Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 1 (2024) 33 - 41 e-ISSN : 2580-0760

Rotation Double Random Forest Algorithm to Predict The Food Insecurity Status of Households

Rais[1], Agus Mohamad Soleh[2], Budi Susetyo[3]
[1,2,3]Department of Statistics, Faculty of Mathematics and Science, IPB University, Bogor, Indonesia
[1]ipb21_rais@apps.ipb.ac.id, [2]agusms@apps.ipb.ac.id, [3]budisu@apps.ipb.ac.id

Abstract

The ensemble tree method has been proven to handle classification problems well. The strength of the ensemble tree technique lies in the diversity and independence between each tree. Increasing the diversity of mutually independent decision trees improves model performance. Various studies propose the development of ensemble tree-based models by forming algorithms that create decision trees that are formed independently of each other and have various inputs. These include random forest (RF), rotation forest (RoF), double random forest (DRF), and the latest is rotation double random forest (RoDRF). RoDRF rotates or transforms data intending to produce better diversity among the learner base. RoDRF applies the variable rotation concept to trees based on the DRF algorithm. Random rotations or transformations on different feature subspaces produce different projections, leading to better generalization or prediction performance. This research aims to compare the performance of RoDRF with RF, RoF, and DRF models on imbalanced data in cases of food insecurity. Class imbalance will be handled with two methods, namely EasyEnsemble and SMOTE-NC. The research results show that the DRF's model with EasyEnsemble techniques produces a model with the best performance among several algorithms tested. Even though the resulting accuracy is 0.62274 and the AUC value is 0.68501, the model can predict each class equally. All algorithms with EasyEnsemble treatment have average AUC values significantly different from each other based on statistical test results. This research also used SHAP to explain variables significantly contributing to the household's food insecurity status model.

Keywords: *rotation double random forest; easyensemble; SMOTE-NC; food insecurity*

How to Cite: Rais, Agus Mohamad Soleh, and Budi Susetyo, "Rotation Double Random Forest Algorithm to Predict The Food Insecurity Status of Households", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 1, pp. 33 - 41, Jan. 2024.
DOI: *https://doi.org/10.29207/resti.v8i1.5540*

1. Introduction

In applying various fields of science, several problems are often encountered, namely determining the appropriate group for an object. In statistics, this problem is included in the case of classification, where objects are predicted to fall into a specific class based on the similarity of the characteristics of the object with other objects. To simplify the classification process, we need an algorithm that can recognize the characteristics of an object and then group it into one class. Machine learning techniques benefit the classification process, especially in large data sets.

The decision tree is a machine learning algorithm commonly used in classification cases because of its simplicity and ease of interpretation, but it still has powerful performance [1]. However, decision trees are susceptible to disturbances in the diversity of input data. Hence, the decision tree model tends to be unstable because the predictions are only based on one decision tree. It impacts the high model variance despite low bias [2]. An ensemble tree can be used to overcome this weakness. One ensemble strategy for classification trees is to build different decision trees using different data clusters and variable subspaces [3]. The ensemble tree approach produces estimates from data that are a combination of estimates produced by several trees [4]. Combining estimates from several trees improves the classification model's performance.

[5] proposed an ensemble tree algorithm, random forest (RF), which constructs more diverse decision trees and is less correlated between trees to increase model accuracy. The RF algorithm applies the bootstrap sampling technique as primary data for each decision tree formed. Then, it selects a subset of variables to

Received: 10-11-2023 | Accepted: 12-01-2024 | Published Online: 30-01-2024

determine the best splitting criteria for each non-terminal node. Each prediction from the tree formed is then aggregated using the majority vote technique. Random forest algorithms have been shown to generally have better model accuracy performance than single classification trees [6]. Furthermore, [7] developed a new ensemble tree method, rotation forest (RoF), which builds mutually independent decision trees. The RoF algorithm builds a rotation matrix so that variables in the data will be transformed and rearranged into new data variables. As a result, the trees formed are more diverse, and the model produced from the rotation forest algorithm is more stable and accurate.

RF algorithm development continues to improve model performance. Among them is research by [8], who developed a new ensemble method: double random forest (DRF). This method is claimed to improve RF performance when the RF model experiences underfitting to data. The research results show that DRF is more accurate than RF on the 34 modelled datasets. The latest development was carried out by [9] by applying the rotation concept from the RoF algorithm to trees based on DRF to produce a rotation double random forest (RoDRF) model. This research compared 12 classification model algorithms tested on 121 datasets and produced the highest average classification accuracy performance of the RoDRF model compared to other algorithms. Applying the principle of variable rotation to a tree based on DRF forms an independent decision tree, thereby simultaneously increasing the diversity and accuracy of each classifier [9]. However, that research has yet to specifically explain the characteristics of the RoDRF algorithm for datasets with imbalanced response variable classes. Meanwhile, imbalanced data conditions can cause the model to be biased towards the majority class, resulting in underfitting or overfitting the model.

In many classification cases, the proportion of response variable classes between one class and another generally has an imbalanced condition or is known as imbalanced data. Imbalanced data handling has been proven to improve model accuracy performance in several studies. These include the use of the EasyEnsemble method in the RF model [10], the SMOTE (Synthetic Minority Oversampling Technique) method in the RoF model [11], and the SMOTE for Nominal and Continuous Features (SMOTE-NC) method in the RF model [12]. In this research, the problem of imbalanced data will be handled by comparing the use of the SMOTE-NC and EasyEnsemble methods.

Empirically, applying a decision tree-based classification algorithm with imbalanced data conditions can be carried out in cases of household food insecurity. Previous studies on food insecurity cases, including [13] and [14], have performed decision tree-based classification modelling with good model accuracy. Food insecurity is one of the crucial issues of the second goal of sustainable development: no hunger. According to data from Statistics Indonesia (BPS), in 2022, there will be 4.85 per cent of households experiencing moderate or severe food insecurity. This figure increased compared to the previous year, namely 4.79 per cent. Ironically, in several provincial areas with abundant natural food resources, including South Sulawesi Province, 3.78 per cent of households experienced moderate or severe food insecurity in 2022. Therefore, we want to apply the classification modelling algorithm to predict household food insecurity status in South Sulawesi Province.

Based on the description above, this research compares the performance of RF, RoF, DRF, and RoDRF models on imbalanced data in cases of household food insecurity in South Sulawesi Province.

2. Research Methods

2.1 Random Forest

The random forest (RF) algorithm is the most popular ensemble tree method. In forming a classification tree, RF applies bootstrap resampling techniques and selects a subset of variables to produce mutually independent trees. The final prediction is obtained by aggregating the predictions produced by each classification tree using a majority vote.

There are several steps in building a model with the RF algorithm. Suppose the number of classification trees to be formed is B. For each classification tree (b=1,…, B), take a random sample (bootstrap sampling) of size n without replacement from the training data (D) to obtain D_b^*. For each tree, perform a partition at the t-th node with the following conditions : (i) randomly select $m \approx \sqrt{p}$ or $m \approx p/3$ explanatory variables, (ii) determine the best splitting criteria, (iii) partition the data at the t-th node based on the splitting criteria in point ii. For each classification tree, repeat steps (i) to (iii) until the stopping criterion is reached to obtain estimation results from one tree. The prediction results from each classification tree are then aggregated with the majority vote using Formula 1.

$$C_B(x) = argmax_j \sum_{b=1}^{B} I(C_B(x) = j) \quad (1)$$

An illustration of the stages in the RF algorithm [5] can be seen in Figure 1.

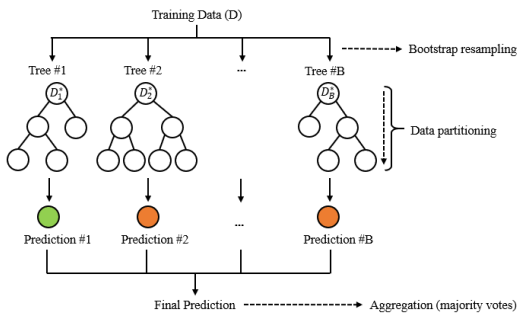

Figure 1. Illustration of the RF algorithm

2.2 Rotation Forest (RoF)

Rotation Forest (RoF) is an ensemble tree method that performs transformations on variables before forming a classification tree. On average, RoF performance will be more accurate if we use continuous variables as predictor variables [15] and cannot produce the best model performance if the predictor variable is dominated by categorical variables [16]. The RoF method uses Principal Component Analysis (PCA) to rotate the explanatory variable axes for which the decision tree will be built. The training data for each tree is formed by first dividing all explanatory variables into K subsets of variables. For each subset of variables, variable extraction will be carried out using PCA. The explanatory variables will be transformed through the formed rotation matrix and rearranged into new data variables to build a decision tree independent of each other [7].

There are several steps in building a model with the RoF algorithm. Suppose the number of classification trees to be formed is B. For each classification tree (b=1,…, B), form a rotation matrix $\boldsymbol{R_i^a}$ by separating the explanatory variable F into K subsets to become $F_{i,j}$ (for j = 1, …, K). For j = 1, …, K, randomly select a subset of classes, then delete the observations on $X_{i,j}$ according to the selected class to form $X_{i,j}^*$. Do a bootstrap of $X_{i,j}^*$ by 75% so that the new observation becomes $X_{i,j}'$. Perform PCA on $X_{i,j}'$ to obtain the matrix $C_{i,j}$ coefficients. Arrange the coefficients obtained into the rotation matrix R_i. Rearrange the columns in R_i to match the original arrangement of the variables, then save them as $\boldsymbol{R_i^a}$. Build the i-th decision tree using $\boldsymbol{XR_i^a, Y}$. All decision tree prediction results are aggregated to obtain the final RoF prediction using Formula 2.

$$\mu_j(x) = \frac{1}{B}\sum_{i=1}^{B} d_{i,j}(xR_i^a), \quad j = 1, \dots, c \qquad (2)$$

An illustration of the stages in the RoF algorithm can be seen in Figure 2.

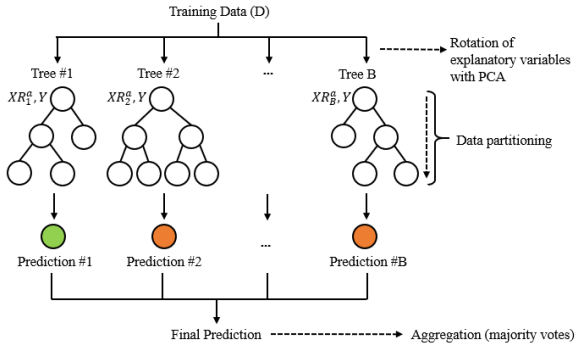

Figure 2. Illustration of the RoF algorithm

2.3 Double Random Forest (DRF)

The double random forest (DRF) algorithm is an alternative solution that can improve model performance when the RF-generated model experiences underfitting [8]. In contrast to RF, DRF can form more giant trees by using all the same training data at the root node of each tree. Bootstrapping is not only done once, like in RF, but at each node when determining the best splitting criteria during the decision tree formation process. As a result, the decision trees formed are more diverse from each other, so predictions from the model tend to be more accurate. An illustration of the stages in the DRF algorithm can be seen in Figure 3.

There are several steps in building a model with the DRF algorithm. Suppose the number of classification trees to be formed is B. For each classification tree (b=1,…, B), use all training data (D) on the root node. For each t-th node), If $n_t > n\,x\,0.1$, form a new data cluster (D_t^*) resulting from bootstrap sampling from D_t. Else, $D_t^* = D_t$. Randomly choose $m \approx \sqrt{p}$ or $m \approx p/3$ explanatory variables from D_t^*. Next, determine the best splitting criteria based on the data cluster D_t^*. Split the D_t data cluster based on the splitting criteria produced from D_t^*. Repeat the steps above until it reaches the stopping criterion so that estimation results are obtained from one tree. The prediction results for each tree are then combined (aggregating) with the majority vote.

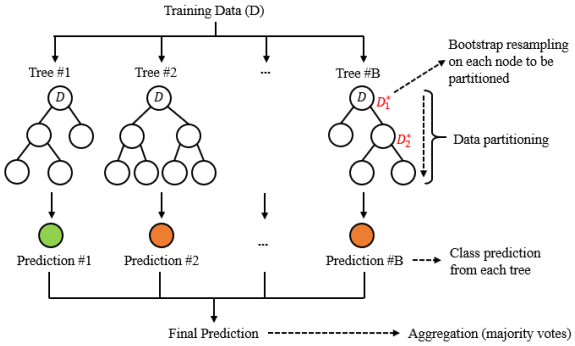

Figure 3. Illustration of the DRF algorithm

2.4 Rotation Double Random Forest (RoDRF)

Rotation Double Random Forest (RoDRF) rotates or transforms data intending to produce better diversity among the learner base. Random rotations or transformations on different feature subspaces produce different projections, leading to better generalization or prediction performance. Rotation is applied to each non-leaf node. RF and DRF algorithms use the concepts of random subspace and bagging to introduce diversity among the base learners of an ensemble. In the RoDRF algorithm, the variable rotation process (PCA) is applied to each non-leaf node to produce a more diverse base of learners. Apart from that, the RoDRF algorithm also has the concept of bagging at each non-leaf node, allowing for greater tree depth to improve performance [9].

There are several steps in building a model with the DRF algorithm. Use all training data D as basic data at the root node of each decision tree T_i, where $i = 1,2, \dots, B$. For each d-th node with training data D_d, if

$N_d > N\ x\ 0.1$, do bootstrap sampling on D_d to form D_d^*. Else, $D_d^* = D_d$. Define "$mtry" = \sqrt{p}$ as the number of variables used in D_d^*. Calculate the total distribution matrix S_d using D_d^*. Calculate all the characteristic roots of the matrix S_d, denoted as V. In the PCA process, do splitting with the best variable. Split D_d data based on splitting with the best variable. Repeat the steps above until the specified stopping criteria are reached. The prediction results for each tree are then combined (aggregating) with the majority vote.

An illustration of the stages in the RoDRF algorithm can be seen in Figure 4.

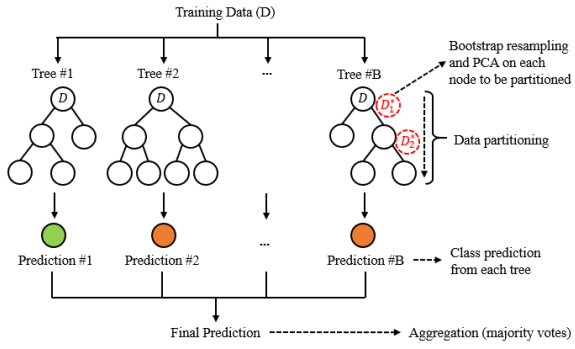

Figure 4. Illustration of the RoDRF algorithm

2.5 EasyEnsemble

EasyEnsemble is an undersampling technique for handling imbalanced data. EasyEnsemble is more than just ordinary undersampling. It repeatedly carries out the undersampling process to form several subsets of data so that every observation from the majority class can be represented. If we only take one subset of the majority class, there will likely be overlooked information from other unselected data [3].

The EasyEnsemble algorithm [10] can be described as follows. First, a data cluster of size S is defined, which consists of a minority class of size P and a majority class of size N where $|P| < |N|$. Second, if T is the number of trees to be created, form a subset N_j from N as many as T using random sampling where $|N_j| = |P|$ so the training data S_j will have balanced classes. The EasyEnsemble method generates T-balanced data subsets. Next, each subset will be modelled according to the previously defined classification model. The output of each subset is the predicted value of each model. Finally, all predicted values will be aggregated using a majority vote to obtain the final prediction.

2.6 Synthetic Minority Over-sampling Technique for Nominal and Continuous Features (SMOTE-NC)

Data imbalance is a condition where the number of observations in one class is much higher (majority class) than in another class (minority class). In machine learning algorithms, imbalanced data conditions can cause the model to be biased towards the majority class, resulting in underfitting or overfitting the model. The Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique that can balance data classes. SMOTE uses information on minority classes to form synthetic observations, thereby increasing the number of observations on minority classes [17]. SMOTE-NC is used if the variables in the data group are a combination of variables with nominal and continuous data types [11].

To generate new observations with SMOTE-NC, the median standard deviation was calculated for all continuous variables in the minority class. Distance calculations are carried out on continuous variables using Euclidean distance, including the median standard deviation that was calculated previously. For categorical variables, this is done by selecting the majority value of the k nearest neighbors [18].

2.7 Evaluation Metric

To measure and compare the performance of classification models, an ideal metric/measure is needed that fits the data's characteristics. One of the appropriate criteria that can be used to measure the performance of different classifiers over an imbalanced data set is the Receiver Operating Characteristic (ROC) graph [19]. From the ROC graph, the AUC (Area Under the Curve) value can be calculated, which states the ability of the classification model to differentiate objects between classes. An AUC value equal to 1 indicates that the model can perfectly differentiate between classes. AUC provides one measure of classifier performance to judge which model is better on average and has been widely used in many imbalanced scenarios [20].

Accuracy measures the model's accuracy in correctly predicting positive and negative classes. Accuracy values can be used to evaluate if the data is in balanced class conditions. The accuracy value of the classification model can be obtained using formula 3. In some cases of empirical classification, both positive and negative classes have the same importance to observe. So, the prediction results from a classification model are also important to consider classification accuracy in the positive and negative classes. It can be seen from the sensitivity and specificity values obtained from Formulas 4 and 5 based on the information in Table 1.

Table 1. Confusion Matrix

Predict	Actual	
	True	False
True	True Positive (TP)	False Negative (FN)
False	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$Spesificity = \frac{TP}{TP+FN} \quad (5)$$

2.8 Shapley Additive Explanations (SHAP)

One way that is often used to interpret tree-based models is Shapley Additive exPlanations (SHAP). SHAP is a method that can be used to explain individual predictions that are theoretically based on Shapley game scores [21]. SHAP can explain the predictions of each x by calculating the contribution of each variable. Apart from that, SHAP can also explain global and local predictions by calculating the Shapley value in formula 6.

$$\phi_j(v) = \phi_j = \sum_{S \subseteq M\{j\}} \frac{|S|!(M-|S|-1)!}{M!} \big(v(S \cup \{j\}) - v(S)\big), \quad j = 1, \dots, M \tag{6}$$

where ϕ_j is the Shapley value for the jth variable, M is the number of predictors, $S \subseteq M\{j\}$ is the subset consisting of |S| predictor variable, $v(S \cup \{j\})$ is the predicted value for all predictor variables, and $v(S)$ is the predicted value without the jth predictor variable. The Shapley value predictor model is represented as variable attribution using the additive method, which in SHAP is explained by the linear model $g(z')$ with the formula explained in Formula 7.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \tag{7}$$

g is the model formed, $z' \in \{0,1\}$ is the coalition vector, M is the number of predictor variables, $\phi j \in \mathbb{R}$ is the Shapley value of the jth variable, and $\phi 0 \in \mathbb{R}$ is the base value of the classification model.

2.9 Food Insecurity

According to Minister of Agriculture Regulation No. 43/2010, food insecurity is a condition of the inability of an individual or group of individuals in an area to obtain sufficient and suitable food for a healthy and active life. Food insecurity can also be interpreted as the condition of an area, community, or household where the level of food availability and safety is insufficient to meet the standards of physiological needs for the growth and health of some people.

Since 2013, the Food and Agriculture Organization (FAO) has socialized a new instrument to measure food insecurity. The instrument is the Food Insecurity Experience Scale (FIES). FIES is a measure of the severity of food insecurity at the household or individual level whose value depends on the answer "Yes" or "No" to eight questions regarding household access to sufficient food [22]. The FIES consists of a series of questions that ask about a person's experience of hunger that they have experienced. The process of food insecurity is applied to eight questions. The level of food insecurity is based on the number of "Yes" answers to the FIES instrument regardless of which specific experience was experienced.

The data used comes from the National Socio-Economic Survey (SUSENAS) carried out by the Central Statistics Agency (BPS) for March 2022, with a total of 22 explanatory variables (Table 2). The explanatory variables used in this research refer to research conducted by [14]. The response variable Y (household food insecurity status) is formed from eight questions about FIES. For each question, there are four possible answers, namely "Yes", "No", "Do not Know", and "Refuse to Answer". A household is categorized as food insecure if one of the eight FIES questions scores 1 (Yes). Before being modelled, the empirical data used in this research will go through a data preparation process. Conduct data exploration to see the composition and characteristics of each variable used. Remove from the data cluster observations with the code "Refuse to Answer" or code "Do not Know" on eight FIES questions. Calculate the value of the response variable (Y) based on whether there is a "Yes" answer to the eight FIES questions. Then, aggregate individual variables into household variables.

Table 2. Research Predictors Features

	Features Name	Scale
X1	Education of Household Head	Continuous
X2	Vulnerable Household	Nominal
X3	Percentage of Family Members Having Savings Account	Continuous
X4	Number of Family Members Illiterate	Continuous
X5	Main Income From the Transferee	Nominal
X6	Ownership of Land	Nominal
X7	Internet Access	Nominal
X8	Access to Outpatient Treatment	Nominal
X9	Grantee of Non-Cash Social Assistance	Nominal
X10	Grantee of Hopeful Family Program	Nominal
X11	Grantee of Prosperous Family Program	Nominal
X12	Grantee of Social Assistance From Local Government	Nominal
X13	Grantee of Health Insurance National Program	Nominal
X14	Grantee of Scholarship Social Program	Nominal
X15	Roof Types	Nominal
X16	Wall Types	Nominal
X17	Floor Types	Nominal
X18	House Size	Continuous
X19	Electricity	Nominal
X20	Types of Cooking Fuel	Nominal
X21	Decent Sanitation	Nominal
X22	Drinking Water Source	Nominal

Generally, the data analysis procedure stages begin with data pre-processing and data exploration. Then, divide the data into 70% training data and 30% testing data. Handling imbalanced data on response variables was carried out using EasyEnsemble and SMOTE-NC. After that, build a model with training data using four machine learning algorithms: random forest, rotation forest, double random forest, and rotation double random forest. Modelling is carried out on training data with and without handling imbalanced data. Then, make predictions on the test data and calculate the level of accuracy of the model predictions that have been formed based on the specified metrics. After that, evaluate the performance of models formed from five classification algorithms. Finally, interpret the model results using SHAP.

3. Results and Discussions

3.1 Model comparison

In this study, the scope of research was limited to the South Sulawesi Province area. The number of households used in modelling was 15088. The proportion of households in the food insecure class was 17 per cent (2504 households) and the remainder (12584 households) were included in the non-food insecure household class. As many as 30 per cent (4528 households) of data were used as testing data and 70 per cent (10560 households) were used as training data. Training data was modelled on three treatment types: data with imbalance conditions, data with SMOTE-NC treatment, and EasyEnsemble treatment. Table 3 shows the data composition in various treatment conditions for handling class imbalance.

Table 3. Household proportions in modelling

Food Insecurity Status	Testing Data	Training Data		
		Imbalance Data	Easy Ensemble	SMOTE-NC
Yes	752 (17%)	1752 (17%)	1752 (50%)	8808 (50%)
No	3776 (83%)	8808 (83%)	1752 (50%)	8808 (50%)
Total	4528	10560	3504	17616

Of the predictor variables used in this study, four of them are continuous variables. So, tests are carried out to see the correlation between these continuous variables. Figure 5 shows that the four continuous variables have a very weak correlation with a value range of -0.2 to 0.17, so it can be concluded that there is no multicollinearity between continuous variables.

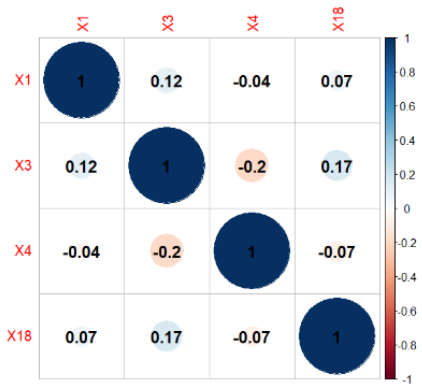

Figure 5. Correlation value between continuous variables

All models in this study were built using standard parameter values without hyperparameter tuning during modelling. Even though the model that will be produced from each algorithm is not the model with the best performance, modelling using standard parameters is expected to show the competitiveness of the model performance produced from each variety of decision tree algorithms. Each algorithm has different concepts of independence and diversity of tree ensembles, which, of course, impacts the performance of the resulting model. Table 4 shows the default parameter values in building models using the RF, RoF, DRF, and RoDRF algorithms.

Table 4. Default Parameter Value by Algorithm

Algorithm	Parameter	Values	Description
RF and DRF	Mtry	$\sqrt{p}$	the number of (randomly selected) variables to consider at each node
	Ntree	500	number of trees in the ensemble
	Nodesize	1	minimum size of terminal nodes
RoF	K	number of feature/3	number of subset features
	L	10	number of trees
RoDRF	minleaf	2	the minimum amount of samples in a leaf
	nvartosample	$\sqrt{p}$	the number of (randomly selected) variables to consider at each node
	Ntree	50	number of trees in the ensemble

Table 5 compares model performance on data with imbalance classes and on data treated using the EasyEnsemble and SMOTE-NC methods. The modelling process for each algorithm was carried out in 50 repetitions to determine the consistency of performance models produced by each algorithm. The resulting model has a fairly high accuracy value on data with an imbalance class for each algorithm. However, the resulting sensitivity value tends to be very small (0.00000 – 0.07162) compared to the specificity value. It shows that the RF, RoF, DRF, and RoDRF models tend to be biased in predicting positive classes (households with food insecurity status) in data with imbalanced classes.

Table 5. Comparison of model performance measures according to algorithms and imbalanced data handling treatment (50 replications validation)

Algorithm of Machine Learning	Average performance			
	Accuracy	Sensitivity	Specificity	AUC
RF	0.82921	0.07162	0.98008	0.68842
RoF	0.83392	0.00000	1.00000	0.50000
DRF	0.83327	0.04207	0.99011	0.69573
RoDRF	0.83310	0.03445	0.99163	0.51304
RF_SMOTE-NC	0.77799	0.25000	0.88314	0.67230
RoF_SMOTE-NC	0.72518	0.39707	0.79526	0.59644
DRF_SMOTE-NC	0.77915	0.23551	0.88742	0.66756
RoDRF_SMOTE-NC	0.77708	0.19991	0.89202	0.54597
RF_easyensemble	0.62274	0.67471	0.61239	0.68501
RoF_ easyensemble	0.61328	0.64085	0.60779	0.67051
DRF_ easyensemble	0.62688	0.67423	0.61745	0.69242
RoDRF_easyensemble	0.61528	0.64785	0.60868	0.67805

In the case classification of household food insecurity status, we hope that the model formed can predict both classes well. Households that should be food insecure but are predicted not to be food insecure or vice versa

impact government policies that are not on target. For this reason, the balance of the model in predicting positive and negative classes is an important point.

Modelling on data with class imbalance handling produces better model performance. The sensitivity value of the model increases when class imbalance is handled with SMOTE-NC. However, a fairly large gap exists between the sensitivity and specificity values. The ability of each algorithm to predict the positive class is still in the range of 0.19991 to 0.39707. In contrast to the data's EasyEnsemble treatment, the results show almost balanced sensitivity and specificity values. The model produced from the four algorithms with EasyEnsemble treatment can predict positive and negative classes more equally. For all algorithms, the model formed with EasyEnsemble treatment is better than the model formed with SMOTE-NC treatment on the data regarding prediction balance for each class. So, the next discussion will compare the model performance of the four algorithms on data with EasyEnsemble treatment.

A comparison of the performance of the four models with the EasyEnsemble treatment was carried out by looking at the AUC value of each model. To find out whether there is a significant difference in AUC values between the four algorithms, an ANOVA test was carried out. The initial hypothesis (H_0) in testing is that the four algorithms have no significant difference in AUC values. Based on the results of the ANOVA test, a p_value of 0.0000 was obtained, so the decision taken was Reject H_0. This means that with a 95 per cent confidence level, there is sufficient evidence to state that at least one algorithm has a significantly different average AUC performance compared to other algorithms. Testing continued with the Tukey HSD test to test the differences in each pair of algorithms. The test results of comparing AUC values for the four algorithms with EasyEnsemble treatment can be seen in Table 7.

Table 7. Tukey HSD test results

Algorithm	Diff	wer	upper	p adj
RF - DRF	.00741	-0.01222	-0.00259	0.00055
RoDRF - DRF	.01437	-0.01919	-0.00956	0.00000
RoF - DRF	.02191	-0.02672	-0.01709	0.00000
RoDRF - RF	.00696	-0.01178	-0.00214	0.00133
RoF - RF	.01450	-0.01932	-0.00968	0.00000
RoF - RoDRF	.00753	-0.01235	-0.00272	0.00042

Based on the table above, all algorithm pair comparisons do not contain the value 0 between the upper and lower limits. In other words, the overall average AUC values between the four algorithms are significantly different from each other. The performance of the DRF model based on AUC values is, on average, better than the RF, RoF, and RoDRF models. Furthermore, the boxplot technique can be used to see the consistency of the algorithm's performance against the resulting models. Figure 6 shows the model performance AUC values distribution for each algorithm presented in boxplot form. The four algorithms have a variety of AUC values that are not much different. Boxplots in the RoF algorithm tend to form a wider range, so the algorithm has more varied AUC values than others.

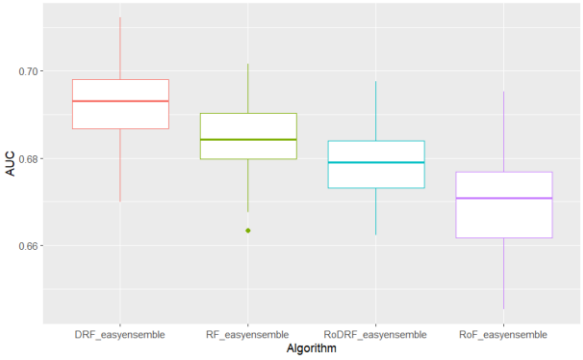

Figure 6. Boxplot: distribution model performance by algorithm

Based on all the comparisons that have been carried out, the results show that the DRF_easyensemble model is generally the best model based on the AUC value. The DRF_easyensemble model also has better accuracy values and can predict positive and negative classes more equally.

In this case, the RoDRF algorithm could not outperform the DRF algorithm. The variable rotation process in the RoF and RoDRF algorithms has yet to improve model performance significantly. This is because most of the variables used in this study are categorical, while the PCA process requires numerical variable types when rotating variables. Even though the process of converting categorical variables into dummy variables has been carried out, several studies have stated that PCA will only be effective on categorical variables if it has a variance structure (the data is binary).

3.2 Model Interpretation

Figure 7 shows ten variables that significantly contribute to determining the status of household food insecurity in South Sulawesi Province. In the SHAP plot, the red graph indicates the contribution of variables in determining the negative class (households not food insecure). Meanwhile, the green graph indicates the contribution of variables in determining the positive class (food insecure households).

Based on Figure 7, households with an average house size (X18) of 105 tend to be included in the category of households that are not food insecure. Apart from that, several other variables that tend to contribute as characteristics of households that are not food insecure include households that have land assets (X6=1), have parquet/vinyl/wood/board/ and similar types of house flooring (X17=3), the average percentage of household members who have savings above 25 per cent (X3=25), and drinking water source is refill water (X22=4).

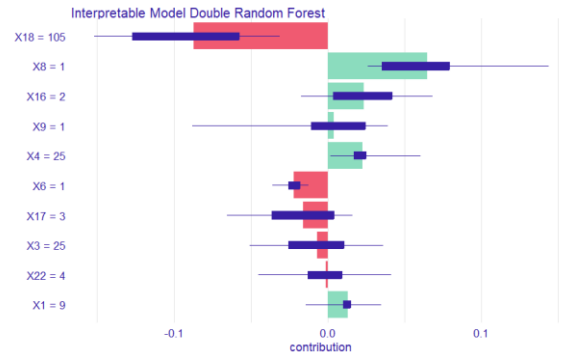

Figure 7. SHAP plot: The variable with the most significant contribution

The variables that tend to contribute as characteristics of food insecure households are households with a household member who is sick but not outpatient (X8=1). Apart from that, households with house walls made from woven bamboo/wood/board/woven bamboo plaster (X16=2), Households Receiving Non-Cash Social Assistance (X9=1), percentage of illiterate household members above 25 per cent (X4=25), and The average length of schooling of household heads is nine years/junior high school completion (X1=9) is a variable that contributes to the characteristics of food insecure households.

The results of model interpretation with SHAP can provide an overview of variables with a large contribution that characterizes each class in the classification of household food insecurity status in South Sulawesi Province.

4. Conclusions

The performance comparison results of the four algorithms show a tradeoff between the accuracy value and the model's ability to predict positive and negative classes in a balanced manner. The more balanced the sensitivity and specificity values produced by the model, the total model accuracy value tends to decrease. Handling class imbalance using the EasyEnsemble technique produces a model with better performance than the SMOTE-NC technique. The EasyEnsemble technique can maximize all the information in the data, thereby producing a model with more accurate predictions.

The accuracy value of the DRF_easyensemble model as the best model is 0.62688 on average, with an AUC value of 0.69242. This means that the variables used in this research cannot properly differentiate the food insecurity status of households in South Sulawesi Province, although the variables used in this research refer to the previous study's variables used for household food insecurity status in other regions. This shows that the characteristics of households in each region are different, so modelling needs to be done using variables that suit regional characteristics. This research also strengthens several previous studies that stated that modelling with the RoF algorithm (and RoDRF) is not the right choice when categorical variables dominate the predictor variables in the dataset.

For further research, the use of numerical variables that are related to household food insecurity status could be considered. Apart from that, exploration related to handling unbalanced data is also an important thing that can be the focus of further research, such as carrying out a combination of undersampling and oversampling to balance data.

References

[1] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *International Journal of Information and Decision Sciences*, vol. 12, no. 3, pp. 246–269, 2020, doi: 10.1504/ijids.2020.108141.

[2] J. Fitzgerald, "Bias and Variance Reduction Strategies for Improving Generalisation Performance of Genetic by A thesis for the PhD Degree," University of Limerick, 2014. [Online]. Available: http://www0.cs.ucl.ac.uk/staff/W.Langdon/ftp/papers/jmfitz-thesis.pdf

[3] H. Du, Y. Zhang, L. Zhang, and Y. C. Chen, "Selective Ensemble Learning Algorithm for Imbalanced Dataset," *Computer Science and Information Systems*, vol. 20, no. 2, pp. 831–856, 2023, doi: 10.2298/CSIS220817023D.

[4] N. Thomas Rincy and R. Gupta, "Ensemble learning techniques and its efficiency in machine learning: A survey," *2nd International Conference on Data, Engineering and Applications, IDEA 2020*, 2020, doi: 10.1109/IDEA49133.2020.9170675.

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

[6] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modelling," *Expert Syst Appl*, vol. 134, pp. 93–101, 2019, doi 10.1016/j.eswa.2019.05.028.

[7] 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.

[8] S. Han, H. Kim, and Y. S. Lee, "Double random forest," *Mach Learn*, vol. 109, no. 8, pp. 1569–1586, Aug. 2020, doi: 10.1007/s10994-020-05889-1.

[9] M. A. Ganaie, M. Tanveer, P. N. Suganthan, and V. Snasel, "Oblique and rotation double random forest," *Neural Networks*, vol. 153, pp. 496–517, Nov. 2022, doi: 10.1016/j.neunet.2022.06.012.

[10] S. Abdullah and G. Prasetyo, "Easy Ensemble With Random Forest To Handle Imbalanced Data in Classification," *Journal of Fundamental Mathematics and Applications (JFMA)*, vol. 3, no. 1, pp. 39–46, 2020, doi: 10.14710/jfma.v3i1.7415.

[11] J. Wijaya, A. M. Soleh, and A. Rizki, "Penanganan Data Tidak Seimbang pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB," *Xplore: Journal of Statistics*, vol. 2, no. 2, pp. 32–40, 2018, doi: 10.29244/xplore.v2i2.99.

[12] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information (Switzerland)*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[13] T. A. E Ramadhani, B Sartono, A F Hadi, S 'Ufa, "Comparison of Main Characteristics of Food Insecurity Using Classification Tree and Random," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 7, no. 4, pp. 2486–2497, 2022, doi: https://doi.org/10.33395/sinkron.v7i4.11852.

[14] H. Dharmawan, B. Sartono, A. Kurnia, A. F. Hadi, and E. Ramadhani, "A Study of Machine Learning Algorithms To

Measure the Feature Importance in Class-Imbalance Data of Food Insecurity Cases in Indonesia," *Communications in Mathematical Biology and Neuroscience*, vol. 2022, pp. 1–25, 2022, doi: 10.28919/cmbn/7636.

[15] A. Bagnall, M. Flynn, J. Large, J. Line, A. Bostrom, and G. Cawley, "Is rotation forest the best classifier for problems with continuous features?," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.06705

[16] B. Sartono, M. Raharjo, and C. Suhaeni, "Empirical Study on the Predictive Power of Rotation Forest," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Nov. 2018. doi: 10.1088/1755-1315/187/1/012053.

[17] E. C. Gök and M. O. Olgun, "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," *Neural Comput Appl*, vol. 33, no. 22, pp. 15693–15707, 2021, doi: 10.1007/s00521-021-06189-y.

[18] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "SMOTE: Potensi dan Kekurangannya Pada Survei," *E-Jurnal Matematika*, vol. 10, no. 4, p. 235, Nov. 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[19] A. J. Bowers and X. Zhou, "Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes," *J Educ Stud Placed Risk*, vol. 24, no. 1, pp. 20–46, 2019, doi: 10.1080/10824669.2018.1523734.

[20] C. Halimu, A. Kasem, and S. H. S. Newaz, "Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification," *ACM International Conference Proceeding Series*, no. Mcc, pp. 1–6, 2019, doi: 10.1145/3310986.3311023.

[21] H. Chen, I. C. Covert, S. M. Lundberg, and S. Lee, "Algorithms to estimate Shapley value feature attributions," *arXiv:2207.07605v1 [cs.LG]*, no. Section 3, 2022, doi: https://doi.org/10.48550/arXiv.2207.07605.

[22] A. Saint Ville, J. Y. T. Po, A. Sen, A. Bui, and H. Melgar-Quiñonez, "Food security and the Food Insecurity Experience Scale (FIES): ensuring progress by 2030," *Food Secur*ity, vol. 11, no. 3, pp. 483–491, 2019, doi: 10.1007/s12571-019-00936-9.

