Scientific Journal of Informatics
Vol. 11, No. 1, Feb 2024
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

A Comparative Study of Random Forest and Double Random Forest Models from View Points of Their Interpretability

Adlina Khairunnisa[1], Khairil Anwar Notodiputro[2*], Bagus Sartono[3]
[1, 2, 3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Pertanian Bogor, Indonesia

Abstract.
Purpose: This study aims to compare the performance of ensemble trees such as Random Forest (RF) and Double Random Forest (DRF) from view points of interpretability of the models. Both models have strong predictive performance but the inner working of the models is not human understandable. Model interpretability is required to explain the relationship between the predictors and the response. We apply association rules to simplify the essence of the models.
Methods: This study compares interpretability of RF and DRF using association rules. Each decision tree formed from each model is converted into if-then rules by following the path from root node to leaf nodes. The data was selected in such a way that they were underfit data. This is due to the fact that DRF has been shown by other researchers to overcome the underfitting problem faced by RF. A Simulation study has been conducted to evaluate the extracted rules from RF and DRF. The rules extracted from both models are compared in terms of model interpretability based on support and confidence values. Association rules may also be applied to identify the characteristics of poor people who are working in Yogyakarta.
Result: The simulation results revealed that the interpretability of DRF outperformed RF especially in the case of modelling underfit data. On the other hand, using empirical data we have been able to characterize the profile of poor people who are working in Yogyakarta based on the most frequent rules.
Novelty: Research on interpretable DRF is still rare, especially the interpretation model using association rules. Previous studies focused only on interpreting the random forest model using association rules. In this study, the rules extracted from the random forest and double random forest models are compared based on the quality of the rules extracted.

Keywords: Interpretability, Association rules, Rule extraction, Double random forest
Received December 2023 / **Revised** February 2024 / **Accepted** February 2024

INTRODUCTION

Statistical learning is a technique for understanding data [1]. It can be used to classify objects and predict outcomes. A commonly used method in the classification problem is Tree-based methods introduced by Breiman et al. [2]. Single tree models such as CART can be interpreted easily because the prediction logic is transparent. It can be followed by observing the rules formed at each split of the decision tree. CART is known as an unstable method because slightly changes in the training data can lead to biased prediction results [3]. One of the ensemble tree methods used to deal with such unsteadiness is Random Forest (RF).

RF is a popular ensemble tree method that is highly accurate [4]. There are two randomization processes in RF such as constructing multiple decision trees using different training data by sampling with replacement (bootstrapping) and randomly selecting a subset of predictors for each split. Randomly selecting some predictors at each split reduces the correlation between the trees, which improves predictive performance. Accordingly, RF has the advantages of overcoming overfitting and being insensitive to outliers [5]. RF can also handle unbalanced data [6].

DRF is a new ensemble tree method similar to RF. DRF may outperform RF when the RF model is underfitting [7]. It is characterized by a relative test accuracy of less than 1 which means that the RF tree size is not large enough to perform well. The relative test accuracy is the accuracy of RF when a given *nodesize* is divided by the accuracy when the *nodesize* is set to the default value of 1. DRF differs from RF in that it uses all training data instead of bootstrapped training data. The use of all training data results in

*Corresponding author.
Email addresses: khairil@apps.ipb.ac.id (Notodiputro)
DOI: 10.15294/sji.v11i1.48721

more unique observations in the nodes, which makes the trees larger than the trees constructed by RF. Bootstrap sampling and random variable selection at each node are used to find the best splitting rule. These steps can add randomness to the splitting rule. This creates a more diverse tree in DRF than in RF. Therefore, DRF may have better predictive performance than RF for datasets where RF is underfitting.

Previous study evaluated DRF with single classification trees, bagging, Samme and RF on 34 datasets [7]. The findings showed that DRF significantly outperformed other methods when RF underfitted the data. A study using DRF method was also performed by Aldania et al. [8] to classify the 2015 Indonesian industrial classification code (KBLI) from the "I" category (accommodation and food services activities) sourced from the 2016 economic census (SE2016) listing results. The results showed that DRF performed well.

Although the ensemble tree method provides strong predictive performance, it has limitation in terms of model interpretability. Due to its lack of interpretability, the ensemble tree method is often called a black box model because it is difficult to interpret [9]. Interpretability in machine learning is the ability of a model to be understood by humans [10]. Interpretability is essential to understand the data based on the model's predictions. Additionally, model interpretability is required to explain the factors that influence the model's predictions.

Each interpretation method attempts to capture the pattern of the model to reduce the risk of misinterpretation of the model predictions. Breiman et al. [4] used the variable importance approach to identify the most important variables in the RF model. However, it does not explain the interaction between the predictors and the response. Another approach to explain the black box model is the decision tree approach [11]. Decision tree can be easily interpreted. The decision tree is converted into an association rule in the form of if-then by following the path from the root node to the leaf nodes. This rule consists of two parts, condition (the "if" part) and prediction (the "then" part). An example of if-then rule is if today is weekend and today is sunny then go on holiday.

Association rules can be applied in ensemble trees such as RF in interpretable trees (inTrees) framework [12]. Association rules is well-suited for interpreting the structure of models that implement a forest of trees. This method can extract simple rules from a complex set of trees. The inTrees framework works by extracting and classifying a set of rules from each decision tree. The rules extracted from the ensemble tree are combination of the rules extracted from all decision trees. The set of rules represents the relationship between the predictors and response in the form of if-then rules. Consequently, the rules can further be used for prediction.

Association rules were used to identify the necessity of biopsy and surgery for thyroid patients [13]. In the economic field, rules extracted from the model were used to identify the factors that lead to economic recessions in the United States [14]. In the social field, rules extracted from RF model showed that the rules that most characterize the poverty status of households in Tasikmalaya City are house wall materials and main source of drinking water, house wall materials and cooking fuel, and house wall materials and motorcycle ownership [15]. Interpretation using association rules may also be applied to the issue of the working poor in Yogyakarta. According to the National Socio-Economic Survey (Susenas) in March 2022, the poverty rate in Yogyakarta is 11.34%, the highest in Java. In contrast, the unemployment rate in Yogyakarta is 4.06%, which is relatively low. This indicates that there are working poor in Yogyakarta. Therefore, it is important to identify the factors that contribute to working poverty in Yogyakarta. Extracting rules from the ensemble tree can be used to develop policies to reduce the risk of being poor.

According to the previous background, RF and DRF are a number of decision trees that use aggregation to achieve more accurate predictions than single decision tree. However, having a powerful prediction is insufficient for understanding the model's predictions. Interpretation of the model is needed to understand how the predictors affect the response. For this reason, association rules are applied to RF and DRF to interpret the models. This study compares the extracted rules based on the rule quality measures. Moreover, this study also applies the association rules method to the empirical data.

METHODS

Data

The data used consists of simulation data and empirical data. Simulation data is used to evaluate the performance of RF and DRF from the view points of model interpretability. Simulated data is generated

with a number of classes for the response variable (Y) is 2 and the number of predictors is 4 (X_1, X_2, X_3, X_4). X_1 is continuous data generated from uniform distribution in the interval [0,1]. X_2, X_3, and X_4 are nominal data, each consisting of 5 categories, 4 categories, and 2 categories, respectively. X_2, X_3, and X_4 are generated from discrete uniform distribution. Furthermore, X_2, X_3, and X_4 are converted into dummy variables for identifying the contribution of each category to the response. For illustration, X_2 is divided into D_{22}, D_{23}, D_{24}, D_{25} with reference category is $X_2 = 1$. D_{24} is assigned a coefficient equal to 2 and D_{25} is assigned a coefficient equal to 3, while D_{22} and D_{23} is assigned a coefficient equal to 0. The coefficient indicates that $X_2 = (4,5)$ has a higher contribution to being classified as $Y = 1$ than $X_2 = (1,2,3)$. This is applicable for other predictors.

Table 1. Description of predictors

Predictors	Dummy Variable	Category	Coefficient	Effect on Y = 1	Description
X_1	-	-	3	Has effect	Continuous data
X_2	D_{22}	$X_2 = 2$	0	No effect	Nominal data with the reference category $X_2 = 1$
	D_{23}	$X_2 = 3$	0	No effect	
	D_{24}	$X_2 = 4$	2	Has effect	
	D_{25}	$X_2 = 5$	3	Has effect	
X_3	D_{32}	$X_3 = 2$	0	No effect	Nominal data with the reference category $X_3 = 1$
	D_{33}	$X_3 = 3$	2	Has effect	
	D_{34}	$X_3 = 4$	3	Has effect	
X_4	D_{42}	$X_4 = 2$	0	No effect	Nominal data with the reference category $X_4 = 1$

The simulation data used in this study is 5,000 observations with the equation of X_1, X_2, X_3, X_4 follows:

$$z = -3.5 + 3X_1 + 0D_{22} + 0D_{23} + 2D_{24} + 3D_{25} + 0D_{32} + 2D_{33} + 3D_{34} + 0D_{42} \quad (1)$$

Then, the response (Y) is generated from the Bernoulli distribution based on the probability values calculated from the following equation:

$$P(Y = 1) = \frac{\exp(z)}{1 + \exp(z)} \quad (2)$$

Data generated when RF may underfit is required to DRF modelling. The dataset is obtained from RF that RF underfits the data. RF is underfit if the relative test accuracy of RF is less than 1 [7]. The relative test accuracy is the accuracy of RF when the given *nodesize* is divided by the accuracy of RF when the *nodesize* is set to default value 1. The *nodesize* used are $0{,}01n;\ 0{,}02n;\ 0{,}03\ n; \ldots;\ 0{,}09n; 0{,}1n$ and 1 with n is the number of the training data. Training data is used for modelling the RF while test data is used to calculate the relative test accuracy at each *nodesize* setting. If all relative test accuracy is less than 1, RF is underfit on simulated data. The data generation is based on trial and error until 100 underfit datasets are obtained.

Empirical data is used as an implementation of the association rules method to identify the characteristics of the working poor in Yogyakarta. The empirical data used in this study was sourced from the National Socio-Economic Survey (Susenas) in 2022, collected by the Statistics Indonesia (BPS). The response used in this study is the poverty status of workers in Yogyakarta. Poverty status refers to the concept of both poverty and employment. The measurement of poverty is based on the individual's capacity to fulfill both food and non-food needs [16]. Meanwhile, individuals classified in the worker category are those aged 15 and above, engaged in activities to earn or assist in earning income, for a minimum of one hour (consecutively) per week, or have a job but did not work due to holidays, leave, illness, etc [17]. Therefore, a poor worker is someone who is employed but resides in a household below the poverty line. The variables used are based on previous studies [18], [19].

Table 2. Variables used in empirical data

Variable	Description	Scale
Poverty status of worker	1: Poor worker; 0: Non-poor worker	Nominal
Age	Age of worker	Ratio
Gender	1: Male; 2: Female	Nominal
Marital status	1: Never married; 2: Married; 3: Divorce; 4: Widowed	Nominal
Educational level	1: No education; 2: Primary school; 3: Secondary school; 4: High school; 5: University	Ordinal
Place of birth	1: Yogyakarta; 2: others	Nominal
Residence of 5 years ago	1: Yogyakarta; 2: others	Nominal
Literacy ability	1: Able; 2: Unable	Nominal
Functional disability	1: Exists; 2: Not exist	Nominal
Internet use	1: Use internet; 2: Not use internet	Nominal
Job sector	1: Agriculture; 2: Mining and quarrying; 3: Construction; 4: Industry; 5: Electricity, gas, and water; 6: Trade, accomodation, and restaurants; 7: Transport and communication; 8: Other services	Nominal
Working hours	Weekly working hours	Ratio
Employment status of worker	1: Self-employed; 2: Employer with unpaid worker; 3: Employer with paid worker; 4: Employee; 5: Freelancer; 6: Family worker/unpaid worker	Nominal
Proportion of working household member	The proportion of working to the total household member	Ratio
Home ownership	1: Own a home; 2: others	Nominal
Access to credit	1: Has access; 2: No access	Nominal

Model

RF and DRF are used as classification models. RF constructs decision trees using bootstrap sampling from the training data and randomly selects variables subset ($m \approx \sqrt{p}$) at each node. This step is performed k times to construct k decision trees. The final prediction is obtained by aggregating the predictions from k decision trees through majority voting. In contrast, the DRF model uses the entire training dataset rather than bootstrap results. This results in more unique observations at each node, leading to larger trees compared to RF. DRF introduces additional randomness into the tree-building process through bootstrap sampling and random variable selection at each node. The process is carried out until k decision trees are obtained. Aggregating the predictions of k decision trees uses majority voting to obtain the final prediction for the response variable class.

As an illustration, consider simulated data where the response variable (Y) takes on values of either class 1 or 0, and the explanatory variables (X) include X_1, X_2, X_3, and X_4. The illustration of the classification models RF and DRF for predicting the response variable classes $Y = 1$ or $Y = 0$ is presented in Figures 1 and 2.

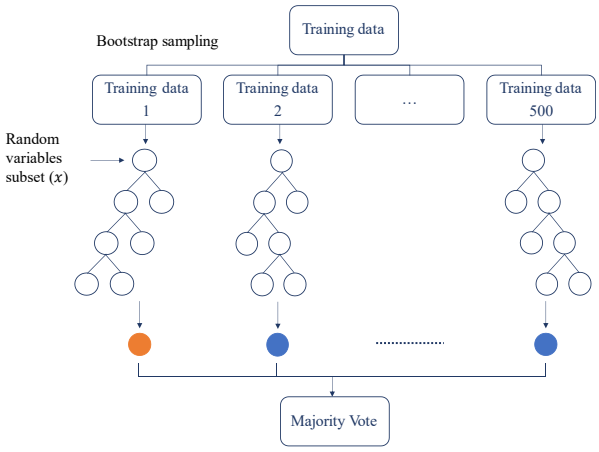

Figure 1. Illustration of RF model

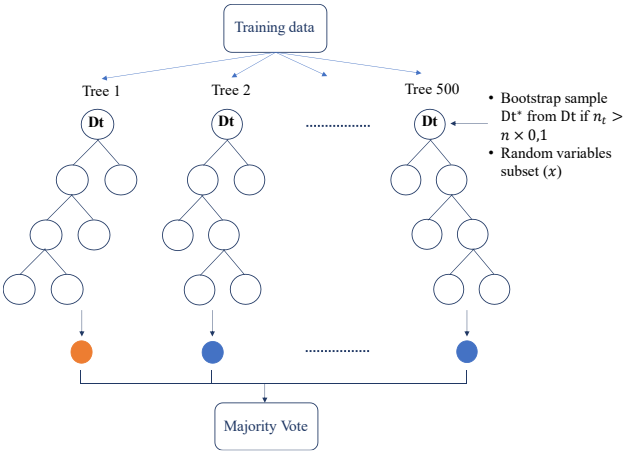

Figure 2. Illustration of DRF model

Classification analysis using RF and DRF is carried out by constructing 500 trees. Each tree will predict the probability of a data being classified into the j class. The final class prediction is aggregating of 500 trees by taking majority vote. $C_B(x)$ is the final prediction of class response based on predictors (x). b is the index of the tree $b = 1, \dots, 500$. j is the response class, $j = 0{,}1$. $I(C_b(x) = j)$ is an indicator function, which is 1 if the tree predicts class $Y = 1$ or $Y = 0$, 0 otherwise.

$$C_B(x) = argmax_j \sum_{b=1}^{500} I(C_b(x) = j), \qquad j = 0{,}1 \tag{3}$$

Next, rules are extracted from each tree in the RF and DRF models. The combination of rules from multiple trees is converted into if-then form.

Rule extraction

The inTrees approach [12] utilizes association rule techniques to explain the relationships between predictors and response variable in ensemble tree models. This approach involves extracting rules from each decision tree, identifying combinations of variables with their most frequent values, and assessing the quality of the extracted rules. The objective of this approach is to provide information that is more easily understood and interpreted.

Ensemble tree is a set of k decision tree. Figure 3 illustrates a decision tree of an ensemble tree. Each node represents splitting and paths from the root node to the leaf node [20]. The tree in Figure 3 has five leaf nodes, so there are five paths from the root node to leaf nodes 1 to 5. The following paths from the root node to the leaf node yield rules that identify the relationship between predictors and response. Accordingly, each decision tree in ensemble tree model will have different number of rules due to the number of rules extracted depending on the number of leaf nodes generated in each decision tree.

Rules are generally expressed as X ⇒ Y. X is the condition and Y is the prediction. Based on Figure 3, the rules extracted from the decision tree are as follows:

1. The first leaf node has a rule $\{X_1 = 5, X_2 = 4, \text{dan } X_3 = 1 \Rightarrow Y = 1\}$
2. The second leaf node has a rule $\{X_1 = 5, X_2 = 4, \text{dan } X_3 = 2 \Rightarrow Y = 0\}$
3. The third leaf node has a rule $\{X_1 = 5 \text{ dan } X_2 = 3 \Rightarrow Y = 1\}$
4. The fourth leaf node has a rule $\{X_1 = 1 \text{ dan } X_4 > 0{,}5 \Rightarrow Y = 1\}$
5. The fifth leaf node has a rule $\{X_1 = 1 \text{ dan } X_4 \leq 0{,}5 \Rightarrow Y = 0\}$

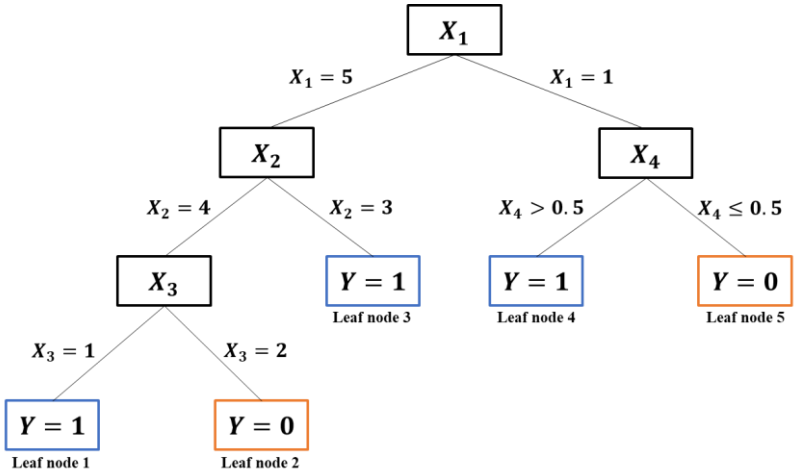

Figure 3. Structure of decision tree

The metrics used to measure the quality of rules include support and confidence values. Support indicates the proportion of occurrences of the combination of X and Y in the dataset, while confidence indicates how often Y appears in rules containing X (Han et al., 2012). In terms of model interpretation, support is defined as the percentage of rules appearing in the trees generated by the RF and DRF models. Confidence represents the accuracy level of a rule in correctly predicting the class of the response variable. High support implies that a rule frequently appears in the trees formed by the model and high confidence indicates that the rule is able to correctly predict the class of the response variable.

Data analysis procedure

The following are the stages of analysis in this study:

1. For simulated data, simulated data is generated according to the simulation data procedure when RF is underfit. For empirical data, the data preprocessing is conducted by classifying the status of the working poor according to the BPS definition and regrouping the categorical variables. Identification of the possibility that RF is underfit is carried out on empirical data.
2. Divide the data into two parts, 80% training data and 20% test data. The simulation data consists of 5,000 observations, divided into training data with 4,000 observations and test data with 1,000 observations. Meanwhile, the empirical data comprises 6,520 observations, with training data consisting of 5,216 observations and test data consisting of 1,304 observations. Specifically for empirical data, balancing of the training data is performed using SMOTE-NC and the identification of the RF model resulting in an underfit model.
3. RF and DRF modelling utilizes the training data. For simulation data, the parameters used are default. For empirical data, the parameters used are the results of hyperparameter tuning using 10-fold cross validation.
4. Conducting the performance evaluation of predictions by calculating AUC values on the test data.
5. Extracting rules from each tree formed in RF and DRF models.
6. In the simulation data, the process is carried out on 100 underfit datasets. The purpose is to evaluate the interpretability performance of RF and DRF models based on the support and confidence values. For empirical data, model interpretation uses the best model. The rules extraction involves identifying frequent variable interactions and analyzing the rules based on support and confidence values.

RESULTS AND DISCUSSIONS

Simulation study

Modelling using simulated data selected in such a way as to reflect RF conditions results in underfit models. The underfit data was used to build the model because the use of DRF provides good accuracy when the RF is underfitting [7]. In modeling simulated data, the parameters set are default values. In general, default values result in good model performance [23]. RF and DRF modeling and rule extraction were performed on 100 underfit datasets. The modeling results show that the RF model produces an AUC of 96.64% and DRF produces an AUC of 96.69%. Then, the results of the paired t-test on the AUC value show a p-value of less than 0.05 in the underfit simulation data (Table 3). This means that there is a significant difference

in the average AUC value between RF and DRF. Although the difference in AUC value between RF and DRF is very small, DRF shows better prediction performance than RF on the underfit datasets.

Table 3. The model performance of simulated underfit datasets

Model	AUC	$\bar{x}_d$	t-statistic	p-value
DRF	96.69%	0,052	2,26	0,026*
RF	96.64%			

*Significant at 0.05

After RF and DRF modelling, rules are extracted from 500 trees formed in RF and DRF. From the extracted rules, there are 67 identical rules from all 100 underfit datasets. Then, the most frequent variable interaction is measured based on support and confidence values. The support value is the percentage of occurrence of a rule in the tree formed in the model, while the confidence value shows the accuracy of the prediction results of a rule.

The support and confidence values of the 100 datasets when RF is underfit were analysed using paired t-test. This test is to determine the significant difference of RF and DRF performance in terms of model interpretability using association rules. The null hypothesis is that the RF and DRF do not have significantly different performance based on rule quality metrics. The alternative hypothesis tested is that both models have significantly different performance in terms of model interpretability.

Table 4. Comparison of RF and DRF based on support and confidence values

Rule	Condition	Pred	Support			Confidence		
			$\bar{x}_d$	p-value	DRF vs RF	$\bar{x}_d$	p-value	DRF vs RF
1	$0{,}15 < X_1 \leq 0{,}85$ & $X_3 = 4$	1	-0,002	0,000*	RF	0,000	-	-
2	$X_1 \leq 0{,}15$ & $X_2 = (1,2,3)$	0	0,002	0,000*	DRF	0,000	-	-
3	$X_1 \leq 0{,}15$ & $X_3 = (1,2)$	0	0,001	0,000*	DRF	0,000	-	-
4	$X_2 = 5$ & $X_3 = (3,4)$	1	0,001	0,002*	DRF	-0,000	0,320	-
5	$X_2 = (4,5)$ & $X_3 = 4$	1	0,002	0,000*	DRF	0,000	0,158	-
6	$X_2 = 5$ & $X_3 = 3$	1	0,002	0,000*	DRF	0,000	0,940	-
7	$X_2 = 4$ & $X_3 = 4$	1	0,003	0,000*	DRF	-0,000	0,606	-
8	$X_2 = (4,5)$ & $X_3 = 3$	1	0,002	0,000*	DRF	-0,002	0,209	-
9	$X_2 = 4$ & $X_3 = (3,4)$	1	0,001	0,000*	DRF	0,000	0,986	-
10	$X_2 = 4$ & $X_3 = 3$	1	0,001	0,000*	DRF	-0,002	0,243	-
11	$X_1 > 0{,}15$ & $X_2 = 5$ & $X_4 = 1$	1	-0,002	0,000*	RF	0,012	0,000*	DRF
12	$X_1 > 0{,}15$ & $X_2 = 5$	1	-0,003	0,000*	RF	0,009	0,000*	DRF
13	$X_1 > 0{,}15$ & $X_3 = 4$ & $X_4 = 2$	1	-0,002	0,000*	RF	0,010	0,001*	DRF
14	$X_1 > 0{,}15$ & $X_3 = 4$	1	-0,004	0,000*	RF	0,013	0,000*	DRF
15	$X_1 > 0{,}15$ & $X_3 = 4$ & $X_4 = 1$	1	-0,003	0,000*	RF	0,011	0,000*	DRF
16	$X_2 = (1,2,3)$ & $X_3 = 3$	0	0,002	0,000*	DRF	0,002	0,024*	DRF
17	$X_2 = 4$ & $X_3 = (1,2)$	0	0,001	0,035*	DRF	0,003	0,007*	DRF
18	$X_1 \leq 0{,}85$ & $X_2 = 5$	1	-0,003	0,000*	RF	0,016	0,000*	DRF
19	$X_1 \leq 0{,}75$ & $X_2 = 5$	1	0,002	0,000*	DRF	0,006	0,056	-
20	$X_1 \leq 0{,}85$ & $X_3 = 4$	1	-0,002	0,000*	RF	0,011	0,000*	DRF
21	$X_2 = 5$ & $X_4 = 2$	1	-0,002	0,000*	RF	0,011	0,000*	DRF
22	$X_2 = 5$	1	0,000	0,977	-	0,012	0,000*	DRF
23	$X_2 = 5$ & $X_4 = 1$	1	-0,001	0,000*	RF	0,015	0,000*	DRF
24	$X_2 = 4$ & $X_4 = 2$	1	-0,001	0,000*	RF	0,012	0,000*	DRF
25	$X_3 = 4$	1	0,000	0,775	-	0,013	0,000*	DRF
26	$X_3 = 4$ & $X_4 = 1$	1	-0,002	0,000*	RF	0,013	0,000*	DRF
27	$X_1 \leq 0{,}75$ & $X_3 = (1,2)$	0	0,003	0,000*	DRF	-0,006	0,090	-
28	$X_1 \leq 0{,}15$ & $X_4 = 1$	0	0,000	0,272	-	0,004	0,203	-
29	$X_1 \leq 0{,}15$	0	0,002	0,000*	DRF	0,000	0,783	-

*Significant at 0.05

Table 4. Comparison of RF and DRF based on support and confidence values (cont.)

Rule	Condition	Pred.	Support			Confidence		
			$\bar{x}_d$	p-value	DRF vs RF	$\bar{x}_d$	p-value	DRF vs RF
30	$0{,}15 \leq X_1 \leq 0.75$	0	0,001	0,001*	DRF	-0,002	0,525	-
31	$X_1 \leq 0{,}15$ & $X_4 = 2$	0	0,000	0,188	-	0,003	0,139	-
32	$X_1 \leq 0{,}75$ & $X_2 = (1,2,3)$	0	0,002	0,000*	DRF	-0,002	0,415	-
33	$0{,}75 < X_1 \leq 0{,}85$	1	0,000	0,620	-	0,004	0,207	-
34	$X_1 \leq 0{,}75$ & $X_3 = (1,2,3)$	0	0,000	0,308	-	-0,005	0,098	-
35	$X_1 > 0{,}75$ & $X_3 = (1,2)$	1	0,001	0,015*	DRF	-0,005	0,119	-
36	$X_2 = (1,2,3,4)$ & $X_3 = 3$	0	-0,000	0,628	-	-0,002	0,276	-
37	$X_2 = (4,5)$	1	0,005	0,000*	DRF	0,010	0,000*	DRF
38	$X_1 \leq 0{,}75$ & $X_2 = (1,2,3,4)$	0	0,001	0,154	-	-0,006	0,062	-
39	$X_3 = (1,2)$ & $X_4 = 1$	0	0,001	0,016*	DRF	-0,001	0,852	-
40	$X_3 = (1,2)$	0	0,005	0,000*	DRF	0,003	0,191	-
41	$X_3 = (1,2)$ & $X_4 = 2$	0	0,001	0,000*	DRF	0,006	0,116	-
42	$X_2 = (1,2,3)$ & $X_4 = 1$	0	0,001	0,000*	DRF	0,002	0,563	-
43	$X_2 = (1,2,3)$	0	0,006	0,000*	DRF	0,002	0,379	-
44	$X_2 = (1,2,3)$ & $X_4 = 1$	0	0,001	0,000*	DRF	0,002	0,609	-
45	$X_1 > 0{,}75$ & $X_2 = (1,2,3)$	1	0,002	0,000*	DRF	-0,003	0,310	-
46	$X_3 = (1,2,3)$ & $X_4 = 2$	0	-0,001	0,001*	RF	0,009	0,074	-
47	$X_3 = (1,2,3)$	0	-0,001	0,067	-	0,007	0,021*	DRF
48	$X_3 = (1,2,3)$ & $X_4 = 1$	0	-0,001	0,000*	RF	0,009	0,066	-
49	$X_1 > 0{,}75$	1	0,001	0,226	-	0,002	0,190	-
50	$X_3 = (3,4)$	1	0,003	0,000*	DRF	0,008	0,026*	DRF
51	$X_1 \leq 0{,}45$	0	0,002	0,006*	DRF	-0,013	0,000*	RF
52	$0{,}45 \leq X_1 \leq 0{,}75$	0	0,001	0,041*	DRF	-0,021	0,000*	RF
53	$X_1 > 0{,}75$ & $X_3 = 3$	1	0,001	0,001*	DRF	-0,004	0,056	-
54	$X_2 = 4$ & $X_3 = (1,2,3)$	0	-0,004	0,000*	RF	-0,001	0,633	-
55	$0{,}85 < X_1 \leq 0{,}95$	1	-0,001	0,001*	RF	0,002	0,558	-
56	$X_2 = (1,2,3,4)$ & $X_4 = 1$	0	0,001	0,036*	DRF	-0,003	0,616	-
57	$X_2 = (1,2,3,4)$	0	0,002	0,001*	DRF	-0,001	0,697	-
58	$X_2 = (1,2,3,4)$ & $X_4 = 2$	0	0,001	0,000*	DRF	0,005	0,350	-
59	$X_1 > 0{,}15$ & $X_2 = (1,2,3)$	0	0,003	0,000*	DRF	0,009	0,035*	DRF
60	$X_1 > 0{,}15$ & $X_3 = (1,2,3)$	0	0,000	0,511	-	0,001	0,910	-
61	$X_1 > 0{,}75$ & $X_3 = (1,2,3)$	1	-0,002	0,000*	RF	-0,004	0,317	-
62	$X_1 \leq 0{,}45$ & $X_3 = 3$	0	0,000	0,237	-	-0,007	0,050*	RF
63	$X_1 \leq 0{,}15$ & $X_3 = 4$	1	0,000	0,314	-	-0,001	0,514	-
64	$X_1 \leq 0{,}65$	0	-0,001	0,033*	RF	-0,018	0,001*	RF
65	$X_1 \leq 0{,}75$ & $X_4 = 1$	0	0,001	0,080	-	-0,005	0,137	-
66	$X_1 \leq 0{,}75$	0	0,005	0,000*	DRF	-0,007	0,000*	RF
67	$X_1 \leq 0{,}75$ & $X_4 = 2$	0	0,001	0,006*	DRF	-0,001	0,796	-

*Significant at 0.05

Table 4 shows the statistical testing results based on the support values from the RF and DRF models. Of the 67 corresponding rules, the difference in support values were analysed to evaluate the performance of RF and DRF. A positive difference in support value ($\bar{x}_d$) between DRF and RF indicates that the interpretability of the DRF model is better than RF and vice versa. DRF outperforms RF based on the support value based on 35 rules formed which implies that 52.24% of the rules are positive and significant. On the contrary, RF outperforms DRF only for 18 corresponding rules. This shows that the rules that extracted from DRF model appear more frequently compared to RF.

Furthermore, the rules formed in the model on underfit datasets are also evaluated based on the confidence values. Most of the confidence value differences of the rules do not show significant differences between RF and DRF (Table 4). Within the 67 corresponding rules, there are 19 rules with significantly positive confidence, while only 5 RF extracted rules are negative and significant. This indicates that the rules extracted from RF and DRF tend to have similar predictive accuracy in correctly predicting the response variable class on underfit datasets.

Nevertheless, the rules extracted from RF and DRF should be supported by substantial evidence indicating the accuracy of these rules. The support value becomes a crucial metric, reflecting how frequently a rule appears in the trees formed within the model. From the earlier analysis regarding support values, the majority of rules extracted from DRF exhibit higher support values compared to RF on underfit datasets. This suggests that rules extracted from DRF are more reliable in explaining the relationship between the predictors and the response variable than rules extracted from RF. Overall, DRF demonstrates superior performance over RF in interpreting the model based on support and confidence values on underfit data.

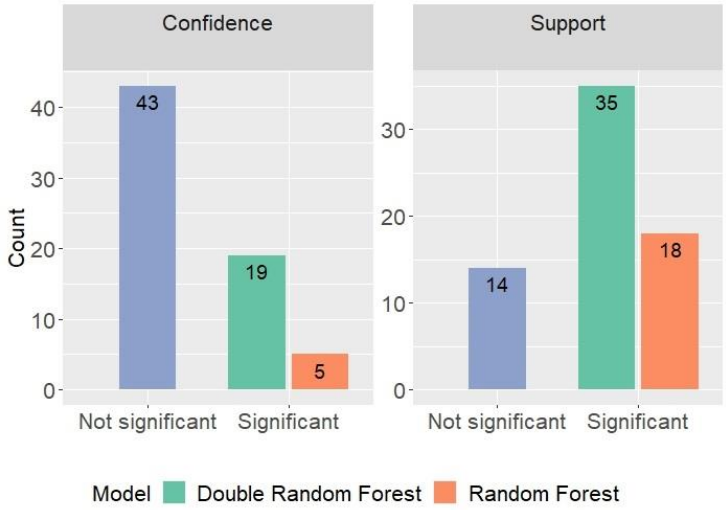

Figure 4. Summary of rule quality measures

Empirical study

Empirical data regarding the working poor in Yogyakarta consists of 6,520 observations, with the response variable classifying individuals into poor workers and non-poor workers. In Table 5, the proportion of worker poverty status indicates that the percentage of poor workers in Yogyakarta is 10.25% of the total observations. This condition indicates data imbalance, thus requiring a balancing process using the Synthetic Minority Oversampling Technique Nominal-Continuous (SMOTE-NC) on the training data. SMOTE-NC is applied to dataset that consist of both categorical and continuous variables [24]. Synthetic data for continuous variables is generated by randomly selecting k nearest neighbors and calculating the Euclidean distance. Synthetic observations are then created along the straight line connecting minority observations and the selected nearest neighbors. Synthetic data for categorical variables is generated based on the most frequent category among the k-nearest neighbors.

Table 5. Proportion of worker poverty status data in Yogyakarta

Poverty status of worker	Proportion
Poor workers	0.1025
Non-poor workers	0.8975

Identification of RF producing an underfit model is conducted prior to constructing the classification model. The procedural steps involve calculating the accuracy value on the test set, similar to the approach used in the simulation data. The relative test accuracy of the RF for all given *nodesize* is less than 1. This implies that RF underfits the empirical data of Yogyakarta. Therefore, DRF is applicable to the empirical data to improve the performance of RF. Afterwards, hyperparameter tuning is carried out before modelling DRF. The process of selecting optimal parameter values is to achieve the most accurate performance [25]. The hyperparameters used in DRF are *nodesize* (minimum number of observations in a leaf node) and number of trees. The combinations of nodesize used in the modeling are 5, 6, 7, 8, 9, and 10, while the combinations of the number of trees used in the modeling are 500 and 1000. The optimal parameter values obtained from 10-fold cross validation are 1,000 trees and *nodesize* equal to 1. These parameters are then used for modelling the DRF.

Furthermore, model evaluation is conducted using test data to obtain AUC, sensitivity, and specificity values (Table 6). The AUC value is 73.85%. This indicates that the DRF model performs quite well in classifying poor workers in Yogyakarta [26]. In addition to the AUC value, sensitivity, and specificity values are also considered in evaluating the predictive performance of a model. The sensitivity value shows that 75.48% of poor workers in DI Yogyakarta are predicted as poor workers. The specificity value indicates that 68.67% of non-poor workers in DI Yogyakarta are predicted as non-poor workers.

Table 6. The model performance of DRF model on working poor dataset

Data	AUC	Sensitivity	Specificity
Training data	79.67%	80.36%	79.54%
Test data	73.85%	75.48%	68.67%

After modelling the DRF, rules are extracted from 1,000 trees The rule extraction from the model resulted in 45,859 rules. Out of these, 206 frequently occurring rules explain interactions between variables. Among these rules, 151 predict poor workers, while 55 predict non-poor workers.

Table 7. Characteristics of the working poor in Yogyakarta based on the top 5 highest support

Rule	Working hours	Place of birth	Proportion of working household member	Marital	Residence of 5 years ago	Job sector	Access to credit	*Sup*	*Conf*
1	≤ 34,5	Yogyakarta	-	-	-	-	-	0.03	0.97
2	≤ 34,5	Yogyakarta	≤ 0,65	-	-	-	-	0.02	0.97
3	-	Yogyakarta	≤ 0,65	Married	-	-	-	0.02	0.98
4	-	-	≤ 0,65	-	Others	-	-	0.02	1.00
5	-	-	-	-	-	Industry	No access	0.02	1.00

Table 7 describes interactions between variables that frequently occur based on the top 5 highest support values with confidence above 95%. The confidence value represents the accuracy of the rules in predicting poor workers. The support value is the percentage of rule occurrences in the trees formed within the model. Here are the characteristics of workers classified as poor in Yogyakarta:

1. If someone works less than or equal to 34.5 hours per week and was born in Yogyakarta, then they are predicted as poor workers. A support of 0.03 indicates that this rule appears in 30 trees out of 1,000 trees created in the model, with a rule accuracy of 97% ($Confidence = 0.97$).
2. If someone works less than or equal to 34.5 hours per week, was born in Yogyakarta, and the proportion of working household member is less than or equal to 0.65, then they are predicted as poor workers. A support of 0.02 indicates that this rule appears in 20 trees out of 1,000 trees created in the model, with a rule accuracy of 97%.
3. If someone was born in Yogyakarta, the proportion of working household member is less than or equal to 0.65, and the marital status is married, then they are predicted as poor workers. A support of 0.02 indicates that this rule appears in 20 trees out of 1,000 trees created in the model, with a rule accuracy of 98%.
4. If the proportion of working household member is less than or equal to 0.65 and did not live in Yogyakarta 5 years ago, then they are predicted as poor workers. A support of 0.02 indicates that this rule appears in 20 trees out of 1,000 trees created in the model, with a rule accuracy of 100%.
5. If someone works in the industrial sector and does not have access to credit, then they are predicted as poor workers. A support of 0.02 indicates that this rule appears in 20 trees out of 1,000 trees created in the model, with a rule accuracy of 100%.

CONCLUSION

The results of the simulation study indicate that the DRF model outperforms the RF model in predicting underfit simulation data. Further analysis was conducted by extracting rules from both the RF and DRF models. DRF demonstrates superior performance compared to RF in underfit simulation data in terms of model interpretation based on the support and confidence values of the generated rules. The analysis related to model interpretation based on support and confidence values provides a deeper understanding of the superiority of DRF over RF in terms of accurate prediction and stronger interpretation of the relationship

between explanatory and response variables. The results of the empirical study reveal that the most frequent variable interactions in predicting poor workers in Yogyakarta is the working hours of employees less than the normal working hours (35 hours per week) and being born in Yogyakarta.

REFERENCES

[1] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, 'Classification and Regression Trees. Wadsworth & Brooks/Cole', *Advanced Books & Software*, 1984.

[3] E. Belli and S. Vantini, 'Measure inducing classification and regression trees for functional data', *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 5, pp. 553–569, 2022.

[4] L. Breiman, 'Random forests', *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, 'Random Forests and Decision Trees', *International Journal of Computer Science Issues(IJCSI)*, vol. 9, Sep. 2012.

[6] A. More and D. P. Rana, 'Review of random forest classification techniques to resolve data imbalance', in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, IEEE, 2017, pp. 72–78.

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

[8] A. N. A. Aldania, A. M. Soleh, K. A. Notodiputro, and others, 'A Comparative Study of CatBoost and Double Random Forest for Multi-class Classification', *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 129–137, 2023.

[9] L. Auret and C. Aldrich, 'Empirical comparison of tree ensemble variable importance measures', *Chemometrics and Intelligent Laboratory Systems*, vol. 105, no. 2, pp. 157–170, 2011, doi: 10.1016/j.chemolab.2010.12.004.

[10] F. Doshi-Velez and B. Kim, 'Towards a rigorous science of interpretable machine learning', *arXiv preprint arXiv:1702.08608*, 2017.

[11] O. Bastani, C. Kim, and H. Bastani, 'Interpreting blackbox models via model extraction', *arXiv preprint arXiv:1705.08504*, 2017.

[12] H. Deng, 'Interpreting tree ensembles with inTrees', *International Journal of Data Science and Analytics*, vol. 7, no. 4, pp. 277–287, Jun. 2019, doi: 10.1007/s41060-018-0144-8.

[13] L. Radebe, D. C. M. van der Kaay, J. D. Wasserman, and A. Goldenberg, 'Predicting Malignancy in Pediatric Thyroid Nodules: Early Experience With Machine Learning for Clinical Decision Support', *The Journal of Clinical Endocrinology & Metabolism*, vol. 106, no. 12, pp. e5236–e5246, Dec. 2021, doi: 10.1210/clinem/dgab435.

[14] P. Cadahia Delgado, E. Congregado, A. Golpe, and J. C. Vides, 'The Yield Curve as a Recession Leading Indicator. An Application for Gradient Boosting and Random Forest', *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 7–19, Mar. 2022, doi: 10.9781/ijimai.2022.02.006.

[15] H. Ilma, K. Notodiputro, and B. Sartono, 'Association Rules in Random Forest for the most Interpretable Model', *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0185–0196, Apr. 2023, doi: 10.30598/barekengvol17iss1pp0185-0196.

[16] Badan Pusat Statistik, *Indikator Kesejahteraan Rakyat 2022*. Jakarta: [BPS] Badan Pusat Statistik, 2022.

[17] Badan Pusat Statistik, *Keadaan Angkatan Kerja di Indonesia Agustus 2022*. Jakarta: [BPS] Badan Pusat Statistik, 2022.

[18] K. C.-K. Cheung and K.-L. Chou, 'Working Poor in Hong Kong', *Social Indicators Research*, vol. 129, no. 1, pp. 317–335, Oct. 2016, doi: 10.1007/s11205-015-1104-5.

[19] F. Faharuddin and D. Endrawati, 'Determinants of working poverty in Indonesia', *Journal of Economics and Development*, vol. 24, no. 3, pp. 230–246, Jan. 2022, doi: 10.1108/JED-09-2021-0151.

[20] A. Takemura and K. Inoue, 'Generating explainable rule sets from tree-ensemble learning methods by answer set programming', *arXiv preprint arXiv:2109.08290*, 2021.
[21] R. Agrawal, R. Srikant, and others, 'Fast algorithms for mining association rules', in *Proc. 20th int. conf. very large data bases, VLDB*, Santiago, Chile, 1994, pp. 487–499.
[22] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edition. Boston: Morgan Kaufmann, 2012.
[23] A. Liaw, M. Wiener, and others, 'Classification and regression by randomForest', *R news*, vol. 2, no. 3, pp. 18–22, 2002.
[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[25] P. Probst, M. N. Wright, and A.-L. Boulesteix, 'Hyperparameters and tuning strategies for random forest', *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, p. e1301, 2019.
[26] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science & Business Media, 2011.