RESEARCH ARTICLE • OPEN ACCESS

Optimizing Random Forest Parameters with Hyperparameter Tuning for Classifying School-Age KIP Eligibility in West Java

Silfiana Lis Setyowati, Asyifah Qalbi, Rafika Aristawidya, Bagus Sartono, and Aulia Rizki Firdawanti

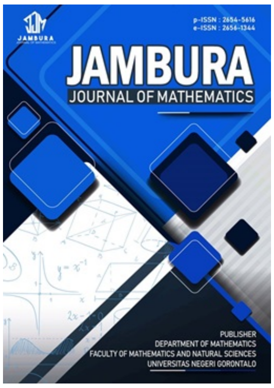
JAMBURA
JOURNAL OF MATHEMATICS

Volume 7, Issue 1, Pages 40–48, February 2025

Received 9 December 2024, Revised 16 January 2025, Accepted 21 January 2025, Published 1 February 2025

To Cite this Article : S. L. Setyowati, et al.,"Optimizing Random Forest Parameters with Hyperparameter Tuning for Classifying School-Age KIP Eligibility in West Java", *Jambura J. Math*, vol. 7, no. 1, pp. 40–48, 2025, *https://doi.org/10.37905/jjom.v7i1.28736*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS



Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

Jambura Journal of Mathematics, Volume 7, Issue 1, Pages 40–48, February 2025 *https://doi.org/10.37905/jjom.v7i1.28736*

Research Article

Optimizing Random Forest Parameters with Hyperparameter Tuning for Classifying School-Age KIP Eligibility in West Java

Silfiana Lis Setyowati[1,2,*], **Asyifah Qalbi**[1], **Rafika Aristawidya**[1], **Bagus Sartono**[1], **and Aulia Rizki Firdawanti**[1]

[1] *Study Program of Statistics and Data Science, IPB University, Bogor, Indonesia*
[2] *Ministry of Higher Education, Science, and Technology, Indonesia*

ARTICLE HISTORY

Received 9 December 2024
Revised 16 January 2025
Accepted 21 January 2025
Published 1 February 2025

KEYWORDS

Kartu Indonesia Pintar (KIP)
Random Forest
SMOTE
Optimal Parameter
Hyperparameter tuning

ABSTRACT. *Random Forest is an ensemble learning algorithm that combines multiple decision trees to generate a more stable and accurate classification model. This study aims to optimize Random Forest parameters for classifying school-age students' eligibility for the Kartu Indonesia Pintar (KIP) in West Java, based on economic factors. The research uses secondary data from the 2023 National Socio-Economic Survey (SUSENAS) of West Java, with a sample size of 13,044 individuals. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) is applied. Hyperparameter tuning through grid search identifies the optimal combination of parameters, including the number of trees (ntree), random variables per split (mtry), and terminal node size (node_size). Model performance is evaluated using balanced accuracy, sensitivity, and specificity. Results indicate that the optimal parameters (mtry = 5, ntree = 674, node_size = 26) yield a balanced accuracy of 65.47%. Significant variables include PKH status, floor area of the house, source of drinking water, and building material type. The model accurately identifies students in need of educational assistance. In conclusion, optimizing Random Forest parameters improves the accuracy of KIP eligibility classification, supporting educational equity policies in West Java. These findings provide a foundation for developing more effective beneficiary selection systems for educational aid.*

1. Introduction

Education is a fundamental right of every Indonesian citizen, as stipulated in Article 31 paragraph 1 of the 1945 Constitution which states that "Every citizen has the right to education." In fact, access to education in Indonesia is still unequal, especially for children from poor and vulnerable families. This inequality is reflected in the Gross Participation Rate (APK) and Pure Participation Rate (APM) data at various levels of education, which varies in each region in Indonesia. Nationally, BPS data for 2023 [1] shows that the APK at the SD/MI level reached 104.53% with an APM of 98.20%, indicating that most children aged 7-12 years attended the appropriate level. However, at the SMP/MTs level, the APK dropped to 94.69% and the APM to 83.61%, while at the SMA/SMK/MA level the decline was even more significant, with an APK of only 79.07% and an APM of 59.01%. This decline shows that the higher the level of education, the more children are not attending school at their age. This limited access to education is influenced by economic constraints. The same condition also occurs in West Java Province, which is one of the provinces with the largest population in Indonesia. BPS data for 2023 shows that the APK at the SD/MI level in West Java was 104.46% with an APM of 96.75%. Meanwhile, at the SMP/MTs level, the APK reached 99.61% and the APM dropped to 76.73%, and at the SMA/SMK/MA level, the APK was recorded at 90.94% with an APM of only 69.70%. Compared to the national average, West Java has higher APK and APM at the SMA/SMK/MA level, but there are still around 30% of school-age children who have not attended the appropriate level. This decline in education participation indicates a serious challenge in ensuring equitable access to education, especially at the secondary level.

Many school-age children, especially from poor and vulnerable families, are unable to fully enjoy their right to education due to financial constraints. This inequality is one of the main causes of educational inequality. The government has attempted to launch the Kartu Indonesia Pintar (KIP) Program as a strategic effort to improve access to education. The program is designed to provide assistance for children from poor and vulnerable families to continue their education up to secondary and tertiary levels [2]. KIP aims to ensure that no child drops out of school due to financial constraints, in line with the government's vision of improving education equity and the quality of human resources.

The KIP program has played a significant role in improving access to education for children from poor and vulnerable families in Indonesia. Prior to the launch of KIP, data showed that APK and APM at the secondary education level were still low. Based on data from the Central Bureau of Statistics (BPS), in 2014, the APK of SMA/SMK/MA level nationally only reached 78.02%, while the APM was at 60.67% [3]. After KIP was launched in 2015, there was a significant increase in school participation. In 2018, the APK at the same level increased to 82.84%, while the APM reached

[*] Corresponding Author.

Email : *silfianalissetyowati@gmail.com* (S. L. Setyowati)
Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

63.07% [4]. This increase shows that the KIP program has successfully encouraged more children to continue their education to higher levels and reduced dropout rates. This is supported by a report from the Ministry of Education and Culture (MoEC) which notes the positive impact of KIP on reducing dropout rates. At the primary education level, the number of children who dropped out of school decreased from 60,066 in the 2015/2016 academic year to 32,127 in the 2017/2018 academic year [2]. This data indicates that the implementation of KIP has helped to address inequality in access to education in Indonesia, increase school participation, and support the government's vision of education equity. Although this program has been successful in addressing the education gap, its implementation in the field still faces various challenges, especially related to inaccurate targeting. Based on a report by the Central Bureau of Statistics [1] and found that many KIP recipients do not meet the eligibility criteria, children from poor families who actually need assistance are often not accommodated. This targeting inaccuracy causes the program to be ineffective, so improving the accuracy of KIP distribution is an urgent need. The government needs to develop a more accurate and transparent data-based selection mechanism so that this program truly targets the groups in need.

To address the challenges of ensuring equitable access to education and improving the accuracy of programs like Kartu Indonesia Pintar (KIP), machine learning methods such us random forest, offer an effective solution. Random forest is an ensemble learning method that combines many decision trees to produce a more accurate model. Random forest has several advantages such as the ability to handle data with complex variables and stable performance when there is noise in the data. Previous research shows the superiority of the random forest algorithm in terms of classification. Nabillah et al. [5] in his research concluded that random forest provides the highest accuracy of 78.02% compared to other algorithms such as k-nearest neighbors, naive bayes classifier, and C4.5 algorithm in the classification of educational aid recipients. Another study conducted by Luchia et al. [6] concluded that with chi-square and information gain-based feature selection, random forest achieved 99% accuracy, outperforming the support vector machine (SVM) algorithm which reached 98%.

On the other hand, the success of the random forest algorithm is highly dependent on the parameters used, such as the number of trees, the depth of the trees, and the number of features selected at each branch. Erlin [7] and Mualfah et al. [8] highlighted that without parameter optimization, Random Forest performance can degrade, especially when used on unbalanced or complex data. Based on this, it means that parameter optimization is an important step to maximize model performance, one of the techniques that can be used to optimize parameters is hyperparameter tuning. Research conducted by Tan et al. [9] shows that the combination of random forest algorithm with Synthetic Minority Oversampling Technique (SMOTE) method with hyperparameter tuning can effectively overcome class imbalance in wireless sensor network intrusion detection. This underscores the importance of proper method selection and parameter optimization to maximize the performance of random forest algorithms in various application contexts. However, previous studies did not specifically compare the effectiveness of combining SMOTE with grid search tuning and SMOTE with hyperparameter tuning. This gap highlights the need for further research to evaluate and compare these approaches, as the choice of tuning method can significantly influence the overall performance of random forest algorithms when dealing with imbalanced datasets.

This research uses the random forest algorithm to improve the classification accuracy of Kartu Indonesia Pintar (KIP) recipient eligibility. Considering the ongoing challenges in accurately targeting and providing fair access to education aid, this research is essential to enhance the efficiency and success of the KIP program. By addressing these challenges, the study aims to ensure that educational aid is distributed appropriately, prioritizing the most disadvantaged and vulnerable populations, thereby supporting the broader goal of reducing educational inequality in Indonesia. This study applies hyperparameter tuning to optimize performance of random forest model using two key approaches: grid search and random search. Grid search systematically combines key parameters, including the number of trees (n_tree), the number of features per split (m_try), and the minimum terminal node size (node_size). In contrast, random search randomly samples hyperparameter combinations within predefined ranges. Both methods are utilized to compare their performance in improving the Random Forest model. Each combination of hyperparameters is evaluated using cross-validation, with metrics such as balanced accuracy, sensitivity, and specificity used to determine the optimal configuration. By optimizing the random forest parameters, this study aims to produce a more accurate model for use in the selection process of education aid recipients. This approach differs from previous studies, which focused more on method comparisons or feature selection. Furthermore, this study refines previous research by evaluating and comparing the performance of optimal SMOTE random forest parameters using both grid search and random search methods. The goal of this research is to optimize the random forest parameters to enhance the classification accuracy of school-age KIP recipients in West Java based on economic conditions. The results of this study are expected to produce an accurate model that can provide policy recommendations for educational equity, particularly in West Java.

2. Methods

2.1. *Research Procedures*

The research flow is exploring and pre-processing data, splitting data, modelling using random forest with hyperparameter tuning on training data, evaluation model on training data and testing data using sensitivity, specificity, and balanced accuracy. The best model is selected based on the high value of sensitivity, specificity, and balanced accuracy. based on the best model, the important variables that become the constituent factors in the observed response variables are obtained. This research was conducted following the steps of the research flowchart in Figure 1.

2.2. *Data Source*

The data used in this study is primary data from the 2023 National Socio-Economic Survey (SUSENAS) of West Java Province conducted by BPS. The population in this study consists of

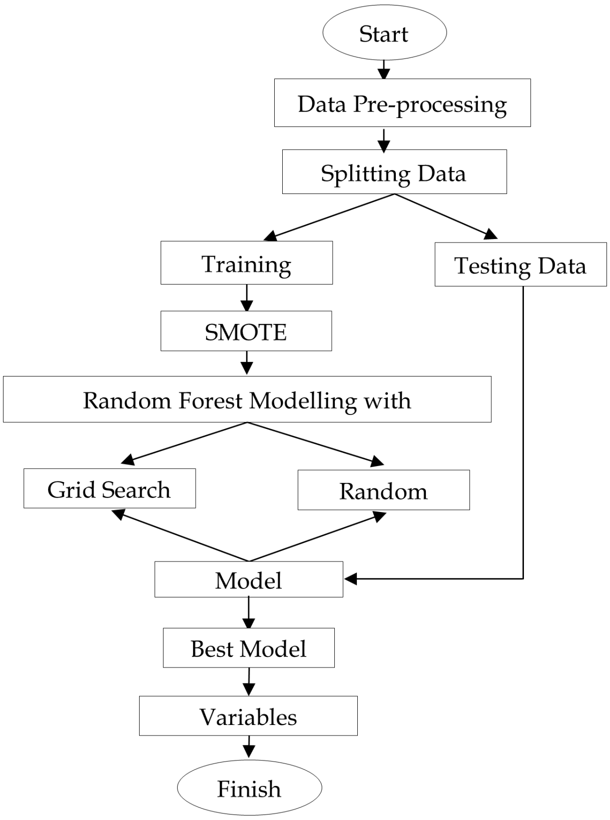


Figure 1. Research procedure

Table 1. Predictor variables in the study

Code	Variables	Measurement Scale
R1101	health insurance ownership status	Binary
R1802	residential building ownership status	Binary
R1804	floor area of residential building	Numeric
R1806	main house material of the widest house roof	Nominal
R1807	main house material of the widest house wall	Nominal
R1808	main house material of the widest house floor	Nominal
R1810A	main water source for drinking	Nominal
R1817	main types of fuel for cooking	Nominal
R1901A	KUR receipt status	Binary
R1901B	Sources of credit loans at commercial banks other than KUR	Nominal
R2001A	ownership status of 5.5 kg LPG gas	Binary
R2001B	refrigerator ownership status	Binary
R2001C	AC ownership status	Binary
R2001D	water heater ownership status	Binary
R2001E	landline phone ownership status (PTSN)	Binary
R2001F	computer/laptop ownership status	Binary
R2001G	gold/jewelry ownership status (min 10 grams)	Binary
R2001H	motorcycle ownership status	Binary
R2001I	boat ownership status	Binary
R2001J	motorboat ownership status	Binary
R2001K	car ownership status	Binary
R2001L	flat screen television ownership status (min 30 inch)	Binary
R2001M	land ownership status	Binary
R2101A	largest source of financing in households	Nominal
R2203	Recipient Status of PKH	Binary

school-age children, aged 6-18 years, totaling 13.044 individuals. The response variable is the ownership status of the Kartu Indonesia Pintar (KIP) among school-age children, while the predictor variables are factors that are believed to influence the KIP ownership status. There are 25 predictor variables, which are detailed in Table 1.

2.3. *Data Analysis*

The analysis procedure is performed using the R software. The data pre-processing stage is carried out to convert raw data into a form that can be understood by the system. The modeling stage is conducted to build the model, generate predictions, and evaluate to obtain the best model. The steps in the data analysis for this study are as follows:

1. Data pre-processing
   (a) Prepare the KIP data and predictor variables from the 2023 SUSENAS data of West Java.
   (b) Categorize the status of school-age KIP recipients in West Java.
   (c) Perform exploration to understand the characteristics of the data.
   (d) Remove highly correlated variables to avoid multicollinearity. Multicollinearity can affect stability [10].
   (e) Split the data into two parts, with 80% for training and 20% for testing.
   (f) Identify potential imbalance issues in the data. If data imbalance is detected, apply data handling techniques such as Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance. Fernandez at al. [11] demonstrate the superior effectiveness of SMOTE in improving classification performance compared to other oversampling techniques, as the synthetic data produced is more representative and helps the model generalize better to unseen data.
2. Random Forest Modelling with Hyperparameter Tuning
   Modeling is carried out using the random forest technique, with data handled using SMOTE, and the following parameters are set:
   i. Parameter Combination with grid search:
      1) n-tree
         The number of trees tested are 100, 500, and 1000 to evaluate the model's stability. A larger number of trees generally provides a more stable model but requires longer computation time [12].
      2) m-try
         This value represents the number of random predictor variables considered at each node split. m-try is set to 2, 4, and 6 based on the recommendation [13], which suggests setting m-try to $\sqrt{p}$, where p is the number of predictor variables. Experiments with low (2), medium (4), and high values are expected to illustrate the impact of random predictor variables on each node in terms of the model's performance.

3) node_size

This value represents the minimum terminal node size. It is set to 1, 5, and 10 to evaluate the model's granularity. A small value (1) allows the model to capture details well but may be prone to overfitting, while larger values (5 and 10) help reduce overfitting while maintaining model performance [14].

ii. Hyperparameter tuning is performed to determine the best parameter values using a 5-fold cross-validation method with random search.

3. Model Evaluation

   Evaluation is conducted to measure the model's performance in classifying the classes correctly. In this study, the positive class refers to school-age children who do not have a KIP, and the negative class refers to children who possess a KIP. Model evaluation focuses on sensitivity, specificity, and balanced accuracy.

   (a) Sensitivity

   Sensitivity is the model's ability to correctly classify the positive class, in this case, school-age children who do not have a KIP. Sensitivity is important to ensure the model does not misidentify children who should not be eligible for KIP as eligible recipients. It is also crucial for identifying groups that deserve attention because they are not receiving KIP despite being eligible.

   (b) Specificity

   Specificity shows the model's ability to correctly identify the negative class, which in this case is school-age children who have a KIP. Specificity is important to ensure children who already have KIP are not misclassified as not receiving KIP. This metric helps identify errors in classifying children who are already beneficiaries of the program.

   (c) Balanced Accuracy

   Balanced Accuracy is the average of sensitivity and specificity. This metric is suitable for imbalanced data, as it gives equal weight to performance on both classes. In this case, balanced accuracy provides a fair view of both KIP recipients and non-recipients, thus avoiding bias towards the dominant class.

4. Selecting the Best Model

   The best model is determined by comparing the evaluation metrics from the SMOTE Random Forest model with parameters from step (2.a.i), and the Random Forest model with hyperparameter tuning. The model with the highest metric values is considered the better model.

5. Variables Importance

   The best model that has been obtained is then used to identify important variables that affect the response variable. These important variables are then used for interpretation.

2.4. *Random Forest*

Random forest is an ensemble learning technique used for classification modeling by combining bootstrap aggregating (bagging) and random feature selection [15]. The steps in building a random forest classification model are as follows:

1. Create n-tree decision trees from the training data.
2. For each tree, perform bootstrap sampling from the training data and build the decision tree from this bootstrap sample.
3. For each split in the tree:
   (a) Select a random number of m-try predictor variables.
   (b) Determine the splitting point.
   (c) Split the data into two branches.
4. Repeat this process for all trees until completion.
5. Generate the ensemble of all trees through majority voting.

2.5. *Synthetic Minority Oversampling Technique (SMOTE)*

The SMOTE process is done by determining the minority class in the data, then determining the distance of k nearest neighbors obtained by calculating the Euclidean distance between minority data. Next, generate synthetic data on the line connecting k nearest neighbors and generate random points on the line [16].

3. Results and Discussion

3.1. *Data Pre-Processing*

Data pre-processing in the research is filtering and cleaning the variables to be used from the available SUSENAS data into a dataset that is ready to be used for further analysis. This stage encompasses both data preparation and exploratory data analysis.

3.1.1. Data Preparation

In this stage, KIP data from the 2023 National Socio-Economic Survey (SUSENAS) for West Java Province was processed, including relevant predictor variables. The categorization of KIP ownership status was performed to separate the data into two classes: ”Receiving KIP” and ”Not Receiving KIP.”

3.1.2. Data Exploration

One important step in data preprocessing is the examination of multicollinearity. A correlation check is carried out between categorical independent variables to ensure that there are no high linear relationships between variables that could affect the analysis results.

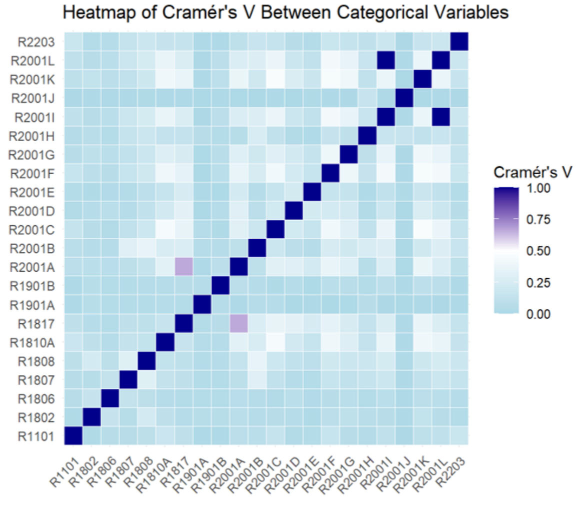


Figure 2. Multicollinearity check

The correlation between predictor variables was examined to identify high linear relationships. Based on Figure 2, it

Table 2. Evaluation metrics for parameter combinations with grid search

No	ntree	mtry	node_size	Accuracy	Balanced Accuracy	Sensitivity	Specificity
1	100	2	1	0.826	0.659	0.897	0.421
2	100	2	5	0.819	0.661	0.886	0.436
3	100	2	10	0.828	0.666	0.896	0.436
4	100	4	1	0.849	0.653	0.933	0.374
5	100	4	5	0.848	0.648	0.933	0.362
6	100	4	10	0.848	0.636	0.938	0.335
7	100	6	1	0.847	0.612	0.948	0.276
8	100	6	5	0.847	0.614	0.946	0.281
9	100	6	10	0.852	0.616	0.953	0.278
10	500	2	1	0.819	0.655	0.889	0.421
11	500	2	5	0.813	0.660	0.879	0.441
12	500	2	10	0.820	0.661	0.888	0.433
13	500	4	1	0.851	0.649	0.937	0.362
14	500	4	5	0.849	0.646	0.935	0.357
15	500	4	10	0.850	0.642	0.939	0.345
16	500	6	1	0.852	0.612	0.954	0.271
17	500	6	5	0.850	0.616	0.950	0.281
18	500	6	10	0.852	0.623	0.950	0.296
19	1000	2	1	0.819	0.660	0.886	0.433
20	1000	2	5	0.816	0.660	0.882	0.438
21	1000	2	10	0.818	0.659	0.885	0.433
22	1000	4	1	0.850	0.644	0.938	0.350
23	1000	4	5	0.850	0.647	0.937	0.357
24	1000	4	10	0.850	0.64	0.939	0.342
25	1000	6	1	0.851	0.612	0.952	0.271
26	1000	6	5	0.853	0.616	0.953	0.278
27	1000	6	10	0.851	0.617	0.951	0.283

was found that R2001L (ownership of flat-screen televisions) and R2001I (ownership of boats) had a perfect correlation (correlation value = 1). Therefore, the variable R2001I was removed to avoid multicollinearity issues that could affect the model's stability. Next, the data imbalance in the dataset was examined, as shown in Figure 3.

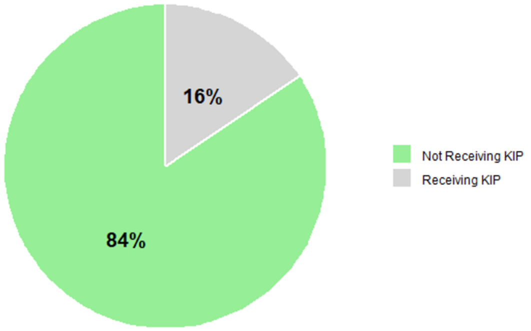


Figure 3. KIP acceptance status

Figure 3 shows that 84% (11.013 observations) were classified as "Not Receiving KIP," while only 16% (2.031 observations) were classified as "Receiving KIP." This imbalance could bias the model toward the majority class. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied.

3.2. SMOTE Random Forest Modeling

The Random Forest modeling for classifying the eligibility of school-age KIP recipients was carried out involving 24 predictor variables believed to influence eligibility. The dataset of 13,044 individuals was divided into two sets: 80% for training data and 20% for test data. The class imbalance issue was addressed using the SMOTE technique. In this stage, Random Forest modeling was performed with several parameter combinations, namely the number of trees (n_tree) tested at values of 100, 500, and 1000; the number of random predictor variables used for each tree split (m_try) set to 2, 4, and 6; and the minimum node size (node_size) set to 1, 5, and 10. These parameter combinations resulted in 27 models, as shown in Table 2, with the optimal combination expected to yield the best model performance for classifying KIP eligibility.

Based on the Random Forest modeling results with various parameter combinations shown in Table 2, most models with ntree values of 100 and 500 achieved high accuracy, ranging from 0.813 to 0.852. The best-performing model, combination 9 (ntree 100, mtry 6, node_size 10), achieved an accuracy of 0.852. However, this improvement in accuracy was often accompanied by a decrease in specificity, increasing the risk of false positives. This indicates that while the model is generally good at predicting outcomes, it may neglect the minority class.

The best model (combination 9) had an accuracy of 0.852 but a low specificity of 0.278, meaning it effectively predicted the positive class but struggled with the negative class, leading to increased false positives. Models with ntree 100 and mtry 6 showed high sensitivity (up to 0.95), indicating strong performance in predicting the positive class. High sensitivity often occurs with smaller node_size values, which may cause the model to focus on the majority (positive) class. For example, model 7 (node_size 1) showed high sensitivity (0.947) but low specificity

(0.275), suggesting the model is good at predicting positive outcomes but weak at identifying the negative class.

On the other hand, models with a larger node_size (e.g., node_size 10) generally had higher specificity but lower sensitivity. This indicates the model is good at identifying the negative class but struggles to detect the positive class, resulting in more false negatives. The highest balanced accuracy was observed in combination 3 (ntree 100, mtry 2, node_size 10), which reached 0.666. This model strikes a good balance between sensitivity (0.896) and specificity (0.435). Although its specificity is slightly lower than some other models, its high balanced accuracy suggests it handles class imbalance well, making it effective at identifying both positive and negative classes fairly. This is crucial for the KIP eligibility classification.

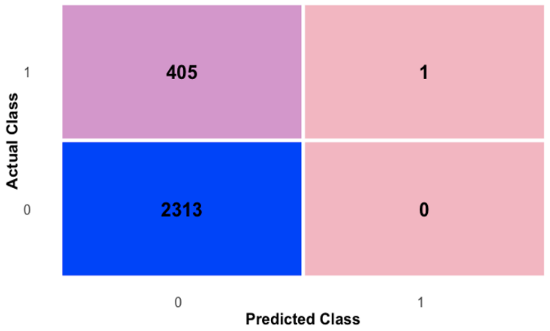


Figure 4. Confusion matrix for SMOTE random forest model

The optimal parameter combination, namely n_tree = 100, m_try = 2, and node_size = 10, was used to build the model on the test data. The confusion matrix is presented in Figure 4. The model performs well in identifying Class 0. It correctly classifies 2.313 instances (True Positive), but also makes 405 classification errors, where instances that should belong to Class 0 are misclassified as Class 1 (False Positive). This suggests a tendency for the model to incorrectly identify Class 0 as Class 1, which could lead to overfitting on the minority class. On the other hand, for Class 1, the model correctly classifies data as positive (True Positive), but fails to detect any other instances. This is because, in the testing data, only one instance belongs to Class 1.

The Out-of-Bag (OOB) Error Rate is 14.94%, indicating that the model correctly classified approximately 85.06% of the data not involved in the training process (out-of-bag sample). As detailed in Table 3, the accuracy of the model is 85.10%, meaning it correctly classifies the majority of the test data. The sensitivity value of 1 indicates the model successfully identifies all positive class data. However, the specificity is extremely low at 0.24%, suggesting that the model almost completely fails to recognize the negative class. The balanced accuracy of 50.12% indicates that the model is not performing well in classifying both classes proportionally.

Table 3. Evaluation metrics of the model on test data

Evaluation Metric	Value
Accuracy	0.851
Balanced Accuracy	0.501
Sensitivity	1
Specificity	0.002

3.3. *SMOTE Random Forest Modeling with Hyperparameter Tuning*

The model was tuned using hyperparameter tuning with 5-fold cross-validation. The optimal parameters obtained are shown in the Table 4.

Table 4. Hyperparameter tuning with random search

mtry	ntree	node_size	Metric	Mean
9	**1202**	**29**	**accuracy**	**0.811**
9	791	7	accuracy	0.810
9	1388	16	accuracy	0.810
9	437	13	accuracy	0.810
9	1974	15	accuracy	0.809
5	**674**	**26**	**balanced accuracy**	**0.665**
5	485	37	balanced accuracy	0.665
5	271	10	balanced accuracy	0.665
5	1634	32	balanced accuracy	0.665
4	163	28	balanced accuracy	0.664

Based on the Table 4, the optimal parameters for building the classification model, according to balanced accuracy, are mtry = 5, trees = 674, and node_size = 26. The confusion matrix for the predictions on the test data is shown in Figure 5.

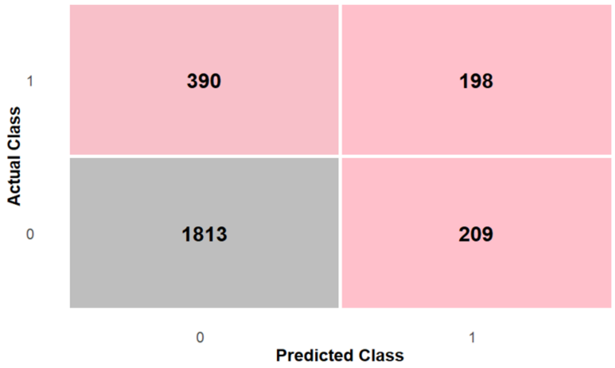


Figure 5. Confusion matrix with hyperparameter tuning

Based on Figure 5, the model's ability to predict class 0 (Not Receiving KIP) with 1813 correct predictions is good, while it performs reasonably well in predicting class 1 (Receiving KIP) with 198 correct predictions. The model evaluation metrics on the test data are shown in Table 5.

Table 5. Model evaluation metrics on test data

Evaluation Metric	Value
Accuracy	0.771
Balanced Accuracy	0.655
Sensitivity	0.823
Specificity	0.487

The evaluation metrics on the test data, as shown in Table 5, indicate that the model built using the optimal parameters from hyperparameter tuning has a balanced accuracy of 0.655 (65.47%), which is fairly good for making predictions on the test data. The sensitivity and specificity values are 82.30% and 48.65%, respectively.

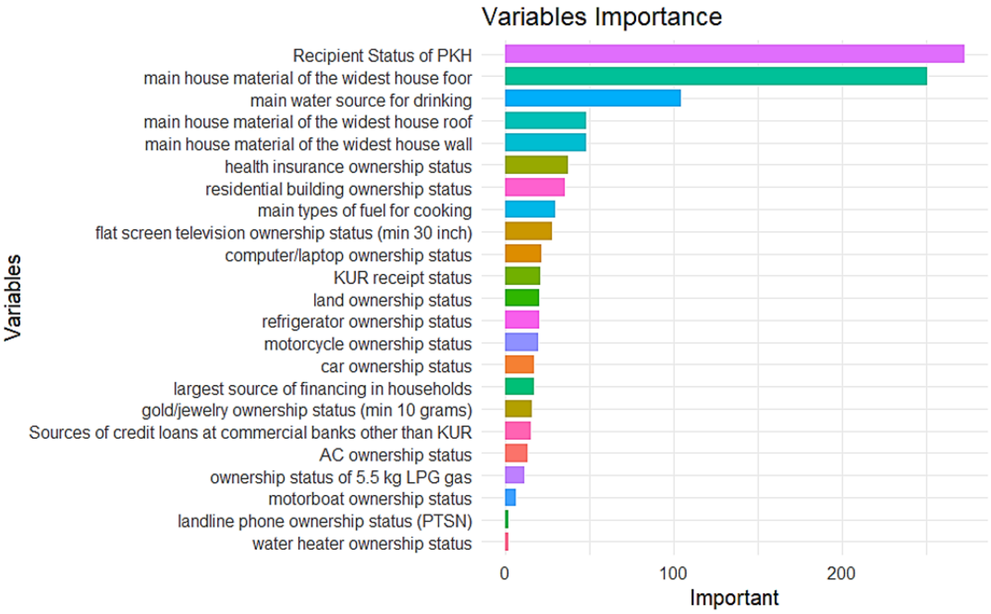
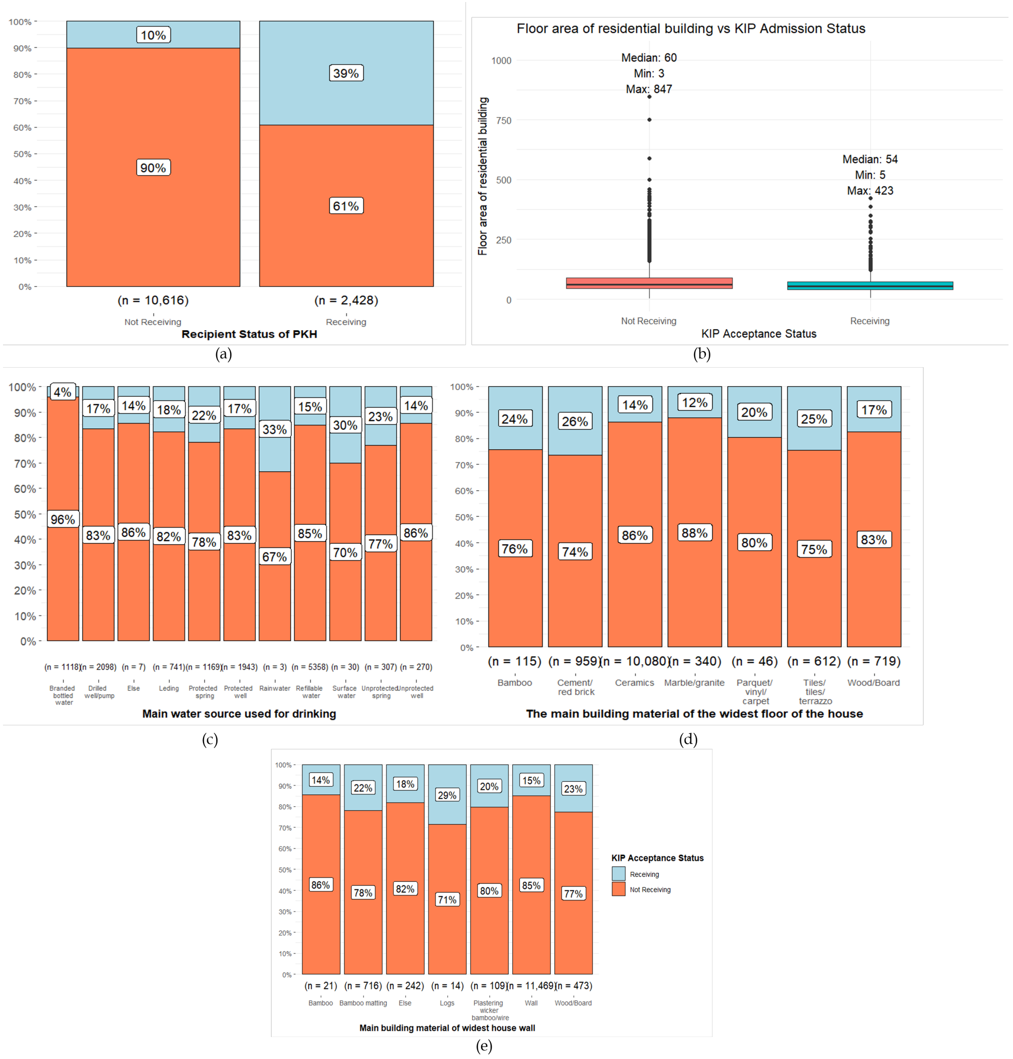


Figure 6. Important variables



(a)

(b)

(c)

(d)

(e)

Figure 7. Visualization of the relationship between the five most important variables and KIP acceptance status

3.4. Model Performance Comparison

The model performance is evaluated by comparing the performance of the model with parameter combinations and the model with hyperparameter tuning. The comparison of model performance is presented in the Table 6.

Table 6. Comparison of model performance on test data

Evaluation Metric	Parameter Combination	Hyperparameter tuning
Accuracy	0.851	0.771
Balanced Accuracy	0.501	0.655
Sensitivity	1	0.823
Specificity	0.002	0.487

Based on Table 6, the model using hyperparameter tuning performs better than the model with parameter combinations. The balanced accuracy shows that the model with hyperparameter tuning is better at recognizing both class 0 (Not Receiving KIP) and class 1 (Receiving KIP).

3.5. Variable Importance

Based on the model with the optimal parameters mtry = 5, trees = 674 and node_size = 26, the important variables are presented in Figure 6. The five most important variables in building the Random Forest classification model are: the PKH (Program Keluarga Harapan) acceptance status, the floor area of the residence, the main source of drinking water, the primary building material for the largest floor area of the house, and the primary building material for the largest wall area of the house. The explanation of these variables can be seen in Figure 7.

Based on Figure 7, the following explanations can be provided:

1. Program Keluarga Harapan (PKH) as the most important variable in forming the classification model aligns with the government's program, which states that one of the requirements for receiving KIP is participation in PKH. However, an interesting observation from the figure is that there are 947 KIP recipients who are also PKH recipients, whereas 1.062 KIP recipients do not participate in PKH. This could suggest targeting issues or that KIP recipients who do not receive PKH may belong to special categories, such as orphans, people with disabilities, or victims of natural disasters.
2. The largest floor area of the house for KIP recipients is 423 m^2, while the largest floor area for individuals who do not receive KIP is 847 m^2. This suggests that most of those who receive or do not receive KIP are aligned with their housing conditions.
3. The majority of both KIP recipients and non-recipients use similar main sources of drinking water. According to BPS (Statistics Indonesia), someone is considered poor if their main drinking water source comes from unprotected wells, springs, rivers, or rainwater.
4. Most KIP recipients and non-recipients share similar types of flooring materials used in their homes. According to BPS, someone is considered poor if the flooring in their home is made of dirt, bamboo, or wood.
5. Most KIP recipients and non-recipients also have similar types of wall materials used in their homes. According to BPS, someone is considered poor if the walls of their home are made from bamboo or low-quality wood, or if the walls are unplastered brick.

4. Conclusion

The optimal parameters for building the Random Forest model for classifying KIP eligibility based on balanced accuracy are mtry = 5, ntree = 674, and node_size = 26. The use of hyperparameter tuning is more effective for determining the optimal parameters, but it requires a long computation time. The five most important variables affecting the classification model for KIP recipients in West Java, based on economic conditions, are PKH, the floor area of the building, the main source of drinking water, the primary material of the largest floor, and the primary material of the largest wall.

Author Contributions. **Silfiana Lis Setyowati:** Conceptualization, methodology, analysis, and writing.—original draft preparation, writing—review and editing. **Asyifah Qalbi:** Conceptualization, methodology, analysis, and writing.—original draft preparation, writing—review and editing. **Rafika Aristawidya:** Conceptualization, methodology, analysis, and writing.—original draft preparation, writing—review and editing. **Bagus Sartono:** Supervision, data acquisition, and project review. **Aulia Rizki Firdawanti:** Supervision, data acquisition, and project review. All authors have read and agreed to the published version of the manuscript.

Acknowledgement. This research is supported by IPB University. The authors would like to express their appreciation to the editors and reviewers for their careful review, insightful feedback, and constructive suggestions that have contributed significantly to improving the quality of this paper.

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest related to this article.

References

[1] Badan Pusat Statistik (BPS), "Statistik Pendidikan Indonesia: Data Angka Partisipasi Sekolah," 2023. [online]. Available: https://www.bps.go.id.

[2] Kementerian Pendidikan dan Kebudayaan (Kemendikbud), "Laporan Penurunan Angka Putus Sekolah melalui Program Indonesia Pintar," 2018. [online]. Available: https://jendela.kemdikbud.go.id.

[3] Badan Pusat Statistik (BPS), "Angka Partisipasi Kasar (APK) dan Angka Partisipasi Murni (APM) pada berbagai jenjang pendidikan di Indonesia," 2014. [online]. Available: https://www.bps.go.id.

[4] Badan Pusat Statistik (BPS), "Angka Partisipasi Kasar (APK) dan Angka Partisipasi Murni (APM) pada berbagai jenjang pendidikan di Indonesia," 2018. [online]. Available: https://www.bps.go.id.

[5] P. Nabillah, I. Permana, M. Afdal, F. Muttakin, and A. Marsal, "A Comparative Study of the Performance of KNN, NBC, C4. 5, and Random Forest Algorithms in Classifying Beneficiaries of the Kartu Indonesia Sehat Program," *JUSIFO (Jurnal Sistem Informasi)*, vol. 10, no. 1, pp. 17–26, 2024.

[6] N. T. Luchia, M. Mustakim, N. Noviarni, K. Sussolaikah, and T. Arifianto, "Feature Selection In Support Vector Machine And Random Forest Algorithms For The Classification Of Recipients Of The Smart Indonesia Program," in *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, IEEE, 2024, pp. 1–6.

[7] Erlin, "Optimasi Parameter Random Forest pada Dataset Tidak Seimbang. Jurnal Ilmu Komputer," *Jurnal Ilmu Komputer*, vol. 8, no. 2, pp. 87–96, 2022.

[8] D. Mualfah, W. Fadila, and R. Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 2, pp. 107–113, 2022.

[9] X. Tan *et al.*, "Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm," *Sensors*, vol. 19, no. 1, p. 203, 2019, doi: 10.3390/s19010203.

[10] K. I. Sundus, B. H. Hammo, M. B. Al-Zoubi, and A. Al-Omari, "Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset," *Inform Med Unlocked*, vol. 33, p. 101088, 2022.

[11] A. Fernández, S. Garcia, F. Herrera, and N. V Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[12] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 3, p. e1301, 2019, doi: 10.1002/widm.1301.

[13] Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S.-S. Ho, "ForesTexter: An efficient random forest algorithm for imbalanced text categorization," *Knowl Based Syst*, vol. 67, pp. 105–116, 2014, doi: 10.1016/j.knosys.2014.06.004.

[14] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014, doi: 10.48550/arXiv.1407.7502.

[15] L. Breiman, "Random forests," *Mach Learn*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] A. F. Anjani, D. Anggraeni, and I. M. Tirta, "Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ," *Jurnal Nasional Teknologi Dan Sistem Informasi*, vol. 9, no. 2, pp. 163–172, 2023.