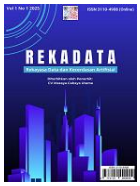

Rekayasa Data dan Kecerdasan Artifisial (REKADATA)
Volume: 2, **Issue:** 1, August 2026
Pages: 140-147
URL: https://journal.mazayacahayautama.com/index.php/rd
ISSN: 3110-4908 (ONLINE)

URL ISSN

Application of Random Forest with SMOTE and Random Search for Food Security Classification in Indonesia

Susi Hasriani [1], Nariza Wanti Wulan Sari [1]*, Pratama Yuly Nugraha[1], M. Fathurahman[1], Meiliyani Siringoringo [1]

[1] *Statistics, Mulawarman University, Indonesia*

*Corresponding Author: nariza@fmipa.unmul.ac.id

Abstract: Food security is an important aspect of sustainable development and one of the key priorities of the Sustainable Development Goals (SDGs). Differences in food security conditions across regencies and cities in Indonesia require accurate classification to support the identification of vulnerable regions and appropriate policy-making. The random forest method is a classification method with good predictive performance, which can be improved through hyperparameter optimization using random search. However, its classification performance may decline when applied to imbalanced datasets. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance. This study aimed to classify the food security levels of regencies and cities in Indonesia in 2024 using random forest with SMOTE and random search hyperparameter optimization and to evaluate model performance based on accuracy, precision, recall, and F1-score. The data were obtained from the Food Security and Vulnerability Atlas (FSVA), comprising 514 regencies and cities, with food security level as the response variable and eight predictor variables. The dataset was divided into training and testing sets using proportions of 80:20 and 90:10. SMOTE was applied with K = 5, while random search was conducted using 20 hyperparameter combinations to determine the optimal model. The results showed that the 80:20 data partition outperformed the 90:10 partition, achieving an accuracy of 88.35%, precision of 83.33%, recall of 50.00%, and F1-score of 62.50%. The best model correctly classified 10 food-insecure and 81 food-secure regencies and cities. These findings provide information for identifying regions based on food security levels and support more targeted food security policy-making.

Keywords: classification, food security, random forest, random search, SMOTE.

Introduction

Food security is one of the important aspects of sustainable development and is addressed in the Sustainable Development Goals (SDGs), particularly Goal 2, Zero [1]. Food security reflects the fulfillment of people's food needs in terms of quantity, quality, safety, diversity, nutritional value, equitable distribution, and affordability. This condition needs to be considered because it is closely related to people's ability to live healthy, active, and productive lives sustainably. Therefore, identifying food security levels across regions is necessary to determine priority areas and support the formulation of targeted policies [2]. One of the instruments used to describe these conditions is the Food Security and Vulnerability Atlas (FSVA), which maps food vulnerability based on indicators of food availability, accessibility, and utilization [3]. The information provided by the FSVA can serve as a basis for identifying and classifying regions according to their food security levels.

To classify regions based on their food security levels, a machine learning approach, particularly supervised learning, can be used to learn patterns from labeled data and determine classes based on their characteristics [4]. One suitable method is random forest, an ensemble method that combines multiple decision trees to produce more reliable classifications [5]. The application of random forest to food security has been carried out to predict the Food Security Index in Indonesia. The results showed that random forest outperformed multiple linear

regression, as indicated by lower MSE and RMSE values [6]. These findings demonstrate the potential of random forest to process food security indicators and produce accurate predictions or classifications.

In addition to selecting an appropriate classification method, the characteristics of the data and model configuration also need to be considered. One of the challenges in classification is an imbalanced class distribution, in which the number of observations in one class is substantially higher than in another, potentially causing the model to favor the majority class and perform poorly in identifying the minority class [7]. One approach to address this problem is the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic observations for the minority class to obtain a more balanced class distribution [8], [9]. In addition to class imbalance, the performance of random forest can be affected by the selection of hyperparameters [10]. Random search is one approach for optimizing hyperparameters by randomly selecting combinations of parameter values from a predefined search space [11]. Therefore, SMOTE and random search can serve complementary roles in addressing class imbalance and optimizing model configuration, respectively.

The use of SMOTE and random search has demonstrated promising results in previous classification studies. [12] showed that the application of SMOTE to random forest improved classification performance on imbalanced data. Meanwhile, [13] reported that random search optimization improved the performance of random forest for stunting prediction, with accuracy increasing from 90.7% to 96.33%. These findings indicate that both approaches can contribute to improving model performance. However, their application in the context of food security still requires further investigation, particularly in integrating SMOTE to address class imbalance and random search to optimize random forest hyperparameters for classifying the food security levels of regencies/cities in Indonesia in 2024.

Based on the above discussion, this study aims to classify the food security levels of regencies/cities in Indonesia in 2024 using random forest with SMOTE and hyperparameter optimization through random search. In addition, this study aims to evaluate the performance of the resulting model based on accuracy, precision, recall, and F1-score.

Methodology

The type of research used in this study is non-experimental research, as the researcher did not apply any treatment to the research objects when obtaining the data, but instead used existing data. This study was conducted by collecting food security level data from the 2024 FSVA Indonesia dataset obtained from the National Food Agency (Badan Pangan Nasional), which was then used as the research object.

Research Dataset

The population in this study consists of regencies and cities in Indonesia. The sample comprises 514 regencies/cities in Indonesia in 2024. The research variables consist of a response variable and several predictor variables, which are presented in Table 1.

Table 1. Research variables

Variable	Description	Measurement Scale	Unit
Y	Food security level 0 = food-insecure 1 = food-secure	Nominal	-
X_1	Percentage of the population below the poverty line	Ratio	Percent
X_2	Percentage of households allocating more than 65% of their expenditure to food	Ratio	Percent
X_3	Percentage of households without access to electricity	Ratio	Percent

X_4	Percentage of households without access to clean water	Ratio	Percent
X_5	Average years of schooling among women aged over 15 years	Ratio	Years

Variable	Description	Measurement Scale	Unit
X_6	Population-to-health-worker ratio	Ratio	People/Health Worker
X_7	Life expectancy	Ratio	Years
X_8	Stunting	Ratio	Percent

Data analysis was conducted using the random forest method, with RStudio used for data processing and Python for visualizing the resulting decision trees. The analysis stages included:

1. Performing descriptive statistical analysis of the research variables.
2. Dividing the data into training and testing sets with proportions of 80:20 and 90:10.
3. Detecting class imbalance in the training data using the Imbalance Ratio (IR), which is calculated using the following equation.
$$IR = \frac{C_{maj}}{C_{min}} \quad (1)$$
Where C_{maj} represents the number of observations in the majority class, and C_{min} represents the number of observations in the minority class [14].
4. Addressing class imbalance using SMOTE with K=5, where K represents the number of nearest neighbors used to generate synthetic data based on the following equation.
$$x_{syn} = x_r + (x_{nn} - x_r) \times \delta \quad (2)$$
Where x_{syn} is the synthetic data generated, x_{nn} is the duplicated data, x_r is the nearest neighbor of x_{nn} and δ is a random value between 0 and 1 [15].
5. Applying random search to evaluate combinations of $mtry$, the number of trees (a), node size, and max node. The $mtry$ parameter was determined based on the number of predictor variables, as shown in Equation (3)
$$mtry = \begin{cases} {}^1/_2\sqrt{m} \\ \sqrt{m} \\ 2\sqrt{m} \end{cases} \quad (3)$$
Where m represents the total number of predictor variables [16]. The combination with the lowest classification error was selected as the optimal hyperparameter configuration.
6. Constructing the final random forest model using the optimal hyperparameters and performing predictions on the testing data.
7. Evaluating model performance using accuracy, precision, recall, and F1-score.
8. Determining the best-performing model and drawing conclusions.

Results and Discussions

Results

Dataset Characteristics

The characteristics of the research data are presented using descriptive statistics, including the response and predictor variables. An overview of the food security levels of regencies/cities in Indonesia in 2024, classified into two categories, is presented in Figure 1.

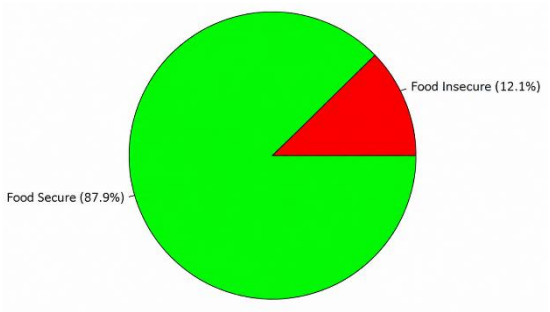

Figure 1. Distribution of food security levels

Based on Figure 1, of the 514 regencies/cities in Indonesia, 452 regencies/cities (87.9%) were classified as food secure, while 62 regencies/cities (12.1%) were classified as food insecure. The green color represents the food secure category, whereas the red color represents the food insecure category.

After presenting the descriptive statistics of food security levels as the response variable, the descriptive statistics of the predictor variables are subsequently presented. The results of the descriptive statistical analysis of the predictor variables are presented in Table 2.

Table 2. Descriptive statistics of predictor variables

Variable	Max	Min	Mean	Standard Deviation
X_1	40.01	2.27	11.50	7.15
X_2	93.04	1.25	24.26	14.36
X_3	79.03	0	2.38	8.03
X_4	99.62	0	28.61	20.05
X_5	12.74	1.28	8.85	1.55
X_6	86.82	0	4.03	8.74
X_7	77.93	55.72	70.20	3.39
X_8	50.20	0	22.35	9.17

Based on Table 2, the conditions of each regency/city vary across the predictor variables. The minimum value of 0 found in several variables does not indicate that the respective condition is completely absent, but may result from very small values being rounded to 0 in the data presentation.

Training and Testing Data Split

The training data are used to build the model, while the testing data are used to evaluate the resulting model. This study uses a total of 514 regencies/cities, which are divided into training and testing data using two proportion scenarios, namely 80:20 and 90:10. The data are randomly divided without replacement, so that each regency/city can only be selected once and assigned to either the training or testing data. This approach ensures that no regency/city appears in both the training and testing datasets. Based on the calculation using the 80:20 proportion, 411 data were used as training data and 103 data as testing data. Meanwhile, using the 90:10 proportion, 463 data were used as training data and 51 data as testing data.

Imbalanced Class and SMOTE

To identify the presence of class imbalance, the Imbalance Ratio (IR) is calculated based on Equation (1). In this study, the majority class is food secure (1), while the minority class is food insecure (0). Based on the calculation, the Imbalance Ratio for the 80:20 data proportion is obtained as follows:

$$IR_1 = \frac{n_1}{n_0} = \frac{361}{50} = 7{,}22,$$

the results of the calculation for the 90:10 data proportion are as follows:

$$IR_1 = \frac{n_1}{n_0} = \frac{361}{50} = 7{,}22.$$

SMOTE was applied only to the training data, while the testing data remained unchanged to ensure an unbiased evaluation of the model. Based on Equation (2), synthetic observations were generated by selecting an observation from the minority class and one of its nearest neighbors, then creating a new observation between the two observations using a random value between 0 and 1. For the 80:20 data split, the initial training set consisted of 411 observations. After applying SMOTE, the synthetic observations were combined with the original training data, resulting in 722 observations. Meanwhile, for the 90:10 data split, the initial training set consisted of 463 observations and increased to 814 observations after SMOTE was applied. Thus, SMOTE increased the number of minority-class observations and produced a more balanced class distribution in the training data.

Random Search Hyperparameter Optimization

Random search was employed to identify the optimal combination of hyperparameters for the random forest method with the aim of improving model accuracy. This method randomly selects combinations of predefined hyperparameter values. One of the hyperparameters used in random forest is the number of predictor variables randomly selected at each model construction ($mtry$), which is calculated based on Equation (3) as follows:

$$mtry = \begin{cases} {}^{1}/_{2}\sqrt{8} = 1{,}41 \approx 1 \\ \sqrt{8} = 2{,}83 \approx 3 \\ 2\sqrt{8} = 5{,}66 \approx 6 \end{cases}$$

Each $mtry$ value obtained was subsequently used to construct a number of trial trees. In this study, the number of trees (a) used was 100, 500, and 1,000. In addition to the random selection of predictor variables and the determination of the number of trees, other constraints can also be applied, such as the minimum number of observations in a node ($node\ size$) and the maximum number of nodes ($max\ node$). The combination of hyperparameter values used in the random forest process is presented in Table 3.

Table 3. Hyperparameter random forest

Hyperparameter	Values
$mtry$	1, 3, 6
a	100, 500, 1000
$node\ size$	1, 5, 10
$max\ node$	10, 20, 30

Based on the hyperparameter combinations presented in Table 3, 20 hyperparameter combinations were evaluated using the mean accuracy obtained from each combination. For the 80:20 data split, the best-performing combination consisted of $mtry = 6$, $ntree = 1000$, $node\ size = 1$, and $max\ nodes = 30$, achieving a mean accuracy of 92.80%. Meanwhile, for the 90:10 data split, the optimal combination consisted of $mtry = 6$, $ntree = 1000$, $node\ size = 5$, and $max\ nodes = 30$, achieving a mean accuracy of 93.49%. The optimal hyperparameter combinations were subsequently used to construct the random forest models for each data-splitting scenario.

Classification Performance of the Random Forest with SMOTE and Random Search

After prediction using the 80:20 data split with the random forest method and the optimal hyperparameters, namely $mtry = 6$, $ntree = 1000$, $node\ size = 1$, and $max\ nodes = 30$, the classification results were evaluated using a confusion matrix. Category 0 represents food insecurity, while category 1 represents food security. The confusion matrix presents the distribution of correctly and incorrectly classified regencies/cities based on their actual and predicted categories. The resulting confusion matrix is presented in Table 4.

Table 4. Confusion Matrix for the 80:20 Data Split

		Classification Category		Total
		Food Insecurity (0)	Food Security (1)	
Actual Category	Food Insecurity (0)	10	10	20
	Food Security (1)	2	81	83
Total		12	91	**103**

Based on Table 4, the confusion matrix shows that 10 regencies/cities were correctly classified as food-insecure, while 81 regencies/cities were correctly classified as food-secure. Meanwhile, 2 regencies/cities classified as food-insecure were predicted as food-secure, and 10 regencies/cities classified as food-secure were predicted as food-insecure.

Following the evaluation of the classification results using the 80:20 data split, the model was subsequently tested using a 90:10 data split with the optimal hyperparameters, namely $mtry = 6$, $ntree = 1000$, $node\ size = 5$, and $max\ nodes = 30$. The classification results were similarly evaluated using a confusion matrix, as presented in Table 5.

Table 5. Confusion Matrix for the 90:10 Data Split

		Classification Category		Total
		Food Insecurity (0)	Food Security (1)	
Actual Category	Food Insecurity (0)	3	5	8
	Food Security (1)	3	40	43
Total		12	6	**51**

Based on Table 5, the classification results of the food security levels of regencies/cities in Indonesia in 2024 using the random forest method show that 3 regencies/cities were correctly classified as food-insecure, while 40 regencies/cities were correctly classified as food-secure.

These classification results were further evaluated to assess the overall performance of the model using several classification metrics. The random forest with SMOTE and random search was applied to classify food security levels across regencies/cities in Indonesia in 2024. Model performance was compared under two data-splitting scenarios, namely 80:20 and 90:10, using four classification metrics: accuracy, precision, recall, and F1-score. The evaluation results are presented in Table 6.

Table 6. Random forest classification performance

Data Proportion	Accuracy	Precision	Recall	F1-score
80:20	88.35%	83.33%	50%	62.50%
90:10	84.31%	50%	37.50%	42.86%

Based on Table 6, the classification results obtained using the random forest method with SMOTE and Random Search showed different performance under the 80:20 and 90:10 data-splitting proportions. Overall, the models achieved reasonable classification performance under both data-splitting scenarios. However, the relatively lower precision, recall, and F1-score compared with the accuracy indicate that the model's ability to correctly identify food-secure regencies/cities was not yet optimal. In particular, the recall results indicate that some regencies/cities belonging to the food-secure class were misclassified as food-insecure. This condition also resulted in relatively lower F1-scores, indicating that the balance between precision and recall remained suboptimal.

The 80:20 data-splitting proportion consistently achieved higher accuracy, precision, recall, and F1-score than the 90:10 proportion. This indicates that, after addressing class imbalance using SMOTE and optimizing the random forest hyperparameters through random search, the model trained with the 80:20 proportion demonstrated better overall classification performance. Therefore, the 80:20 data-splitting proportion provided the most favorable performance for classifying food security levels across regencies/cities in Indonesia in 2024.

Discussions

The results showed that the random forest method with hyperparameter optimization using random search and class imbalance handling using SMOTE achieved relatively good classification performance. With an 80:20 training-to-testing data split, the model achieved an accuracy of 88.35% and a precision of 83.33%. indicating that the model was able to provide accurate predictions for most of the data. These results are consistent with the study by [13], which demonstrated that hyperparameter optimization using random search could improve the performance of random forest. The findings are also consistent with [12], who showed that the application of SMOTE could improve classification performance on imbalanced data.

Despite achieving relatively high accuracy and precision, this study still has limitations in terms of recall and F1-score. A recall of 50% indicates that the model was not yet able to optimally identify all observations in the target class, resulting in several regencies/cities being misclassified. This condition also affected the F1-score, which reached only 62.50%, as the F1-score considers the balance between precision and recall. Therefore, the relatively high accuracy and precision do not fully indicate that the model can optimally identify all members of the target class. Improving the model's ability to correctly identify observations that are still misclassified is therefore an important aspect to consider.

Conclusion

Based on the results of the analysis and discussion, the application of the random forest method with SMOTE and hyperparameter optimization using random search for classifying the food security levels of regencies/cities in Indonesia in 2024 demonstrated different performance across the two training-to-testing data-splitting proportions. Under the 80:20 split, the model correctly classified 10 regencies/cities as food-insecure and 81 as food-secure, achieving an accuracy of 88.35%, precision of 83.33%, recall of 50.00%, and F1-score of 62.50%. Meanwhile, the 90:10 split achieved an accuracy of 84.31%, precision of 50.00%, recall of 37.50%, and F1-score of 42.86%. The higher values across all evaluation metrics indicate that the 80:20 split provided better overall classification performance than the 90:10 split. Therefore, the random forest model with SMOTE and random search using the 80:20 training-to-testing data split was identified as the best-performing model for classifying food security levels across regencies/cities in Indonesia in 2024.

Acknowledgments

The authors would like to express their gratitude to the Department of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, for the academic support provided throughout this research. The authors also thank the National Food Agency (Badan Pangan Nasional) for providing the Food Security and Vulnerability Atlas (FSVA) data used in this study.

References

[1] A. B. Darmawan, A. R. Sulistyaning, J. Santoso, T. Linggarwati, K. Saadah, dan R. A. Dwianto, "Implementasi Kebijakan SDGs Pemerintah Daerah dalam Mengelola Ketahanan Pangan pada Masa Pandemi Covid-19 (Studi Kasus Desa Pandak, Kec. Baturaden, Kab. Banyumas)," *J. Ketahanan Nas.*, vol. 29, no. 2, hal. 145–165, 2023.

[2] T. Aguilera dan Y. A. Jatmiko, "Pemodelan Tingkat Kerawanan Pangan Rumah Tangga di Indonesia Tahun 2021 dengan Pendekatan Regresi Logistik Ordinal," *Indones. J. Appl. Stat.*, vol. 5, no. 2, hal. 90–108, 2022.

[3] Badan Pangan Nasional, "Peta Ketahanan dan Kerentanan Pangan." [Daring]. Tersedia pada: https://fsva.badanpangan.go.id/

[4] F. S. Pamungkas, B. D. Prasetya, dan I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python," in *PRISMA, Prosiding Seminar Nasional Matematika*, 2020, hal. 689–694.

[5] R. D. Marzuq, S. A. Wicaksono, dan N. Y. Setiawan, "Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree," *J. Pengemb. Teknol. Inf. dan*

Ilmu Komput., vol. 7, no. 7, hal. 3448–3456, 2023.
[6] V. A. Handayani, S. Rahmiati, B. Varischa, dan A. Arrafi, "Comparative Analysis : Multiple Regression and Random Forest Regression in Predicting Food Security Index in Indonesia," *J. Math. Comput. Stat.*, vol. 8, no. 2, hal. 389–400, 2025.
[7] N. N. Sholihah dan A. Hermawan, "Implementation of Random Forest and Smote Methods for Economic Status Classification in Cirebon City," *J. Tek. Inform.*, vol. 4, no. 6, hal. 1387–1397, 2023.
[8] Y. A. Sir dan A. H. H. Soepranoto, "Pendekatan Resampling Data untuk Menangani Masalah Ketidakseimbangan Kelas," *J-ICON (Jurnal Komput. dan Inform.*, vol. 10(1), hal. 31–38, 2022, doi: 10.35508/jicon.v10i1.6554.
[9] A. Syukron, Sardiarinto, E. Saputro, dan P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *J. Teknol. Inf. dan Terap.*, vol. 10, no. 1, hal. 47–50, 2023.
[10] G. N. Elwirehardja, T. Suparyanto, dan B. Pardamean, *Pengenalan Konsep Machine Learning Untuk Pemula*. Yogyakarta: Yogyakarta: Instiper Press, 2023.
[11] Y. Azhar, G. A. Mahesa, dan M. C. Mustaqim, "Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, hal. 15–21, 2021, doi: 10.14710/jtsiskom.2020.13790.
[12] V. Oktaviani, N. Rosmawarni, dan M. P. Muslim, "Perbandingan Kinerja Random Forest dan Smote Random Forest dalam Mendeteksi dan Mengukur Tingkat Stres pada Mahasiswa Tingkat Akhir," *J. Inform.*, vol. 20, no. l, hal. 43–49, 2024.
[13] A. A. Reza dan M. S. Rohman, "Prediction Stunting Analysis Using Random Forest Algorithm and Random Search Optimization," *JITE (Journal Informatics Telecommun. Eng.*, vol. 7, no. 2, hal. 534–544, 2024.
[14] C. S. Touazi, I. Ahriz, N. Niang, dan A. Piperno, "Analisis Perbandingan Teknik Oversampling SMOTE," *J. Perangkat Lunak dan Sist. Komun.*, vol. 21, no. 2, hal. 132–143, 2025.
[15] R. Fatiya, N. Yusliani, M. D. Marieska, dan D. M. Saputra, "Pengaruh Synthetic Minority Oversampling Technique pada Analisis Sentimen Menggunakan Algoritma K-Nearest Neighbors," *J. Linguist. Komputasional*, vol. 5, no. 1, hal. 7–12, 2022.
[16] S. Mahmuda, "Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube," *J. Jendela Mat.*, vol. 2, no. 01, hal. 21–31, 2024, doi: 10.57008/jjm.v2i01.633.