance analysis, the SHAP-based interpretation was further extended by stratifying children according to geographical location, namely urban and rural areas. This stratification was performed in accordance with the observed differences in immunization coverage between urban and rural areas based on SUSENAS 2024 data. This approach aims to examine differences in both the direction and magnitude of feature contributions across regional contexts, while also providing a more detailed identification of distinct feature contribution patterns between children living in urban and rural areas. The results of the SHAP analysis for children in urban and rural areas are presented in Figure 5.

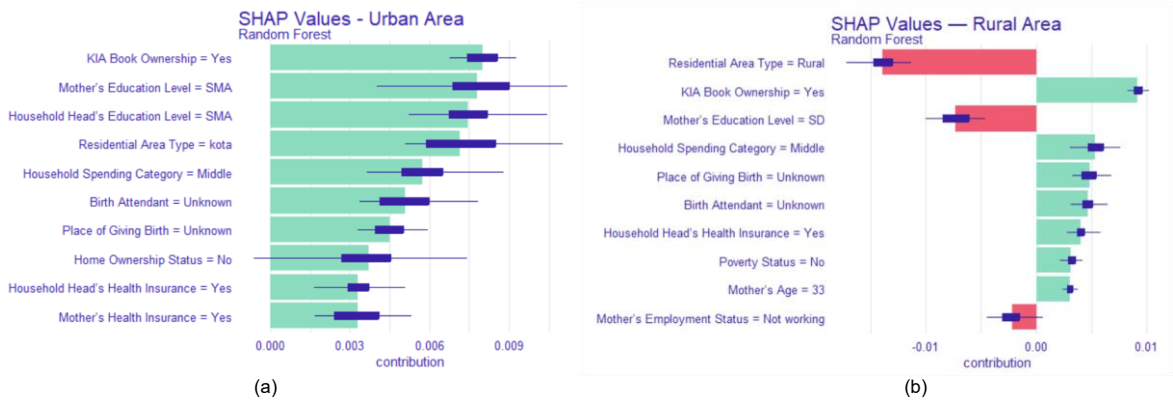

Figure 5 SHAP values plot for: (a) Urban, (b) Rural area

In the SHAP visualization the values on the horizontal axis represent both the magnitude and direction of feature contributions to the model predictions, where positive values (shown in green) increase the likelihood of predicting complete basic immunization, whereas negative values (shown in red) decrease that likelihood. The SHAP analysis results presented in Figure 5 reveal differences in feature contributions between urban and rural areas, reflecting distinct socioeconomic contexts.

In urban areas (Figure 5a), the features with the greatest positive contribution were KIA Book Ownership, Mother's Education Level, and Household Head's Education Level. These features increased the model's tendency to predict complete basic immunization among children. In addition, Household Spending Category and Health Insurance Ownership also showed positive contributions, although with relatively smaller magnitudes. Knowledge-related factors, access to health information, and socioeconomic conditions therefore play important roles in supporting complete immunization coverage in urban areas, as household income, maternal education, residence location, and insurance ownership have also been reported to be associated with children's likelihood of receiving immunization in West Java [26].

In contrast, in rural areas (Figure 5b), the most influential feature was Type of Residential Area (Rural), which showed a negative contribution toward the prediction of complete basic immunization. This finding indicates that living in rural areas remains one of the factors hindering the achievement of complete immunization among children. Such conditions may be associated with various structural barriers, including limited healthcare facilities, unequal access to healthcare services, inadequate infrastructure, and shortages of healthcare workers compared to urban areas. Furthermore, Mother's Education Level in the elementary school category also tended to decrease the probability of complete immunization among children in rural areas. This suggests that lower maternal education levels may affect understanding regarding the importance of immunization, immunization schedules, and access to child health information.

Nevertheless, Ownership of the KIA Book consistently demonstrated a positive influence in increasing the likelihood of complete immunization, similar to the pattern observed in urban areas. This finding is consistent with the role of the KIA Book as a tool for monitoring maternal, infant, and child health up to six years of age [27]. Ownership of the KIA Book indicates that maternal and child health records play an important role in improving parental awareness regarding immunization schedules and child health monitoring. Therefore, strengthening the distribution and utilization of the KIA Book through integrated health posts and primary healthcare services may help improve complete basic immunization coverage. In addition, educational campaigns regarding the importance of completing immunization schedules through the use of the KIA Book may also increase parental compliance with routine immunization programs.

The contribution of Household Expenditure Category also indicates that socioeconomic inequality remains an important barrier to immunization access. Children from households with lower expenditure levels may experience

limitations related to transportation costs, healthcare accessibility, and health literacy, thereby reducing the likelihood of receiving complete immunization. Therefore, more targeted interventions are needed, such as improved access to healthcare services and community-based health education, particularly in rural areas, to help increase complete immunization coverage among children.

Overall, these findings indicate that the determinants of complete basic immunization coverage differ between urban and rural areas. In urban areas, factors related to knowledge and access to health information tend to play a greater role in supporting complete immunization among children. Meanwhile, in rural areas, structural barriers and socioeconomic conditions remain the main factors influencing lower immunization coverage. These findings highlight the importance of designing targeted and region specific immunization interventions that are tailored to the geographical and socioeconomic characteristics of each region.

V. CONCLUSIONS AND SUGGESTIONS

This study identified the determinants of complete basic immunization among children in West Java using three decision tree–based classification algorithms. The results showed that Random Forest combined with SMOTE was retained as the main model after threshold adjustment using Youden's Index. The final model achieved an accuracy of 60.9%, sensitivity of 71.2%, specificity of 50.2%, and balanced accuracy of 60.7%. These results indicate that threshold adjustment improved the model's ability to detect children with incomplete immunization, although a trade-off remained between sensitivity and specificity. Global feature importance analysis revealed that Household Spending Category, KIA Book Ownership, Mother's Age at First Birth, and Mother's Education Level were the most influential factors in predicting complete basic immunization among children. Furthermore, the SHAP analysis stratified by region indicated notable differences in determinants between urban and rural areas. In urban areas, factors related to family health knowledge and literacy were more dominant, whereas in rural areas, structural barriers and socioeconomic conditions played a greater role. These findings confirm that the determinants of basic immunization are contextual and not uniform across regions. Therefore, strategies to improve immunization coverage should be specifically designed to reflect the characteristics and needs of each region.

ACKNOWLEDGEMENT

The authors would like to thank IPB University for supporting this research. Appreciation is also extended to all parties who contributed to and supported the process of writing this paper.

REFERENCES

[1] Kemenkes RI, "Undang-undang Nomor 17 Tahun 2023 tentang Kesehatan," Jakarta, 2023. Accessed: Dec. 06, 2025. [Online]. Available: https://www.peraturan.go.id/Details/258028/uu-no-17-tahun-2023

[2] World Health Organization, "Immunization Agenda 2030: A Global Strategy to Leave No One Behind," 2020. [Online]. Available: https://www.who.int/teams/immunization-vaccines-and-biologicals/strategies/ia2030

[3] Kemenkes RI, "Permenkes Nomor 12 Tahun 2017 tentang Penyelenggaraan Imunisasi," Jakarta, 2022. Accessed: Dec. 06, 2025. [Online]. Available: https://www.peraturan.go.id/id/permenkes-no-12-tahun-2017?.com

[4] Dinas Kesehatan Provinsi Jawa Barat, "Profil Kesehatan Provinsi Jawa Barat Tahun 2024," 2024.

[5] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[6] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[7] A. Y. Perdana, R. Latuconsina, and A. Dinimaharawati, "PREDIKSI STUNTING PADA BALITA DENGAN ALGORITMA RANDOM FOREST."

[8] A. Khairunnisa, K. A. Notodiputro, and B. Sartono, "A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 207–218, Feb. 2024, doi: 10.15294/sji.v11i1.48721.

[9] E. Ramadanti, D. Aprilya Dinathi, C. Christianskaditya, and D. R. Chandranegara, "Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With Hyperparameter Tuning," *Sinkron*, vol. 8, no. 2, pp. 956–963, Mar. 2024, doi: 10.33395/sinkron.v8i2.13530.

[10] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (N. Y).*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[11] "Two Modifications of CNN," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976, doi: 10.1109/TSMC.1976.4309452.

[12] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." [Online]. Available: https://github.com/slundberg/shap

[13] S. I. Oktora, D. Matualage, K. A. Notodiputro, and B. Sartono, "Data-Driven Insights Into Underdeveloped Regencies: SHAP-Based Explainable Artificial Intelligence Approach," *International Journal of Artificial Intelligence Research*, vol. 9, no. 1, May 2025, doi: 10.29099/ijair.v9i1.1399.

[14] R. J. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:14981989

[15] M. A. Gonzalez-Moles *et al.*, "Adhesion molecule CD44 as a prognostic factor in tongue cancer.," *Anticancer Res.*, vol. 23, no. 6D, pp. 5197–202, 2003.

[16] A. Ramadhan, B. Susetyo, and Indahwati, "PENERAPAN METODE KLASIFIKASI RANDOM FOREST DALAM MENGIDENTIFIKASI FAKTOR PENTING PENILAIAN MUTU PENDIDIKAN," *Jurnal Pendidikan dan Kebudayaan*, vol. 4, no. 2, pp. 169–182, Dec. 2019, doi: 10.24832/jpnk.v4i2.1327.

[17] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: 10.35580/variansiunm31.

[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[19] M. Tang *et al.*, "An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes," *Energies (Basel).*, vol. 13, no. 4, p. 807, Feb. 2020, doi: 10.3390/en13040807.

[20] M. Kristanaya, Melinda Putri Azzahra, Trimono, and Mohammad Idhom, "Klasifikasi Status Rujukan Pasien Poliklinik Bandara Berbasis Random Forest dan Interpretabilitas Model Menggunakan SHAP," *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 60–74, Jul. 2025, doi: 10.47709/dsi.v5i1.6078.

[21] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1802.03888

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

[23] A. Nurhopipah and C. Magnolia, "PERBANDINGAN METODE RESAMPLING PADA IMBALANCED DATASET UNTUK KLASIFIKASI KOMENTAR PROGRAM MBKM," *Jurnal Publikasi Ilmu Komputer dan Multimedia*, vol. 2, no. 1, pp. 9–22, Jan. 2023, doi: 10.55606/jupikom.v2i1.862.

[24] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, "Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 36, Dec. 2022, doi: 10.1186/s12911-022-01775-z.

[25] E. F. Schisterman, D. Faraggi, B. Reiser, and J. Hu, "Youden Index and the optimal threshold for markers with mass at zero," *Stat. Med.*, vol. 27, no. 2, pp. 297–315, Jan. 2008, doi: 10.1002/sim.2993.

[26] P. H. Saraswati and A. Gani, "Socio economic factors of child basic immunization: Case of West Java Province," *Jurnal Medicoeticolegal dan Manajemen Rumah Sakit*, vol. 9, no. 1, 2020, doi: 10.18196/jmmr.91111.

[27] Kementerian Kesehatan Republik Indonesia, "Buku Kesehatan Ibu dan Anak," https://ayosehat.kemkes.go.id/buku-kia-kesehatan-ibu-dan-anak.