JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 9 Number 1 | June 2024: 128-137
DOI: 10.15575/join.v9i1.1012

Performance Comparative Study of Machine Learning Classification Algorithms for Food Insecurity Experience by Households in West Java

Khusnia Nurul Khikmah[1], Bagus Sartono[2], Budi Susetyo[3], Gerry Alfa Dito[4]
[1,2,3,4]Department of Statistics, IPB University, Bogor, Indonesia 16680

Article Info

Article history:

Received January 03, 2023
Revised October 27, 2023
Accepted April 21, 2024
Available Online June 25, 2024

Keywords:

Extremely randomized tree
Food insecurity
Gradient boosting
Random forest
Rotation forest

ABSTRACT

This study aims to compare the classification performance of the random forest, gradient boosting, rotation forest, and extremely randomized tree methods in classifying the food insecurity experience scale in West Java. The dataset used in this research is based on the Socio-Economic Survey by Statistics Indonesia in 2020. The novelty of this research is comparing the performance of the four methods used, which all are the tree ensemble approaches. In addition, due to the imbalance class problem, the authors also applied three imbalance handling techniques in this study. The results show that the combination of the random-forest algorithm and the random-under sampling technique is the best classifier. This approach has a balanced accuracy value of 65.795%. The best classification method results show that the food insecurity experience scale in West Java can be identified by considering the factors of floor area (house size), the number of depositors, type of floor, health insurance ownership status, and internet access capabilities.

Corresponding Author:

Khusnia Nurul Khikmah, S.Si.
Department of Statistics, IPB University,
Jl. Raya Dramaga, Babakan, Kecamatan Dramaga, Kabupaten Bogor, Jawa Barat, Indonesia 16680
Email: khusniank@gmail.com

1. INTRODUCTION

Statistics is one of the sciences where the development of analytical methods is swift. This development is supported by its utilization in various fields, such as the economic, health, and social populations. On the other hand, this development also requires researchers, especially statisticians, to conduct further research related to comparing the performance of various methods to obtain information regarding their best performance to achieve research objectives, especially in making decisions and policies. The rapid development of statistical methods in the last twenty years has occurred in the classification method. Starting from the approach method using standard methods, such as k-nearest neighbors [1], to approaches using machine learning, such as random forests [2], and support vector machines [3]. Therefore, it is crucial to select the method with the best performance to be used in the analysis. Statisticians have extensively researched the classification analysis approach using various machine learning methods with different results, such as classification trees, bagging, random forests, rotation forests, extremely randomized trees, and double random forests. Leo Breiman introduced the random forest method in 2001. Based on previous research, the random forest method is claimed to be the best method compared to the decision tree method and convolutional neural network [2]. In addition, random forest is also claimed to be the best classification method compared to support vector machines and artificial neural networks [4].

Another classification approach which is also claimed to be the classification method with the best performance, is gradient boosting. Gradient boosting is a method with ensemble techniques from decision trees introduced by Friedman in 2001. Based on research [5], gradient boosting has better classification performance than k-nearest neighbors, support vector machines, and random forests. In addition, research results [6] claim that gradient boosting has better classification performance than support vector machines, decision trees, and multilayer perceptrons. The best classification method besides the random forest and gradient boosting, in its development, many new techniques are used and with excellent classification performance. Like a rotation forest, namely a random forest using an ensemble technique and principal component analysis to rotate the variable axes to build a decision tree [7]. Extremely randomized trees, namely a classification method with ensemble techniques with high or extreme randomness. Therefore, this study examines the best performance of random forest, gradient boosting, rotation forest, and extremely randomized tree methods.

The real problems we often face are related to the use of classification methods, one of which is in the social field of population, namely the incidence of food insecurity. Incidents of the Food Insecurity Experience Scale/FIES are one of the most critical issues in the social field of the population, where the end goal is sustainable development [8]. This sustainable development focuses on providing food security so that cases of hunger can be handled. Therefore, analyzing issues related to the food insecurity experience scale (FIES) is crucial.

2. METHOD

This research is proposed to compare the performance of machine learning classification models and data imbalance problems by applying the resampling method. In general, the flow of this research is shown in Figure 1.

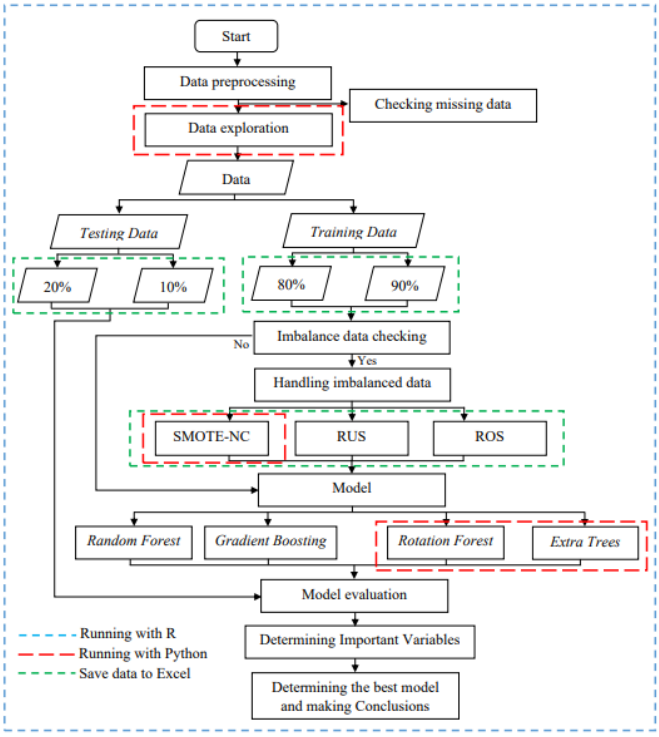

Figure 1. Flowchart of research analysis procedures

2.1. Research

This study used food insecurity experience scale (FIES) data from West Java Province in 2020. The data used is secondary data taken from the Central Bureau of Statistics. The initial data used is 19902 observations. There are 24 explanatory variables and one response variable. The 24 explanatory variables used in the study were the education of the head of the household, the head of the vulnerable household, the number of savers, the number of illiterates, transfer recipients, ownership of land assets, use of internet access, illness but not outpatient care, family hope program, family card. Prosperous, non-cash food assistance, assistance from the local government, health BPJS, regional health insurance, innovative Indonesia program, type of roof, type of floor, type of wall, floor area, electricity, cooking fuel, source of drinking water, proper drinking water, and proper sanitation.

2.2. Random Forest

Random forest is a classification method that combines ensemble techniques with bagging. Breiman first proposed this method in 2001 [9], which is one method that is easy to use and quite effective in classifying. This method generally uses a bootstrapping technique to randomly select n samples from training data to generate new training data samples and return them randomly to train a decision tree. Previous research states that the random forest method has several advantages, such as being effective even if the data used has outliers [10], has better classification performance than the support vector machine method, k-nearest neighbor [11].

Random forest is a popular classification method compared to others because, in its application, it only requires optimization of two parameters, namely $ntree$ and $mtry$. Previous research [12]–[14] showed that the best number of trees ($ntree$) recommended for use in the analysis is 100 trees. Meanwhile, the $mtry$ is the square root of the number of independent variables. In general, the stages of classification using the random forest method are as follows [9], [11], [15]:

1. Determine the number of trees to be formed (k)
2. Perform random sampling with a return to training data of n conservation for each tree
3. Take a random subset of m explanatory variables on each tree, where if p is the number of explanatory variables then $m < p$
4. Repeat steps two and three for k trees
5. The random forest prediction results are calculated using a majority vote from the classification results of k trees.

2.3. Gradient Boosting

Gradient boosting is one of the classification methods with ensemble techniques; where this method was first introduced in 2001 by Freidman [16]. This method was developed with the aim of getting the best classification performance and overcoming the weaknesses of other classification methods. Gradient boosting is built based on a decision tree algorithm using an ensemble technique [6]. Based on previous research, gradient boosting has several advantages, which are resistant to data outliers, have good predictive results, and can be used for all types of data. However, gradient boosting is one of the methods which is sequential in nature, so it has deficiencies in the efficiency of computation time.

Previous studies have stated that gradient boosting has good classification performance, as [17] shows that gradient boosting has better performance than random forests, deep neural networks, and support vector machines. Research [6] also shows that gradient boosting has better performance than support vector machines, k-nearest neighbors, and multilayer perceptrons. Friedman 2001 showed several stages of the working gradient boosting algorithm, namely [16]:

1. Input as much as n training data, where $(x_i, y_i)_{i=1}^{n}$ where y is the response variable with two classes
2. Determine the number of iterations (M)
3. Determine the value of learning rate v, $v \in (0,1)$
4. Determine the loss function (L), where $L\big(y_i, F(x)\big)$
5. Determine the model of the base learner $h(x, \theta)$
6. Initialize the value of $F_0(x)$, where:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

7. For $t = 1$ to M, then:
   a. Calculate the pseudo residue r_{it} for $i = 1,2, \dots, n$, where:

$$r_{it} = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{t-1}(x)}$$

   b. Defines weak learner $h_t(x)$ to be a pseudo residual. Then train the residual pseudo $\{(x_i, r_{it})\}_{i=1}^{n}$
   c. Calculates the value of γ_t, where:

$$\gamma_t = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{t-1}(x_i) + \gamma h_t(x)$$

   d. Update the $F_t(x)$ model, where:

$$F_t(x) = F_{t-1}(x) + \gamma_t h_t(x)$$

e. Output: $F_M(X)$.

2.4. Rotation Forest

Another classification method which is also a development of the previous method, is the rotation forest. Rotation forest is a combined tree method or ensemble classification using principal component analysis to rotate the variable axes to build decision trees [7]. The principal component analysis used to build a decision tree must maintain the completeness of data information. In addition, the primary purpose of using principal component analysis in the rotation forest is only to rotate the variables. This method was first introduced in 2006 by Rodriguez; based on previous research, the rotation forest has several advantages. Namely, it is a development method of bagging and random forest by applying principal component analysis. This method simultaneously increases the accuracy and diversity of each classifier in the ensemble system, rotation forests applying principal component analysis can build decision trees that are independent of each other, and with ten decision trees can produce optimal modelling [7], [18]. Many previous studies stated that the rotation forest has good classification performance, as [19] that the rotation forest produces better accuracy than other classifier ensembles. Rotation forest produces competitive performance compared to random forest, and rotation forest produces more accurate accuracy than AdaBoost and random forest [7]. In general, the rotation forest has algorithm stages which are explained in detail as follows:

Suppose $x = [x_1, \dots, x_p]^T$ is a data point of p variables, and $\boldsymbol{X}$ is a data set consisting of training data in the form of an $N \times p$ matrix. Suppose $y = [y_1, \dots, y_p]^T$ is a vector with class labels on the data. The classifiers in this method are denoted by $D_1, \dots, D_L$ and $\boldsymbol{F} = (\boldsymbol{x}, \boldsymbol{y})^{\boldsymbol{T}}$ as variable clusters. As with other classification methods, in a rotation forest, it is necessary to determine the number of trees to be built, namely L, and then all classifiers can be trained together. The steps taken to form a D_i decision tree; $i = 1, 2, \dots, L$ (Rodríguez et al., 2006):

1. Divide $\boldsymbol{F}$ into K subsets randomly so that each subset has nearly the same number of variables (M_j).
2. For $j = 1$ to K:
   a. Randomly select a class subset
   b. Remove observations on $X_{i,j}$ corresponding to the selected class (eg $X^*_{i,j}$)
   c. Take the bootstrap observation example from $X^*_{i,j}$ then notate it with $X'_{i,j}$
   d. Perform principal component analysis on $X'_{i,j}$ then store principal component coefficients in $a^{(1)}_{i,j}, a^{(2)}_{i,j}, \dots, a^{(M_j)}_{i,j}$
3. Arrange the obtained principal component coefficient vectors into the rotation matrix $\boldsymbol{R}_i$

$$\boldsymbol{R}_i = \begin{bmatrix} \boldsymbol{a}^{(1)}_{i,j}, \boldsymbol{a}^{(2)}_{i,j}, \dots, \boldsymbol{a}^{(M_1)}_{i,j} & [\boldsymbol{0}] & \cdots & [\boldsymbol{0}] \\ [\boldsymbol{0}] & \boldsymbol{a}^{(1)}_{i,j}, \boldsymbol{a}^{(2)}_{i,j}, \dots, \boldsymbol{a}^{(M_2)}_{i,j} & \cdots & [\boldsymbol{0}] \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{0}] & [\boldsymbol{0}] & \cdots & \boldsymbol{a}^{(1)}_{i,j}, \boldsymbol{a}^{(2)}_{i,j}, \dots, \boldsymbol{a}^{(M_K)}_{i,j} \end{bmatrix}$$

4. Rearrange the columns in $\boldsymbol{R}_i$ so that they match the original arrangement of the variables and then save it as $\boldsymbol{R}^a_i$
5. Construct the ith decision tree (D_i) using $(\boldsymbol{X}\boldsymbol{R}^a_i, Y)$
6. Repeat steps 1 to 5 until L decision trees are obtained.

2.5. Extremely Randomized Trees

Extremely randomized trees are a classification method with an ensemble technique first introduced by Geurts in 2006, where the algorithm used is a combination of single trees with extreme randomization [20]. Randomization in this method is done when selecting explanatory variables and determining cut points to separate nodes, where the best cut-point resulting from a node is the result of evaluating the Gini coefficient and the entropy of each variable value. In addition, each extremely randomized tree is formed using all complete training data, which aims to minimize model deviation [21].

In general, extremely randomized trees have two fundamental differences compared to classification methods with other ensemble techniques, such as a random forest. These two differences, namely, the random selection of cut points, determine the separation of nodes. The calculations are carried out using original data (training data), not from data from repeated trials. Therefore, several

previous studies stated that extremely randomized trees have good classification performance. Research [22] concluded that extremely randomized trees are better than the random forest and AdaBoost methods. Research [23] concluded that extremely randomized trees are better than support vector machines. In general, the extremely randomized trees algorithm uses all training data and has the following calcification stages [20]:

1. Stages of selecting the best splitting:
   a. Choose at random m independent variables
   b. Randomly choose k cut-points
   c. Determine the best splitting criteria
   d. Repeating steps a to c until it reaches the stopping criteria so that the prediction results from one tree are obtained
2. Repeating step 1 is repeated until M trees are formed
3. Combining the estimation results obtained from each classification tree using a majority vote.

2.6. Model Evaluation

Model evaluation is a research stage that aims to measure the accuracy of the classification results based on the method used. Thus, the performance of the method used on the classification results can be obtained in the form of numbers, namely the success ratio of classification. Evaluation of the model has several calculations for the value of classification accuracy where classification accuracy can be calculated using several calculations, such as balanced accuracy, sensitivity, and specificity based on the confusion matrix in Table 1 [24].

Table 1. The confusion matrix

Prediction	Actual	
	0	1
0	True Positive (TP)	False Positive (FP)
1	False Negative (FN)	True Negative (TN)

Sensitivity is: $Sensitivity = \frac{TP}{TP+FN}$

Specificity is: $Specificity = \frac{TN}{TN+FP}$

Balanced accuracy is: $balanced\ accuracy = \frac{sensitivity+specificity}{2}$

A method has good classification performance if its sensitivity and specificity values have higher values for the two response variables [25].

3. RESULTS AND DISCUSSION

The initial data used as research data for the Food Insecurity Experience Scale/FIES for West Java Province in 2020 has a response variable in detail with distribution in Table 2.

Table 2. Distribution of response variable

Variable	Description		Total	%
Y	Vulnerable Status	Vulnerable	4322	21.71641
		Not Vulnerable	15580	78.28359

This initial data is then divided into data training and testing data for further classification modelling. Based on the data sharing scenarios, the percentage of vulnerable status, namely the distribution of vulnerable and non-vulnerable, in each data training scenario, it was found that there was an imbalance in the data. Thus, the data training scenarios are handled by unbalanced data using three methods: the synthetic minority over-sampling technique for nominal and continuous, random under sampling, and random oversampling.

First, the data is modelled using a random forest, where hyperparameter tuning is done for modelling. This hyperparameter tuning process is carried out to get the best parameters. In addition, this study also uses 5-fold cross-validation in evaluating model performance. The 5-fold cross-validation evaluates the model as much as five repetitions of each parameter in the hyperparameter tuning process. The model with the best hyperparameters from each scenario of data sharing and handling of

unbalanced data is then evaluated by testing the model on data testing. Evaluation of this model is calculated based on the value of each model's balanced accuracy, sensitivity, and specificity.

The model evaluation results show that the best random forest model is generally obtained from unbalanced data handling by random under-sampling, where specifically in the scenario, the training data distribution is 90%. The model with this scenario has a balanced accuracy of 65.795%, meaning it can detect a vulnerable status of 65.795% with a sensitivity and specificity value of 68.750% and 62.840%. The sensitivity value of 68.750% means that the model error predicts a non-food insecure status even though the household is categorized as food insecure by 31.250%. While the specificity value is 62.840%, meaning that the ability of the model to give negative results to households that are not classified as food insecure is 62.840%. The best model obtained is categorized as good enough in the analysis to identify food insecurity status. In detail, the evaluation results of the random forest model from all scenarios and the best hyperparameter tuning are presented in Table 3.

Table 3. Evaluation value of the random forest model based on the results of the best hyperparameter tuning

Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
Unhandled	80	0.97946	0.06250	0.52098
	90	1.00000	0.00000	0.50000
SMOTE-NC	80	0.76220	0.37040	0.56630
	90	0.74968	0.33333	0.54150
RUS	80	0.62840	0.66320	0.64580
	90	0.62840	0.68750	0.65795
ROS	80	0.83472	0.24653	0.54062
	90	0.85045	0.25000	0.55022

The second model used to classify food insecurity status is gradient boosting. The data used for modelling is done by hyperparameter tuning first. The hyperparameter (ntree) tuning process is carried out to get the best parameters. In addition, this study also uses 5-fold cross-validation in evaluating the performance model. The 5-fold cross-validation evaluates the model as much as five repetitions of each parameter in the hyperparameter tuning process. The scenario of data sharing and unbalanced data handling used in model building is based on the best hyperparameter tuning results. The best hyperparameter tuning results from each scenario of data sharing and unbalanced data handling are then evaluated by model testing by testing the data. Evaluation of this gradient boosting model is calculated based on the value of each model's balanced accuracy, sensitivity, and specificity.

Evaluation of this gradient boosting model generally shows that the model with unbalanced data handling with random under-sampling is the best obtained. Specifically, the model scenario with unbalanced data handling with random under-sampling and distribution of 90% training data has the best accuracy. The model with this scenario has a balanced accuracy of 66.440%, meaning it can detect a vulnerable status of 66.440% with a sensitivity and specificity value of 73.380% and 59.500%. The sensitivity value of 73.380% means that the model error predicts the status of not being food insecure even though the household is categorized as food insecure at 26.620%. Meanwhile, a specificity value of 59.500% means that the ability of the model to give negative results to households that are not classified as food insecure is 59.500%. The best model obtained from these various scenarios is categorized as good enough in the analysis to identify food insecurity status. The results of the evaluation of the gradient boosting model from all scenarios and the best hyperparameter tuning in detail are presented in Table 4.

Table 4. Evaluation value of the gradient boosting model based on the results of the best hyperparameter tuning

Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
Unhandled	80	0.98716	0.03704	0.51210
	90	0.99101	0.03009	0.51055

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
SMOTE-NC	80	0.73813	0.40856	0.57334
	90	0.74005	0.43519	0.58762
RUS	80	0.59720	0.67360	0.63540
	90	0.59500	0.73380	0.66440
ROS	80	0.68740	0.55090	0.61915
	90	0.68420	0.57410	0.62915

The third model used for the analysis of food insecurity status is the rotation forest, where for this modelling, the number of trees built is determined at the start, namely ten trees (K=10), where this value is the result of previous research [18]. Data is modelled with initial parameters for each scenario of unbalanced data sharing and handling. Next, an evaluation of the model is carried out by testing the model on testing data. Evaluation of this model is calculated based on the value of each model's balanced accuracy, sensitivity, and specificity. Evaluation of the rotation forest model by testing data from various predefined scenarios shows that the best model generally obtained is the model with unbalanced data handling using the synthetic minority over-sampling technique for nominal and continuous. Specifically, the model with synthetic minority over-sampling technique for nominal and continuous unbalanced data handling and 80% training data sharing is the best.

The best model obtained from the evaluation results has a balanced accuracy value of 55.9035%, meaning that it can detect a vulnerable status of 55.9035% with a sensitivity and specificity value of 70.6033% and 41.2037%. The sensitivity value of 70.6033% means that the model error predicts the status of not being food insecure even though the household is categorized as food insecure at 29.3967%. While the specificity value is 41.2037%, meaning that the ability of the model to give negative results to households that are not classified as food insecure is 41.2037%. The best model obtained needs to be categorized as insufficient to identify food insecurity status. In detail, the evaluation results of the rotation forest model from all scenarios and the initial parameters are presented in Table 5.

Table 5. Evaluation value of the rotation forest model based on initial parameters

Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
Unhandled	80	0.309027	0.779525	0.544276
	90	0.277777	0.785622	0.531700
SMOTE-NC	80	0.386577	0.704428	0.545503
	90	0.412037	0.706033	0.559035
RUS	80	0.473379	0.398267	0.435823
	90	0.500000	0.401797	0.450899
ROS	80	0.743055	0.201861	0.472458
	90	0.773148	0.181001	0.477075

Extremely randomized trees are the final model used for food insecurity status analysis. The data modelled using extremely randomized trees is preliminarily tuned to the hyperparameter. This hyperparameter tuning process is carried out to get the best parameters. This study uses 5-fold cross-validation in evaluating model performance. The 5-fold cross-validation evaluates the model as much as five repetitions of each parameter in the hyperparameter tuning process.

Other parameters of the highly randomized trees model are determined at the outset, such as the number of trees built, as many as 100, with the value m (maximum variable), used, which is the root of n variables ($\sqrt{24}$). The initial parameters that have been set are used to model the data in each

scenario of data sharing and unbalanced data handling. Next, an evaluation of the model is carried out by testing the model on testing data. Evaluation of this model is calculated based on the value of each model's balanced accuracy, sensitivity, and specificity.

Table 6. Evaluation value of the extremely randomized trees model based on initial parameters

Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
Unhandled	80	0.27486	0.79223	0.533545
	90	0.29795	0.79426	0.546105
SMOTE-NC	80	0.28122	0.80781	0.544515
	90	0.29739	0.81267	0.555030
RUS	80	0.15291	0.70691	0.429910
	90	0.15246	0.70985	0.431155
ROS	80	0.20868	0.72455	0.466615
	90	0.21010	0.72925	0.469675

Based on Table 6, the evaluation value of the extremely randomized trees model by testing data from various predetermined scenarios shows that, in general, the best model obtained is the model with unbalanced data handling using the synthetic minority over-sampling technique for nominal and continuous. Specifically, the model with synthetic minority over-sampling technique for nominal and continuous unbalanced data handling and 90% training data sharing is the best. The best model obtained from the evaluation results has a balanced accuracy value of 55.5030%, meaning that it can detect a vulnerable status of 55.5030% with a sensitivity and specificity value of 81.267% and 29.739%. The sensitivity value of 81.267% means that the model error predicts the status of not being food insecure even though the household is categorized as food insecure at 18.733%. While the specificity value is 29.739%, meaning that the ability of the model to give negative results to households that are not classified as food insecure is 29.739%. The best model obtained is categorized as insufficient to identify food insecurity status.

The best modelling results using four models (random forest, gradient boosting, rotation forest, and extremely randomized trees) and the data distribution and handling scenarios described in section 4.2 are then selected for the best model based on the evaluation value. In detail, the best model of each model and its scenarios are presented in Table 7.

Table 7. The best model evaluation value

Model	Unbalanced Data Handling	Training Data	Specificity	Sensitivity	Balanced Accuracy
RF	RUS	90	0.62840	0.68750	0.657950
GB	RUS	90	0.59500	0.73380	0.664400
RoF	SMOTE-NC	90	0.41203	0.70603	0.559035
ET	SMOTE-NC	90	0.29739	0.81267	0.555030

Based on Table 7 above, the best model obtained is the random forest model with a training data distribution of 90% and handling imbalanced data using random under-sampling. Compared to other models, the best model was chosen based on sound sensitivity and specificity values and highly balanced accuracy. In addition, based on previous research that sensitivity and specificity values are suitable evaluation measures in looking at model performance and determining the best model based on hyperparameters and scenarios that have been made [26], [27]. The best model was obtained: random forest with unbalanced data handling with random under-sampling. Then further analysis was carried out, namely identifying essential variables to determine the contribution of each explanatory variable in predicting food insecurity status. The results of the follow-up analysis in Figure 2 show that the variables floor area, number of savers, type of floor, BPJS ownership status, and internet access

p-ISSN: 2528-1682
e-ISSN: 2527-9165

capabilities are the five variables with the highest ranking (important variables) that characterize the status of food insecurity events.

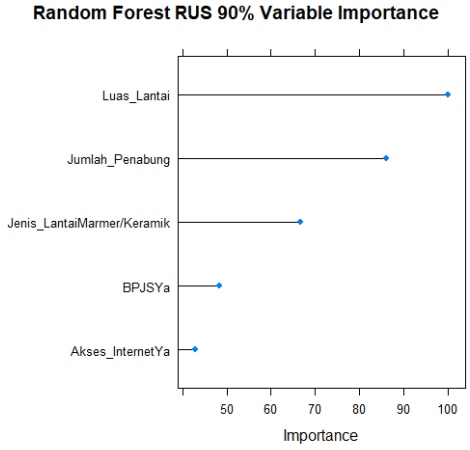

Figure 2. Important variables (top 5) results of the best random forest model

The best model obtained from a comparison of the four models in Table 7 above can detect food insecurity status in West Java Province with an accuracy of 65.795% with floor area, the number of depositors, floor type, BPJS ownership status, and internet access capabilities as variables. Characterize the status of food insecurity events. In addition to answering the comparison of the best models of the four methods used to add to the repertoire of knowledge and research, this research is also expected to become material for consideration by the authorities in making policies related to food insecurity in West Java Province. In addition, the results of this analysis can be used as a basis for classifying eligible households to receive assistance so that it is right on target based on households with food insecurity characteristics based on the analysis results. This research is also expected to solve one of the most critical issues in the social field of population, the end goal of which is sustainable development and providing food security in West Java Province so that cases of hunger can be handled.

4. CONCLUSION

The results of this study indicate that the random forest model is the model that has the best classification performance for food insecurity status data in West Java Province compared to the gradient boosting, rotation forest, and extremely randomized tree models. The best model obtained is a model with a training data-sharing scenario of 90%, and imbalanced data handling is carried out with random under-sampling. This model has an accuracy value of 65.795%. Next, the random forest model is analyzed for essential variables. The analysis of important variables shows that the variables floor area, number of savers, type of floor, BPJS ownership status, and internet access ability characterize food insecurity status in West Java Province. Analysis related to classification modelling with machine learning is then suggested to compare the use of k values in the k-fold cross-validation used. In addition, you can then use a classification model, such as a neural network, to compare model performance.

5. REFERENCES

[1] W. Xing and Y. Bei, "Medical health big data classification based on KNN classification algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2019.
[2] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, "A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification," *Advances in Space Research*, vol. 65, no. 8, pp. 2052–2061, 2020.
[3] A. R. Bagasta, Z. Rustam, J. Pandelaki, and W. A. Nugroho, "Comparison of cubic SVM with Gaussian SVM: classification of infarction for detecting ischemic stroke," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 546, no. 5, p. 052016.
[4] S. Talukdar, P. Singha, S. Mahato, S. Pal, Y.-A. Liou, and A. Rahman, "Land-use land-cover classification by machine learning classifiers for satellite observations—A review," *Remote Sens (Basel)*, vol. 12, no. 7, p. 1135, 2020.
[5] N. Chakrabarty, T. Kundu, S. Dandapat, A. Sarkar, and D. K. Kole, "Flight arrival delay prediction using gradient boosting classifier," in *Emerging technologies in data mining and information security*, Springer, 2019, pp. 651–659.
[6] Z. Tian, J. Xiao, H. Feng, and Y. Wei, "Credit risk assessment based on gradient boosting decision tree," *Procedia Comput Sci*, vol. 174, pp. 150–160, 2020.
[7] M. Juez-Gil, Á. Arnaiz-González, J. J. Rodríguez, C. López-Nozal, and C. García-Osorio, "Rotation Forest for Big Data," *Information Fusion*, vol. 74, 2021, doi: 10.1016/j.inffus.2021.03.007.

[8] M. Anwar, "The Household Food Insecurity Amidst the Covid-19 Pandemic in Indonesia," *JEJAK*, vol. 14, no. 2, pp. 244–260, 2021.
[9] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.
[10] M. Maniruzzaman *et al.*, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," *J Med Syst*, vol. 42, no. 5, pp. 1–17, 2018.
[11] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, p. 2809, 2020.
[12] F. Cánovas-García, F. Alonso-Sarría, F. Gomariz-Castillo, and F. Oñate-Valdivieso, "Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery," *Comput Geosci*, vol. 103, pp. 1–11, 2017.
[13] B. Ghimire, J. Rogan, V. R. Galiano, P. Panday, and N. Neeti, "An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA," *GIsci Remote Sens*, vol. 49, no. 5, pp. 623–643, 2012.
[14] T. N. Phan, V. Kuch, and L. W. Lehnert, "Land Cover Classification using Google Earth Engine and Random Forest Classifier—The Role of Image Composition," *Remote Sens (Basel)*, vol. 12, no. 15, p. 2411, 2020.
[15] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble machine learning*, Springer, 2012, pp. 157–175.
[16] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann Stat*, pp. 1189–1232, 2001.
[17] B. A. Tama and K.-H. Rhee, "An in-depth experimental study of anomaly detection using gradient boosted machine," *Neural Comput Appl*, vol. 31, no. 4, pp. 955–965, 2019.
[18] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A New classifier ensemble method," *IEEE Trans Pattern Anal Mach Intell*, vol. 28, no. 10, pp. 1619–1630, 2006, doi: 10.1109/TPAMI.2006.211.
[19] C. S. Septeria and L. Wachidah, "Klasifikasi Pasien Diabetes Melitus Tipe 1 dengan Metode Rotation Forest," *Prosiding Statistika*, pp. 521–529, 2021.
[20] Pd. Geurts, "Ernst D," *Wehenkel L. Extremely randomized trees. Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
[21] C. Désir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville, "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," *IEEE Trans Biomed Eng*, vol. 59, no. 9, pp. 2677–2683, 2012.
[22] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information*, vol. 11, no. 6, p. 332, 2020.
[23] G. Alfian *et al.*, "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," *Computers*, vol. 11, no. 9, p. 136, 2022.
[24] B. T. Pham *et al.*, "Intergration of Rotation Forest and MultiBoost Ensembles with Forest by Penalizing Attributes for Spatial Prediction of Landslide Susceptibility," 2022.
[25] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit*, vol. 91, pp. 216–231, 2019.
[26] J. H. Kranzler, R. G. Floyd, N. Benson, B. Zaboski, and L. Thibodaux, "Classification agreement analysis of cross-battery assessment in the identification of specific learning disorders in children and youth," *Int J Sch Educ Psychol*, vol. 4, no. 3, pp. 124–136, 2016.
[27] Y. Liu, J. Zhang, C. Gao, J. Qu, and L. Ji, "A sensitivity analysis of attention-gated convolutional neural networks for sentence classification," *arXiv preprint arXiv:1908.06263*, 2019.