BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0185–0196
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0185-0196

ASSOCIATION RULES IN RANDOM FOREST FOR THE MOST INTERPRETABLE MODEL

Hafizah Ilma[1], Khairil Anwar Notodiputro[2*], Bagus Sartono[3]

[1,2,3]*Department of Statistics, IPB University*
Meranti Wing 22 level 4 Dramaga, Bogor, 16680, Indonesia

*Corresponding author's e-mail: * **khairil@apps.ipb.ac.id***

ABSTRACT

Random forest is one of the most popular ensemble methods and has many advantages. However, random forest is a "black-box" model, so the model is difficult to interpret. This study discusses the interpretation of random forest with association rules technique using rules extracted from each decision tree in the random forest model. This analysis involves simulation and empirical data, to determine the factors that affect the poverty status of households in Tasikmalaya. The empirical data was sourced from Badan Pusat Statistik (BPS), the National Socio-Economic Survey (SUSENAS) data for West Java Province in 2019. The results obtained are based on simulation data, the association rules technique can extract the set of rules that characterize the target variable. The application of interpretable random forest to empirical data shows that the rules that most distinguish the poverty status of households in Tasikmalaya are house wall materials and the main source of drinking water, house wall materials and cooking fuel, as well as house wall materials and motorcycle ownership.

Article History:

Received: 17th September 2022
Revised: 12th December 2022
Accepted: 23rd January 2023

Keywords:

Association Rule;
Interpretable Model;
Random Forest;
Rule Extraction.

How to cite this article:

H. Ilma, K. A. Notodiputro and B. Sartono., "ASSOCIATION RULES IN RANDOM FOREST FOR THE MOST INTERPRETABLE MODEL," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0185-0196, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Random forest is an ensemble method for classification and regression derived from a set of decision trees. Each decision tree in a random forest is constructed independently. The final random forest prediction for the classification case is based on aggregating predictions with the majority of votes from all decision trees [1]. Random forest is one of the most popular ensemble methods because random forests can be applied to various prediction problems and produce competitive accuracy. In addition, random forests have customizable parameters, are easy to use, and have the ability to handle small sample sizes and high-dimensional feature spaces [2]. Therefore, the random forest is often used in various applications, such as in agricultural production systems [3], random forest for index tracking [4], and the implements a random forest classifier for Parkinson's disease [5].

Although random forest is good in many areas, it is often criticized for its “black-box” model. This is because there is a trade-off between the accuracy and interpretability of the model [6]. The model's accuracy tends to increase along with the complexity of the rules built by the model. However, the higher the complexity of the model, the more difficult it is for humans to interpret the relationship between the elements in the model.

Interpretability is very important in understanding the relationship among phenomena. It is the degree to which a human can understand the cause of a decision [7]. An interpretable model is a model that can provide a qualitative understanding of the relationship between the value of the independent variable and the resulting response variable [8]. In the case of tree-based classification, the model can be interpreted through tree structures, such as features and thresholds used for splitting in the decision tree model. The decision tree uses a simple if-then decision rule consisting of conditions and predictions. For example, “if it rains today and if it is April (conditions), then tomorrow it will rain (prediction)”. Because a random forest consists of a collection of decision trees, a random forest has a set of rule patterns.

Association rules are the "if-then" statements, a pattern mining technique that aims to get combinations of items that often appear within large data sets and can also be used to find association rules among combinations of items. In this study, the item set referred to the set of decision rules in the random forest model. In [9], a way to get an interpretation in a random forest is proposed. The proposal includes extraction, measurement, and processing the rules generated by a set of decision trees in a random forest. Several studies have used the inTrees framework for tree ensemble interpretation, such as Jimenez et al. [10] using inTrees to explain artificial intelligence in the case of drug discovery and Narayanan et al. [11] using inTrees to understand the characteristics of SSD failures in production datacenters.

The purpose of this study is to enrich the discussion presented by [9] by using different scenarios of simulation data, especially in solving the problem of interpretation in random forests for classification problems. In addition, we also propose an alternative approach in selecting meaningful rules. The interpretable random forest is applied to solve practical problems and determine the factors that affect the poverty status of households in Tasikmalaya.

2. RESEARCH METHODS

2.1 Data and Variables

The data used in this study consisted of simulation data and empirical data. The data simulation design aims to prove the correctness of the association rules method in finding patterns of interest between items. The simulation data is designed in such a way that $Y = 1$ is characterized by conditions ($X_1 = 1$, $X_2 = 1$), ($X_1 = 1$, $X_2 = 1$, $X_3 = 1$), and ($X_1 = 1$, $X_2 = 1$ $X_3 = 2$). While $Y = 0$ is characterized by conditions ($X_1 = 2$, $X_2 = 2$), ($X_1 = 2$, $X_2 = 2$, $X_3 = 1$), and ($X_1 = 2$, $X_2 = 2$, $X_3 = 2$).

Simulation data is generated with the following conditions:

1. Generate three independent variables, with each having two categories so that there are eight combinations of rules formed, as shown in **Table 1**.
2. Select two combinations of rules that characterize the response $y = 1$ and two other combinations of rules that characterize the response $y = 0$.
3. Generate 100 observations based on each combination of rules with the following conditions:

- probability of response $Y= 1$ for two combinations is 0.9,
- probability of response $Y = 1$ for the other two combinations is 0.1,
- probability of response $Y = 1$ for the remaining four combinations is 0.5.

Table 1. The combination of rules formed

X_1	X_2	X_3	$P(Y = 1)$
1	1	1	0.9
1	1	2	0.9
1	2	1	0.5
1	2	2	0.5
2	1	1	0.5
2	1	2	0.5
2	2	1	0.1
2	2	2	0.1

The response variable (y) follows a Bernoulli distribution with the following probability function:

$$(y) = p^y(1-p)^{1-y} \quad (1)$$

with the random variable $y = \{0, 1\}$ and p is the probability of success in one trial.

The empirical case used is to identify household characteristics affecting poverty in Tasikmalaya in 2019. The data used is collected by the Badan Pusat Statistik (BPS), namely, the National Socio-Economic Survey (SUSENAS) data for West Java Province in 2019. The variables used are consisted of one dependent variable and 22 independent variables with categorical data types, as described in **Table 2**.

Table 2. Empirical Data Variables

Code	Variables	Categories
Y	Poor household status	1: Yes; 0: No
X_1	House ownership status	1: Own; 0: others
X_2	House floor area (m^2)	1: Floor area ≤ 35; 2: 35 < floor area ≤ 48; 3: 48 < floor area ≤ 72; 4: floor area > 72
X_3	House roof material	1: Roof tile; 0: others
X_4	House wall material	1: Brick wall; 0: others
X_5	House floor type	1: Ceramic; 0: others
X_6	Ownership of a place to defecate	1: Own; 2: General; 3: None
X_7	The main source of drinking water	1: Bottled water or refill; 2: Well; 3: Springs; 4: Others
X_8	The main source of water for cooking and bathing	1: Bottled water or refill; 2: Well; 3: Springs; 4: Others
X_9	The main source of lighting	1: PLN electricity with a meter; 2: PLN electricity without a meter; 3: Non-PLN Electricity
X_{10}	Cooking fuel	1: LPG 3 KG; 0: others
X_{11}	Receive people's business credit	1: Yes; 0: No
X_{12}	Have a gas cylinder of 5.5 kg or more	1: Yes; 0: No
X_{13}	Ownership of refrigerators	1: Yes; 0: No
X_{14}	Ownership of computers or laptops	1: Yes; 0: No
X_{15}	Ownership of gold jewellery (min 10gr)	1: Yes; 0: No
X_{16}	Ownership of motorcycles	1: Yes; 0: No
X_{17}	Ownership of cars	1: Yes; 0: No
X_{18}	Ownership of flat-screen TVs (min 30 inches)	1: Yes; 0: No
X_{19}	Ownership of land	1: Yes; 0: No
X_{20}	Receiving poor rice	1: Yes; 0: No

Code	Variables	Categories
X_{21}	Receive non-cash food assistance	1: Yes; 0: No
X_{22}	The number of household members	1: household members ≤ 4; 2: 5 ≤ household members ≤ 7; 3: household members > 7

2.2 Model

The model used is a random forest model with 100 decision trees. The random forest as one of the "black box" models causes a lack of model transparency so that the model is difficult to interpret. Although the internal structure of the model is hidden, conceptually, the random forest modeling in this study using the method can be illustrated in **Figure 1**. This study enriches the discussion presented by **[9]**, especially in solving the problem of interpretation in the random forest for classification problems using association rules.

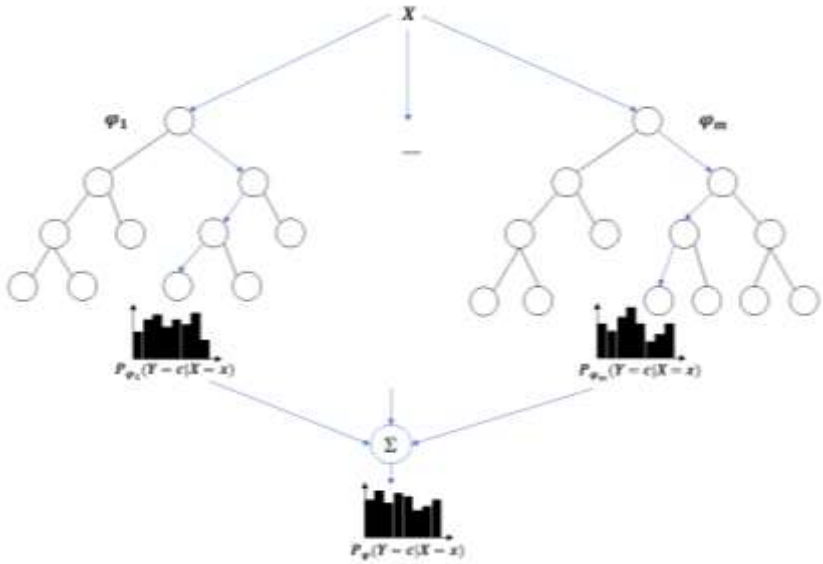

Figure 1. Illustration of the classification process in a random forest

where,

$\boldsymbol{X}$: the predictor variables $(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_{22}})$,
$\boldsymbol{Y}$: the target variable, household poverty status,
$\boldsymbol{c}$: categories of $\boldsymbol{Y}$, poor or non-poor household.
$\boldsymbol{\varphi_m}$: the $\boldsymbol{m}$-th classification tree for $\boldsymbol{m}$-th bootstrap sampling, with $\boldsymbol{m = 1, 2, \ldots, 100}$
$\boldsymbol{P_{\varphi_m}(Y = c|X = x)}$: the prediction probability of $\boldsymbol{Y = c}$ on a certain value of $\boldsymbol{X}$ which is obtained from the $\boldsymbol{m}$-th classification tree, with $\boldsymbol{m = 1, 2, \ldots, 100}$,
$\boldsymbol{\Sigma}$: final prediction using the majority votes.

2.3 Association Rule

Association rule is a technique in pattern mining aimed to obtain repeated relationships in certain data sets so that interesting associations can be obtained between items in the data sets. Association rules are ideally used to explain patterns in data from seemingly independent information repositories, such as relational databases and transactional databases. There are measures of interest to describe the association of the item set in pattern mining **[12]**. An objective measure for association rules of the form $X \Rightarrow$ Y is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. Another objective measure for association rules is confidence, which assesses the degree of certainty of detected association. This is considered a conditional probability $P(Y|X)$, that is, the probability that a transaction containing X also contains Y. More formally, support and confidence are defined as:

$$support(\mathrm{X} \Rightarrow \mathrm{Y}) = P(X \cup Y), \quad (2)$$

$$confidence(\mathrm{X} \Rightarrow \mathrm{Y}) = P(Y|X). \quad (3)$$

2.4 Extracting Rules of Decision Tree

According to the definition in [13], the item set in a decision tree is a rule that is built along the path from the root node to the final node in the decision tree. Below is an example of extracting a branch rule (BR) from a decision tree.

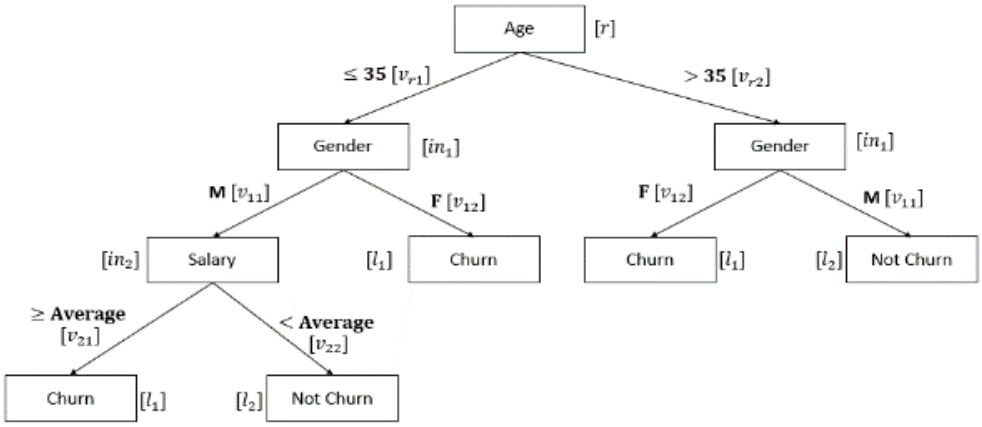

Figure 2. An example of a decision tree structure

Figure 2 provides an example of a tree structure. According to the example, the first rule is:

$$rule_1 = \langle([r] = [v_{r1}])\ ([in_1] = [v_{11}])\ ([in_2] = [v_{21}])([l_1])\rangle$$

where $[r]$ is the root node of the decision tree, $[in_p]$ is the p-intermediate node on the related branch, $[l]$ is the final node (class), and $[v]$ is the value of the intermediate node on the related branch (the value that connects the related node to the next node). Each node and its value pair $([in_1] = [v_{11}])$ are called a branch. In the decision tree in Figure 1, the combination of branch rules formed is as follows:

$$rule_1 = \langle([Age] = [\leq 35])\ ([Gender] = [M])\ ([Salary] = [\geq Average])([Churn])\rangle$$

$$rule_2 = \langle([Age] = [\leq 35])\ ([Gender] = [M])\ ([Salary] = [< Average])([Not\ Churn])\rangle$$

$$rule_3 = \langle([Age] = [\leq 35])\ ([Gender] = [F])([Churn])\rangle$$

$$rule_4 = \langle([Age] = [> 35])\ ([Gender] = [F])([Churn])\rangle$$

$$rule_5 = \langle([Age] = [> 35])\ ([Gender] = [M])([Not\ Churn])\rangle$$

2.5 Data Analysis Procedures

Figure 3 demonstrates the procedures of data analysis used in this paper.

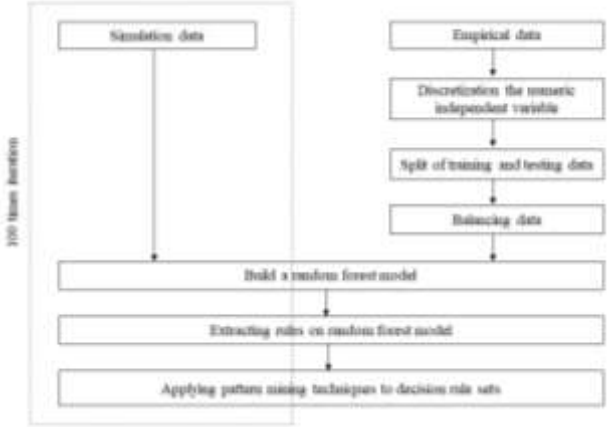

Figure 3. Process Analysis

As shown in **Figure 3**, the generated data was first modeled using a random forest, then the rules were extracted from the random forest and calculated the measure of the rule interest. Meanwhile, for empirical data, the data is preprocessed first, such as discretization, splitting data, and balancing data.

This research uses software R version 4.1.3. The RRF package was used to build a random forest model, and the inTrees package **[14]** extracted the rule tree and calculated the measure of the rule interest. The inTrees package has been used to derive knowledge from tree ensembles such as in **[15]**, **[16]**, and **[17]**.

3. RESULTS AND DISCUSSION

3.1. The Simulation Results

Metrics of Rules

The simulation data consists of three independent variables and one response variable with two categories. Each possible combination of independent variables has been generated 100 times, so the data consisted of 800 rows. The resulted random forest model consisted of 100 decision trees.

Table 3 shows the top 10 rules formed based on the smallest error value. Based on **[9]**, there are three measures of rules' quality, namely length, frequency, and error. Length is the number of an itemset in each condition. Frequency is a measure of the popularity of the rule, which is the proportion of the number of data rows that meet the condition compared to the total number of data rows. The error is the proportion of the number of rows that meet the conditions but have a response variable value which is different from the prediction for the classification problem. For example, for a condition $X_1 = 1$ & $X_2 = 1$, it has a frequency of 0.25 and an error of 0.11. It means that there are 200 rows out of a total of 800 rows of data with conditions $X_1 = 1$ & $X_2 = 1$. There are 11% data with conditions $X_1 = 1$ & $X_2 = 1$ and $Y \neq 1$. This follows the data generation that has been carried out, following the Bernoulli distribution, with the probability $Y = 1$ of 0.9 under these conditions (see **Table 1**). The conclusion obtained from **Table 3** is the six rules with the smallest error are six rules designed to describe the characteristics of the conditions $Y = 1$ and $Y = 0$.

Table 3. Top 10 Rules based on Error Value

Length	Frequency	Error	Condition/ Rule	Prediction
3	**0.125**	**0.100**	**X1=1, X2=1. and X3=1**	**1**
2	**0.250**	**0.110**	**X1=1 and X2=1**	**1**
3	**0.125**	**0.110**	**X1=2, X2=2 and X3=1**	**0**
2	**0.250**	**0.115**	**X1=2 and X2=2**	**0**
3	**0.125**	**0.120**	**X1=2, X2=2 and X3=2**	**0**
3	**0.125**	**0.120**	**X1=1 X2=1 and X3=2**	**1**
2	0.250	0.275	X2=1 and X3=2	1
1	0.500	0.290	X2=1	1
2	0.250	0.290	X1=1 and X3=1	1
1	0.500	0.295	X1=1	1

The Most Frequent Variable Interactions from Rules

As explained in subsection 2.3, the most frequent rules can be seen by two measures of interest, namely, the value of support and the value of confidence. The support value is the probability of a condition from all rules that are formed. The confidence value is a conditional probability, the probability that if a condition occurs, then a particular predictive result occurs. Simulation data generation, random forest modeling, and branch rule extraction were repeated 100 times. After de-duping the same rules, there are 26 unique rule conditions for each iteration. **Table 4** shows the statistical value of the set of rule conditions based on the value of support, confidence, and prediction results with 100 repetitions.

Rules in bold are rules that are set with a probability of 0.9 for each prediction. As shown in **Table 1**, the simulation data with $Y = 1$ characterized by combinations of independent variables i.e. $X_1 = 1$ and $X_2 = 1$, or $X_1 = 1$, $X_2 = 1$ and $X_3 = 1$, or $X_1 = 1$, $X_2 = 1$ and $X_3 = 1$. On the other hand, the simulation data

with $Y = 0$ characterized by combinations of independent variables i.e. $X_1 = 2$ and $X_2 = 2$, or $X_1 = 2$, $X_2 = 2$ and $X_3 = 1$, or $X_1 = 2, X_2 = 2$ and $X_3 = 1$. For example, in **Table 4**, the model predicts the probability that if the conditions are $X_1 = 1, X_2 = 1$ and $X_3 = 1$ then the probability of $Y = 1$ is 100% (confidence). From the 100 iterations of the model, the prediction result of $Y = 1$ in this condition also occurs 100 times (no error). A support value of 5% means the same conditions and predictions are 5% of the total rules formed. The conclusion obtained from **Table 4** is that the six rules that describe the characteristics of $Y = 1$ and $Y = 0$ have an average confidence value of 1.00 with a standard deviation of 0.00. The six rules also predict Y correctly (no errors in 100 iterations). This proves that the association rule technique can extract the rules that characterize the target variable (Y) correctly.

Table 4. The Most Frequent Rules based on Measure of Interest

Condition/rules	Support		Confidence		Number of Predictions (Y=1)
	Average	Standard Deviation	Average	Standard Deviation	
X1=1	0.31	0.05	0.81	0.13	100
X1=1 and X2=1	**0.14**	**0.01**	**1.00**	**0.00**	**100**
X1=1, X2=1 and X3=1	**0.05**	**0.01**	**1.00**	**0.00**	**100**
X1=1, X2=1 and X3=2	**0.05**	**0.01**	**1.00**	**0.00**	**100**
X1=1 and X3=1	0.11	0.03	0.80	0.19	100
X1=1 and X3=2	0.11	0.03	0.81	0.19	100
X2=1	0.30	0.05	0.81	0.13	100
X2=1 and X3=1	0.11	0.03	0.81	0.19	100
X2=1 and X3=2	0.11	0.03	0.80	0.19	100
X1=1, X2=2 and X3=2	0.05	0.01	1.00	0.00	49
X1=2 and X2=1	0.11	0.03	0.81	0.19	49
X1=2, X2=1 and X3=1	0.05	0.01	1.00	0.00	49
X1=2, X2=1 and X3=2	0.05	0.01	1.00	0.00	47
X3=1	0.23	0.03	0.61	0.07	47
X3=2	0.22	0.03	0.59	0.07	47
X1=1 and X2=2	0.11	0.03	0.79	0.19	46
X1=1, X2=2 and X3=1	0.05	0.01	1.00	0.00	45
X1=2	0.31	0.05	0.82	0.14	0
X1=2 and X2=2	**0.14**	**0.01**	**1.00**	**0.00**	**0**
X1=2, X2=2 and X3=1	**0.05**	**0.01**	**1.00**	**0.00**	**0**
X1=2, X2=2 and X3=2	**0.05**	**0.01**	**1.00**	**0.00**	**0**
X1=2 and X3=1	0.11	0.03	0.81	0.19	0
X1=2 and X3=2	0.12	0.03	0.83	0.19	0
X2=2	0.31	0.05	0.82	0.13	0
X2=2 and X3=1	0.11	0.03	0.83	0.19	0
X2=2 and X3=2	0.11	0.03	0.81	0.20	0

3.2. Modeling the Empirical Data

The empirical data consist of 918 observations with 22 explanatory variables and one response variable, the poverty status of households. **Table 5** shows the proportion of poor and non-poor households. Approximately one out of ten households belong to the poor category. Since this is an imbalance condition in the data then a balancing process using the SMOTE (Synthetic Minority Over-sampling) technique needs to be applied to the train data. SMOTE is an oversampling approach by creating a "synthetic" sample. The synthetic sample is created by interpolation between several minority class instances that are within a defined neighborhood **[18]**.

Table 5. The Proportion of Poor Household Status in Tasikmalaya

Poor Household Status	Proportion
Yes	0.099
No	0.901

The random Forest algorithm in this study was conducted on the default hyperparameter, that was, $mtry = \sqrt{number\ of\ predictors} = 5$ and $number\ of\ trees = 100$. The value of $mtry$ defined the number of predictors involved in the best splitting [1]. Table 6 shows the performance measures of the resulting random forest. The modeling of household poverty status using random forest is quite good with an accuracy value of 73.37%.

Table 6. The Performance Measures of the Model

Goodness of Fit	Value
Accuracy	0.7337
Sensitivity	0.7554
Specificity	0.4667
AUC	0.6111

Variables Importance based on Random Forest

Variable importance shows how much a predictor contributes to predicting the response. The measure that can be used to see the importance of a variable in a random forest is the mean decrease in Gini [19]. The higher the value of the mean decrease Gini score, the higher the importance of the variable in the model.

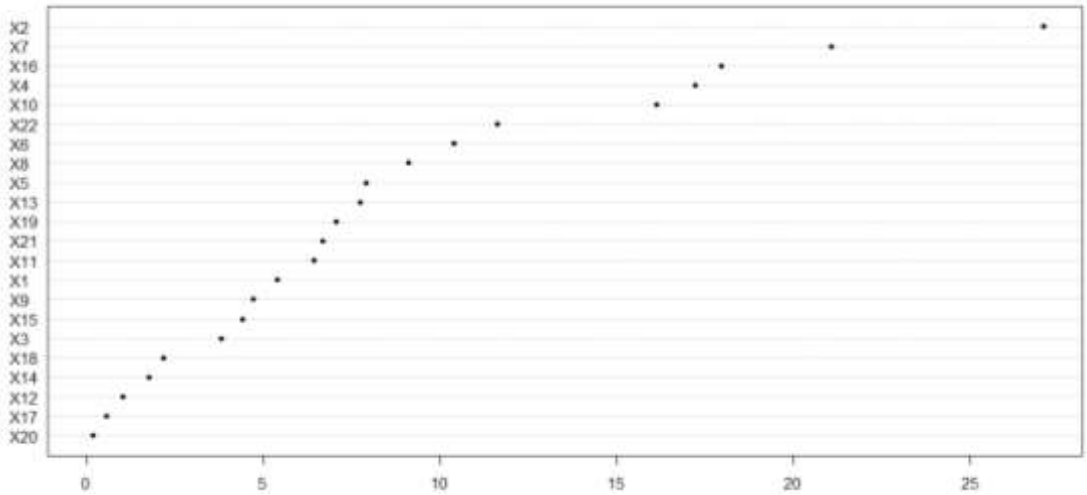

Figure 4. Variables Importance based on Mean Decrease Gini

Figure 4 shows the importance of the variables based on the mean decrease in Gini. The variable house floor area (X_2) is the most important variable in predicting the poverty status of households in Tasikmalaya. The next most important variables are the main source of drinking water (X_7), ownership of motorcycles (X_{16}), house wall material (X_4), and cooking fuel (X_{10}).

The Most Frequent Variable Interactions from Rules based on Measure of Interest

After de-duping the same rules, there are 189 unique rule conditions out of a total of 1275 rules conditions with $2 \leq length \leq 6$ were extracted from the RRF by the condition extraction method in the inTrees package [9]. Of the 189 rules, there are 72 rules with predictions of poor households, the others 117 rules with predictions of non-poor households. Figure 5 shows the distribution of the measure of interest for the 189 unique patterns formed in each class of target.

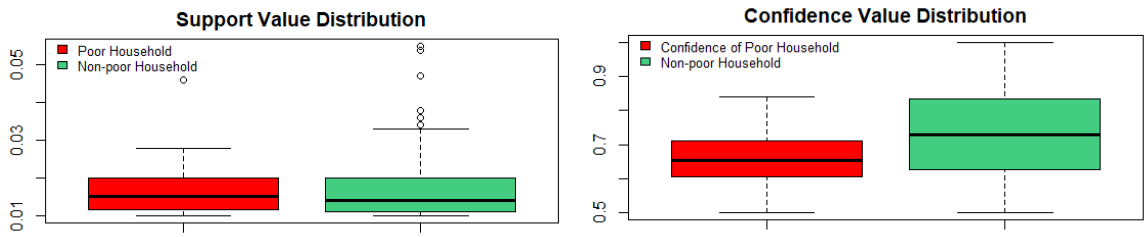

Figure 5. Support and Confidence Value Distribution

As shown in **Figure 5**, the distribution of support values for the two classes of poor and non-poor household status is not significantly different; the median support for the two classes is almost the same. The support values of rules with predictions of poor household status ranged from 0.01 to 0.04, with an average of 0.016. For non-poor household status, the support value for the set of rules ranged from 0.01 to 0.05, with an average of 0.017. Contrariwise, the distribution of confidence values for the two household poverty statuses is slightly different. The confidence value of the set of rules for the poor household ranges from 0.5 to 0.84 with an average of 0.66. Meanwhile, for non-poor households, the confidence value ranges from 0.5 to 1, with an average of 0.73.

Table 7 shows the top rules based on the highest confidence value for the status of non-poor and poor households. For non-poor households, the top three rules have a confidence value of 1.00, but the top rules for poor households only have the highest confidence, respectively, 0.841, 0.828, 0.818, and 0.812. Based on the results in **Table 7**, the rules that most distinguish the poverty status of households in Tasikmalaya are the house wall material and the main source of drinking water. Non-poor households are characterized by house wall material in the 'brick' category and the main source of drinking water in the 'bottled water or refill', 'springs', or other categories (X4=1 and X7= (1,3,4)). Otherwise, poor households are characterized by house wall material in the 'others' category and the main source of drinking water in the 'well' category (X4=0 and X7=2).

Table 7. The Top Rules based on the Highest Confidence Value

Length	Support	Confidence	Condition	Prediction
2	0.013	1	**X4=1 and X7= (1,3,4)**	Non-poor household
2	0.011	1	X10=1 and X7=1	Non-poor household
2	0.01	1	X15=1 and X4=1	Non-poor household
2	0.013	0.841	X10= 0 and X8= (1,2,4)	Poor household
2	0.018	0.828	X10=0 and X19=0	Poor household
2	0.028	0.818	X4=0 and X5=1	Poor household
2	0.023	0.812	**X4=0 and X7=2**	Poor household

Association rule analysis can be used to obtain a set of association rules with minimum support and confidence **[9]**. There are no specific rules to determine the best minimum support and minimum confidence values. Therefore, to select the top rules, first, we separate the rules by target class. Then, we filter the rules based on the multiplication value of the highest support and confidence. **Table 8** shows the top 5 rules in each target class based on the highest support and confidence multiplication value. Based on the results shown in **Table 8**, it can be claimed that the most distinguishing rule characteristics in predicting household poverty status are:

- If a household owns a motorcycle and the wall material of their house is brick, the household is a non-poor household ($X_{16} = 1 \; and \; X_4 = 1 \Rightarrow \mathrm{Y} = 0$). Otherwise, if a household does not own a motorcycle and the wall material of their house is not brick, the household is a poor ($X_{16} = 0 \; and \; X_4 = 0 \Rightarrow \mathrm{Y} = 1$).
- If a household uses 3 KG LPG for cooking fuel and the wall material for their house is brick, it is a non-poor household ($X_{10} = 1 \; and \; X_4 = 1 \Rightarrow \mathrm{Y} = 0$). Otherwise, if a household uses cooking fuel instead of 3 KG LPG and the wall material of their house is not brick, the household is poor ($X_{10} = 0 \; and \; X_4 = 0 \Rightarrow \mathrm{Y} = 1$).

Table 8. Top 5 Rules for Each Target Class based on The Highest ($Support \times Confidence$)

Length	Support	Confidence	Condition	$Support \times Confidence$	Prediction
2	0.055	0.854	**X16=1 and X4=1**	0.047	Non-poor household
2	0.054	0.866	X10=1 and X16=1	0.047	Non-poor household
2	0.047	0.877	**X10=1 and X4=1**	0.041	Non-poor household
2	0.033	0.814	X10=1 and X15=0	0.027	Non-poor household
2	0.038	0.697	X15=0 and X4=1	0.026	Non-poor household
2	0.046	0.781	**X10=0 and X4=0**	0.036	Poor household
2	0.028	0.818	X4=0 and X5=1	0.023	Poor household
2	0.027	0.78	X4=0 and X9=1	0.021	Poor household
2	0.023	0.812	X4=0 and X7=2	0.019	Poor household
2	0.026	0.704	**X16=0 and X4=0**	0.018	Poor household

4. CONCLUSIONS

In this paper, we expand the discussion of the paper [9] by using different scenarios of simulation data, especially on classification problems. This research shows that rules extracted and processed from decision trees in random forests provide results in accordance with the prior expectation. Other alternatives to obtain the best rules by multiplying the support and confidence values are also discussed. The application of interpretable random forest to empirical data shows that the rules that most distinguish the poverty status of households in Tasikmalaya are house wall materials and the main source of drinking water, house wall materials, and cooking fuel, as well as house wall materials and motorcycle ownership.

REFERENCES

[1] I. Nirmala, "Prediction of Undergraduate Student's Completion Status Using Missforest Imputation in Random Forest and XGBoost Models," IPB University, 2021.

[2] G. Biau and E. Scornet, "A Random Forest Guided Tour," *TEST*, 2016.

[3] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–29, 2018, doi: 10.3390/s18082674.

[4] Y. Cao, H. Li, and Y. Yang, "Combining Random Forest and Multicollinearity Modeling for Index Tracking," *Commun. Stat. Comput.*, 2022.

[5] I. Gupta, V. Sharma, S. Kaur, and A. K. Singh, "PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification," *arXiv Prepr. arXiv2203.11287 Search...*, 2022.

[6] A. I. Weinberg and M. Last, "Selecting a Representative Decision Tree from an Ensemble of Decision-Tree Models for Fast Big Data Classification," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0186-3.

[7] T. Miller, "Explanation in Artificial Intelligence: Insights from The Social Sciences," *Artif. Intell.*, vol. 267, 2019, doi: 10.1016/j.artint.2018.07.007.

[8] T. Hastie, R. Tibshirani, G. James, and D. Witten, *An introduction to statistical learning (2nd ed.)*, vol. 102. 2021.

[9] H. Deng, "Interpreting Tree Ensembles with inTrees," *Int. J. Data Sci. Anal.*, vol. 7, no. 4, pp. 277–287, 2019, doi: 10.1007/s41060-018-0144-8.

[10] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug Discovery with Explainable Artificial Intelligence," *Nat. Mach. Intell.*, 2020.

[11] I. Narayanan *et al.*, "SSD Failures in Datacenters: What? When? and Why?," *Proc. 9th ACM Int. Syst. Storage Conf.*, 2016.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.

[13] G. Bakirli and D. Birant, "DTreeSim: A New Approach to Compute Decision Tree Similarity Using re-mining," *Turkish J. Electr. Eng. Comput. Sci.*, 2017.

[14] H. Deng, X. Guan, and V. Khotilovich, "Package 'inTrees,'" 2022.

[15] S. Eskandarian, P. Bahrami, and P. Kazemi, "A Comprehensive Data Mining Approach to Estimate The Rate of Penetration: Application of Neural Network, Rule Based Models and Feature Ranking," *J. Pet. Sci. Eng.*, vol. 156, no. June, pp. 605–615, 2017, doi: 10.1016/j.petrol.2017.06.039.

[16] J. Szlęk, A. Pacławski, R. Lau, R. Jachowicz, P. Kazemi, and A. Mendyk, "Empirical Search for Factors Affecting Mean Particle Size of PLGA Microspheres Containing Macromolecular Drugs," *Comput. Methods Programs Biomed.*, vol. 134, 2016, doi: 10.1016/j.cmpb.2016.07.006.

[17] Gallego-Ortiz, M. C., and A.L, "Using Quantitative Features Extracted from t2-weighted MRI to Improve Breast MRI Computeraided Diagnosis (CAD)," *PLoS ONE 12(11)*, 2017.

[18] A. F. ́andez, S. G. ́ıa, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress andChallenges,

Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, 2018.
[19] H. Han, X. Guo, and H. Yu, "Variable Selection Using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 0, pp. 219–224, 2016, doi: 10.1109/ICSESS.2016.7883053.