khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Performance Comparison of Random Forest, Bagging, and CART Methods in Classifying Recipients of the Family Program in North Aceh

Meri Hari Yanni [1,2], Khairil Anwar Notodiputro[1]*, Bagus Sartono [1]

*correspondence: khairil@apps.ipb.ac.id

[1]Department of Statistics
IPB University
Bogor City

[2]Faculty of Teacher Training and Education
Universitas Bumi Persada
Lhokseumawe City

Abstract- Machine learning is a method in data mining, it is used to study large data patterns through classification methods including Random Forest, Bagging, and CART. The Random Forest method develops the Bagging technique and Decision Tree components (CART) in decision-making. The difference between RF and Bagging is the selection of random features in forming a decision tree. It is only found in RF. Bagging can improve performance, model stability, and reduce variance by forming many different models. The research aims to see the performance of the Random Forest, Bagging, and CART methods in classifying family recipient programs in North Aceh. The results show that the performance of the RF, Bagging, and CART classification methods using the SMOTE technique for handling unbalanced classes is better than before handling unbalanced data. The classification method is evaluated through each model's accuracy, sensitivity, specificity, precision, F1 score, and AUC values. The results show good performance with accuracy values of 90% Smote-RF and 86% Smote Bagging. The best performance was seen in the Smote-RF model which was obtained by tuning the Grid Search CV model parameters with k = 5 and repeat = 1 for a data set proportion of 90:10. This shows that the model can correctly predict all observations with an accuracy percentage of 90% with an average AUC value of 93.52%. On the other hand, the CART method has a very low accuracy value, so the model is less able to accurately predict all observations. Measurement of the level of importance of predictor variables that have the greatest influence in predicting recipient households is the floor area of the house, the number of household members aged 10 years and over, and the type of work of the head of the household.

Keywords: Bagging, CART, Family Program, Random Forest

Article info: *submitted May 16, 2024, revised May 25, 2025, accepted June 24, 2025*

1. Introduction

In the digital age, the exponential growth of big data demands analytical approaches that can effectively manage its complexity and volume. One of the most prominent approaches is machine learning, which offers a flexible framework for building data-driven predictive models. Among its various applications, classification plays a key role. This method aims to group data into specific categories based on patterns derived from historical data, allowing for systematic decision-making across a wide range of fields [1]. Well-known classification algorithms such as Random Forest (RF)**,** Bootstrap Aggregating (Bagging)**,** and Classification and Regression Trees (CART) are valued for their robustness and predictive accuracy in processing large datasets [2]–[4].

One of the key challenges in public policy, especially in developing countries, is the low accuracy in targeting social assistance programs. Mistargeted aid distribution can result in inefficient resource allocation and may exacerbate social inequalities [5]. Classification models in machine learning offer a data-driven approach to improve targeting by leveraging socioeconomic and demographic characteristics [6]. With improved accuracy, limited aid resources can be more precisely directed toward households that truly qualify, increasing the overall impact and sustainability of assistance programs [7].

The Family Program in Indonesia is a conditional cash transfer scheme targeting Extremely Poor Households to promote access to education and health services. Since its implementation in 2012, North Aceh has been one of the regions with the highest number of beneficiaries, with 32,314 households reported. However, official statistics from BPS in 2022 indicate that only 700 households were registered as recipients [8]. Despite the program's intended benefits, field implementation remains problematic ranging from inaccurate recipient data and halted payments to unsynchronized databases and other technical issues [9], [10]. On the recipient side, limited

coverage and inconsistency in assistance indicate potential errors in classification and targeting [11].

This study aims to evaluate the predictive performance of three classification algorithms—RF, Bagging, and CART—in identifying eligible recipients of the Family Program in North Aceh. In addition, it investigates household-level factors that influence eligibility classification. Because the response variable is imbalanced, with relatively few households receiving aid, this study employs the SMOTE (Synthetic Minority Over-sampling Technique) to enhance model training. SMOTE generates synthetic data for the minority class using the K-Nearest Neighbors algorithm, addressing class imbalance during classification [12], [13]. This approach has also been shown to improve the prediction of poverty status in Indonesia [4], [14].

Although RF, Bagging, CART, and SMOTE have been applied in diverse fields such as health [15], [16] and biometrics [17], studies focusing on their use in social assistance targeting—particularly at the local government level in Indonesia—are limited. Therefore, this study seeks to fill this research gap and contribute to the growing body of literature by applying ensemble classification models in the context of poverty targeting.

2. Methods

2.1 The Data

This data uses secondary data originating from the results of Survei Sosial Ekonomi Nasional (SUSENAS) of Badan Pusat Statistik (BPS) Aceh in March 2022. The sample unit in this research is the head of the household in North Aceh. The number of saplings in North Aceh is 700 people.

The variables used in this study consisted of seventeen independent variabels and one response variable. This study's response variable is households that received or did not receive PKH assistance in 2022. All variables used in this research are shown in Table 1.

2.2 The Model

This research uses RF, Bagging, and CART methods for modeling. This method is used to classify households who received or did not receive family programs in North Aceh. The explanatory variables used are variables that influence the status of family program recipients in North Aceh, namely town or village classification, ownership status of residential building, house floor area, house wall materials, main building materials of house floor, defecation facilities, the main water source used for drinking, the main source of household lighting, owning a motorcycle, own land, the largest source of financing in the household, number of household members, household members aged 5 years and above, household members aged 10 years and above, type of work, working status of the household head. These variables consist of four continuous variables and twelve categorical variables.

Before creating the classification model, researchers divided the observational data into three data distribution conditions: (I) 60% training data and 40% testing data, (II) 75% training data and 25% testing data, and (III) 90% training data and 10% testing data. These three conditions will be applied to the RF, Bagging, and CART models.

2.2.1 Random Forest (RF)

RF is an ensemble method that builds many decision trees using the bootstrap technique and makes a final prediction based on majority voting for classification. In each tree, the selection of variables used at each branch is done randomly, thereby increasing model diversity and reducing correlation between trees. Random Forest is known to be effective in handling data with many features and can overcome the problem of overfitting that commonly occurs in single trees.

In the previous explanation, data splitting was performed under three data division conditions. After that, bootstrap resampling was performed on the specified training data. The researcher also set three alternative values for ntree (number of trees), namely 50, 500, and 1000. Meanwhile, the mtry value (number of variables considered for splitting in each tree) was tested using three alternatives, namely 2, 4, and 8. The determination of the mtry value was based on the suggestion from Breiman and Cutler that the ideal number of splitting variables is $\sqrt{p}$, where p is the total number of predictors [18]. The combination of the three data splitting conditions, ntree values, and mtry was then tested to find the optimal model.

Table 1. Research Variables

Symbols	Variables	Category
Y	Recipients of the Family Program	(0) No, (1) Yes
X_1	Town or Village Classification	(1) City, (2) Village
X_2	Ownership status of residential building	(1) Owned, (2) Contract/lease, (3) Rent-free, (4) Agency, (5) Other
X_3	House Floor Area	Square meters
X_4	House Wall Materials	(1) Wall, (2) Plastered woven bamboo/wire, (3) Wood/board, (4) Woven bamboo, (5) Log, (6) Bamboo, (7) Other
X_5	Main building materials of house floor	(1) Marble/granite, (2) Ceramic, (3) Parquet/vinyl/carpet, (4) Tile/seal/terrazzo, (5) wood/board, (6) cement/red brick, (7) bamboo, (8) soil (9) other
X_6	Defecation Facilities	(1) available, used only by own ART, (2) available, used with specific household ART, (3) available, at communal MCK, (4) available, at public MCK/anyone uses, (5) available, ART does not use, (6) no facilities
X_7	The main water source used for drinking	Branded bottled water, (2) Leading refill water, (3) borehole/pump, (4) protected well, (5) unprotected well, (6) protected spring, (7) unprotected spring, (8) Surface water (river/lake/reservoir/pond/irrigation), (9) Rainwater, (10) Other
X_8	The main source of household lighting	(1) PLN electricity with a meter, (2) PLN electricity without a meter, (3) non-PLN electricity, (4) no electricity

X_9	The main type of fuel used for cooking	(0) no cooking at home, (1) Electricity, (2) LPG 5.5 Kg/blue gas, (3) LPG12 kg, (4) LPG 3 Kg, (5) City gas, (6) Biogas, (7) Kerosene, (8) Briquettes, (9) Charcoal, (10) Firewood, (11) Other.
X_{10}	Owning a Motorcycle	(1) Yes, (5) No
X_{11}	Own land	(1) Yes, (5) No
X_{12}	The largest source of financing in the household	(1) working households, (2) remittances/goods, (3) investments (deposits, royalties, stocks, bank interest, and the like), (4) pensioners
X_{13}	Number of Household Members	Total
X_{14}	Household members aged 5 years and above	Total
X_{15}	Household members aged 10 years and above	Total
X_{16}	Type of Work	(1) Agriculture of rice and secondary crops, (2) Horticulture, (3) Plantations, (4) Fisheries, (5) Livestock, (6) Forestry and other Agriculture, (7) Mining, (8) Processing Industry, (9) Electricity, gas, steam/hot water, and cold air supply, (10) Water management, wastewater management, waste management and recycling, and remediation activities, (11) Construction, (12) Wholesale and retail trade, repair and maintenance of cars and motorcycles, (13) Transportation and warehousing, (14) Provision of accommodation and provision of food and drink, (15) Information and communication, (16) Financial and insurance activities, (17) Real estate, (18) Professional, scientific, and technical activities, (18) Rental and leasing activities without option rights, employment, travel agencies, and other business support, (20) Public administration, defense, and compulsory social security, (21) Education, (22) Human health and social activities, (23) Arts, entertainment, and recreation, (24) Other service activities, (25) Activities of households as employers, (26) Activities of international and other extra-international bodies.
X_{17}	Working Status of the household head	(0) No activity, (1) Working

In addition, the mtry value was also adjusted using the Grid Search Cross-Validation technique. The prediction results were obtained using the majority vote technique, and the model that provided the highest accuracy value for each combination of conditions was selected as the best model.

2.2.2 Bagging (Bootstrap Aggregating)

The Bagging method was also used in this study with an approach like Random Forest. Bagging is a method that generates multiple versions of a model from predictors using bootstrap replication techniques, resulting in an aggregated estimator. The purpose of this method is to reduce the variance of the final estimator. Bagging works by forming several models from training data obtained through the bootstrap technique. This approach will provide B different training data sets, from which prediction models will be created for each training data set to obtain the estimator model. The prediction results from each training data set are averaged, which can be written as follows:

$$\hat{f}_{bag}(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(X)$$

The steps involved include dividing the observation data into training data and testing data, just like in RF modeling. Next, a bootstrap resampling process is performed on the training data to form several decision tree models. The final classification result is determined based on an aggregation technique called majority vote. Unlike Random Forest, Bagging does not perform random selection of variable subsets at each split; all predictor variables are fully considered in each tree. Therefore, the Bagging method is more suitable for cases with a limited number of predictor variables, as it tends to produce simpler yet stable models [19].

2.2.3 CART

Likewise with the CART model, because this research uses categorical response variables, it uses a classification tree in the model [20]. In the CART model, the first step is to compile the branch candidates, the preparation is carried out on the complete predictor variables. Second, assess all branch candidates by calculating the amount of Q (S | t). Third, determine the branch candidate that has goodness of fit Φ (S | t). When the decision node no longer exists, the CART algorithm process is stopped. Goodness Φ (S | t) of candidate branch s at decision point t, is defined by the equation.

$$\Phi\,(S\,|t) = 2P_L 2P_R Q\,(S\,|t)$$
$$\Phi\,(S\,|t) = 2P_L 2P_R Q\,(S\,|t) \quad (1)$$

$$Q\,(S\,|t) = \sum_{j=1}^{number\ of\ categories} |P(j|t_L) - P(j|t_R)|$$
$$Q\,(S\,|t) = \sum_{j=1}^{number\ of\ categories} |P(j|t_L) - P(j|t_R)| \quad (2)$$

with $P(j|t_L) = \frac{m}{p} P(j|t_L) = \frac{m}{p}$, $P(j|t_R) = \frac{r}{p} P(j|t_R) = \frac{r}{p}$, $P_L = \frac{c}{d} P_L = \frac{c}{d}$, $P_R = \frac{e}{d} P_R = \frac{e}{d}$. Let $t_L t_L$ is left branch from decision node t, $t_R t_R$ is Right branch from decision node t, m is number of records in category j in the right branch candidate t_R t_R, r is number of records in category j in the right branch candidate $t_L t_L$, p is number of records at decision node t, c is number of records in the left candidate $t_L t_L$, d is number of records in training data, and e is number of records in the left candidate $t_R t_R$.

Because there are unbalanced classes in the response variable, it is possible to produce low accuracy values. So, the SMOTE method is used to overcome class imbalance and increase model accuracy. This method is applied to three classification models, namely the RF, Bagging, and CART models.

2.2.4 Confusion Matrix

Furthermore, the evaluation measures used in this study are the values of accuracy, sensitivity, specificity, F1 Score,

Precision, and AUC. The goodness of evaluation can be analyzed using a confusion matrix [21]. Evaluation is used to see the level of error that occurs in the classification of the sample area so that the percentage of accuracy can be seen. The calculation process for these four measures is based on the confusion matrix given in Table 2.

Table 2. Confusion Matrix

		Prediction Result Data	
		Good	*Bad*
Actual Data	*Good*	True Positive (TP)	False Positive (FP)
	Bad	False Negative (FN)	True Negative (TN)

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

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

$$\text{Specificity} = \frac{TN}{TN+FP} \frac{TN}{TN+FP} \quad (5)$$

Apart from that, we will also look at the proportion of positive cases that are correctly predicted against all positive predictions through precision [22] and the use of the F1 score to detect False Negatives and False Positives in cases of imbalanced data [23], [24]. The formula for calculating precision and F1 Score is as follows.

$$\text{Presisi} = \frac{TP}{TP+FP} \frac{TP}{TP+FP} \quad (6)$$

$$\text{Skor F1} = \frac{2\times presisi\times recall}{presisi+recal} \frac{2\times presisi\times recall}{presisi+recal} \quad (7)$$

2.3 The Procedure of Data Analysis

The analysis steps in this research area:

a. Perform pre-processing. The status of households that in the last year still received the family program was categorized as 1 and 0 as not receiving it.
b. Data exploration. General description of family program recipient data in North Aceh and other independent variables
c. Checking for missing data values
d. Splitting data into training data and testing data: training data is used for modeling and testing data is used to evaluate classification performance.
   i. Splitting data consists of three proportion parts, namely 60:40, 75:25, and 90:10. In RF, set mtry and ntree for each proportion. The combination of mtry, ntree, and splitting data produces accuracy values.
   ii. Tuning hyperparameters of the Grid Search CV model to produce optimal features with the highest accuracy.
   iii. From stages i and ii, the best mtry was obtained through the highest accuracy value.
   iv. Determination of the next RF classification model based on mtry in stage iii.
e. Carry out RF, Bagging, and CART classification modeling using training data.
f. Evaluate classification performance using the values in the confusion matrix from modeling results at stage d, namely sensitivity, specificity, accuracy, precision, F1 score, and AUC for the three classification methods.
g. To improve the proportion of data in the minority class using SMOTE, the SMOTE method is applied to the training data obtained from stage d.
h. Carry out RF, Bagging, and CART classification modeling using training data in stage g
i. Evaluate classification performance using the confusion matrix on the results of stage h.
j. Comparing the performance of RF, Bagging, and CART classification methods before and after handling imbalanced data in terms of sensitivity, specificity, accuracy, precision, and F1 score
k. Interpretation of the relationship between important variables and household status.

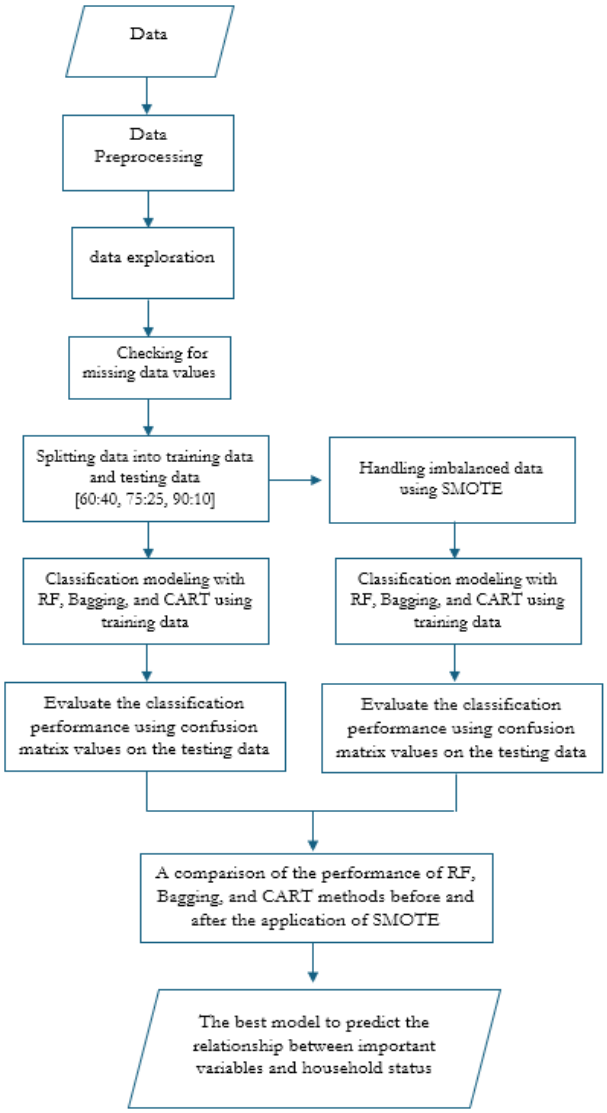


Figure 1. Data Analysis Procedure Flowchart

3. Result

3.1 Data Exploration

The data used is household data in the last year whether they received the Family Program with a sample size of 700 observations. After preprocessing the data, no missing values were found in the observations. The data proportion of

households receiving the family program is 22.29% (156 observations) and households that do not receive PKH is 77.71% (544 observations) as shown in Figure 2. The difference in proportion between receiving and not receiving the family program in Figure 1 shows that the data is unbalanced, where the minority class receives the family program, and the majority class does not receive the family program. this is called unbalanced data. Apart from that, data exploration was also carried out to see the distribution of explanatory variables in each class of response variables, as in Figure 3.

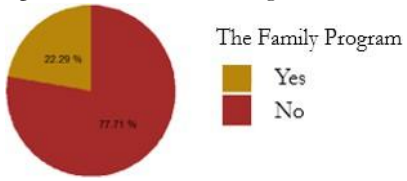


Figure 2. **Pie Chart Percentage of heads of household receiving family program assistance**

3.2 Random Forest Classification Modeling

RF modeling begins by determining the mtry and ntree that will be used. The mtry values are 2, 4, and 8 with p being the number of explanatory variables of 17. The ntree values used are 50, 500, and 1000. Modeling is carried out with 5-fold cross-validation. Apart from that, the formation of the RF model is not only a combination of m and ntree, but there is a combination of dividing the data set, namely the proportion of training data and testing data of 60:40, 75:15, and 90:10. The selection of the optimal combination of mtry, ntree, and split data is seen based on the highest accuracy value. In Table 3, the combination of parameters in random forest modeling that provides the best performance is given.

Based on Table 3, the best RF model is produced if modeling is carried out using mtry = 2 and ntree of 50, 500, 1000 with a proportion of training data and testing data of 90: 10. This combination produces the highest accuracy value compared to the other parameters, namely 79 .71%. This is also comparable to the accuracy value obtained by using the Search CV grid model parameter tuning, namely 79.71% with mtry = 8. Furthermore, compare each mtry and ntree by looking at the sensitivity and specification values. This comparative performance can be seen in Table 4.

Based on Table 4, mtry = 8 obtained a higher sensitivity value compared to mtry = 2, amounting to 33.33%. So mtry = 8 and ntree = 1000 with a proportion of training data and testing data of 90:10 is the basis for forming a classification model. Even though the sensitivity value is quite low, this indicates that mtry = 8 is the most optimal mtry to provide the best RF performance

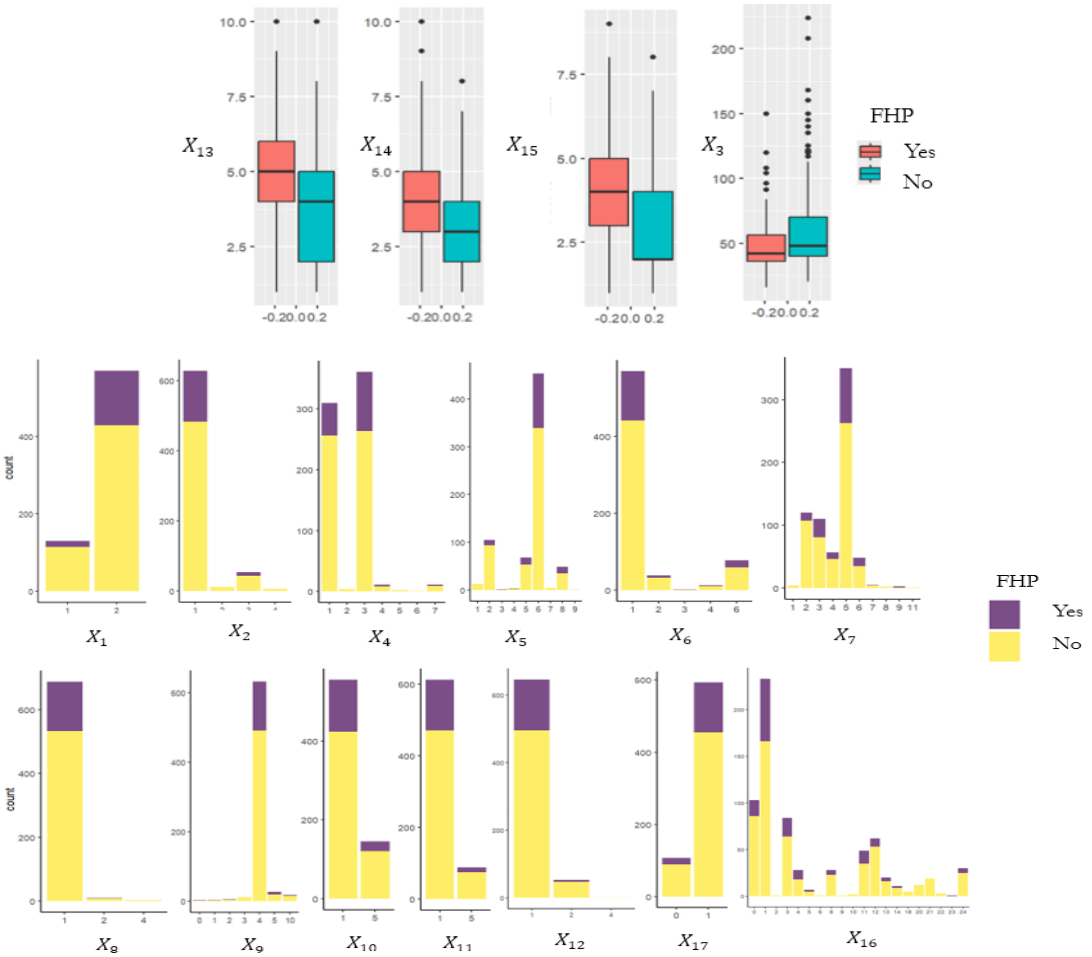


Figure 3. **Distribution of Explanatory Variables in Each Class of Response Variable**

Table 3. Optimal Parameter Combination of Random Forest Models

Proportions (Training Data : Testing Data)	Model		Accuracy
	mtry	ntree	
60 : 40	4	50	75,99%
		500	75,99%
Dimensions:		1000	75,27%
Training Data 421 : 18	2	50	75,99%
Testing Data 279 : 18		500	76,7%
		1000	77,02%
	8	50	76,34%
		500	74,91%
		1000	76,34%
75 : 25	4	50	76.57%
		500	77,71%
Dimensions:		1000	77,71%
Training Data 525 :18	2	50	70%
Testing Data 175 : 18		500	79.43%
		1000	78.86%
	8	50	76.57%
		500	78.86%
		1000	78.86%
90 :10	4	50	71.16%
		500	72.61%
Dimensions:		1000	72.61%
Training Data 631:18	2	50	79.71%
Testing Data 69: 18		500	79.71%
		1000	79.71%
	8	50	78.26%
		500	76.81%
		1000	78.26%
Tuning Parameter model Grid Search CV: with k=5 and repeated =1, obtained mtry Optimal = 8			79,71%

One of the causes of low RF sensitivity values is the problem of unbalanced data. In Table 4, the RF sensitivity value is 0.333. This value is small enough to see the performance of the classification model, namely the model is only able to correctly classify 33.3% of households that receive the family program. Class imbalance can affect prediction results, so it is necessary to handle it using the SMOTE technique on training data. Figure 4 is a line diagram that presents differences in household class proportions in the training data after handling class imbalance. After using SMOTE, the household class becomes balanced because of the synthetic data creation process for the majority class, namely the class that does not receive the family program.

Table 4. RF Performance Based on Optimal Mtry with Ntree = 1000

RF Performance	Mtry = 2	Mtry = 8 (Grid Search CV Model)
Accuracy	79,71%	79,71%
Sensitivity	20%	33,33%
Specificity	90,29%	90,74%

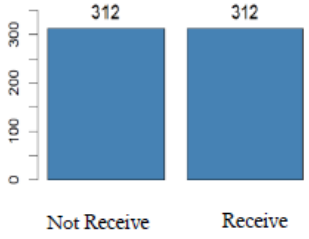


Figure 4. The Proportion of Classes Receiving Family Programs After Implementing the SMOTE Method

3.3 Classification Analysis of RF, Bagging and CART with SMOTE Method

After the training data is handled using SMOTE, the classification modeling process is then carried out in RF, Bagging, and CART. The model formed is then tested on test data. The test results form a confusion matrix as presented in Figure 5 for each classification method with and without SMOTE.

In Figure 5, the RF method for the 69 heads of households observed, the RF classification model was only able to correctly predict 5 heads of households who received aid, and 10 households who received aid were predicted incorrectly as not receiving family program aid, while for heads of households. Those who did not receive assistance were correctly predicted not to receive assistance in as many as 49 households. Thus, this model is not good for use in classification models. Furthermore, the confusion matrix in the RF-SMOTE model was obtained from the results of 69 heads of household who were observed, the RF-SMOTE classification model was able to correctly predict 15 heads of household receiving assistance, and no households receiving assistance were predicted incorrectly. Meanwhile, 47 heads of households who did not receive assistance were predicted correctly and 7 heads of households did not receive actual family program assistance but were predicted to receive it.

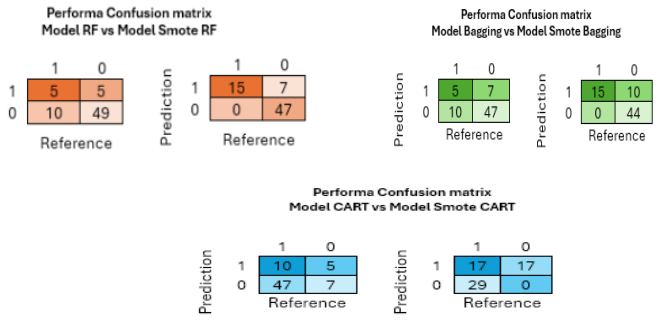


Figure 5. The Proportion of Classes Receiving Family Programs After Implementing the SMOTE Method

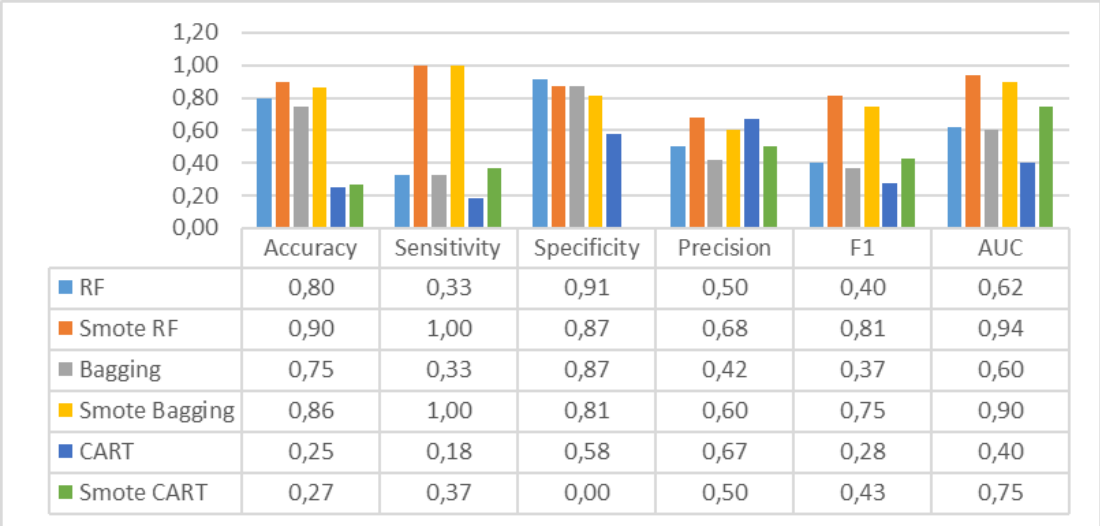
3.4 Classification Analysis of RF, Bagging and CART with SMOTE Method

The performance of the classification method was evaluated by calculating the accuracy, sensitivity/recall, specificity, precision, F1 score, and AUC values. These values can be a comparative measure of how well the classification method is in predicting the observation class in the test data. The calculation results of various performance measures of the classification methods are then compared in one table and graph as presented in Figure 5.

In Figure 6 the sensitivity values of the RF, Bagging, and CART methods after SMOTE increase. This shows that the performance of the classification method that has been SMOTE is better than before SMOTE. The highest sensitivity value reaches 1.0 on RF and Bagging models that have been SMOTE. This shows that the model correctly predicted households receiving family programs in the last year, namely 100%. However, the ability of the CART classification method after SMOTE to identify households receiving family programs is only limited to 0.4 or 40%. On the other hand, this classification method is very good for predicting households that do not receive family program assistance. This is indicated by specificity values above 80%, except for the Bagging, CART, and SMOTE CART models. Overall, the highest accuracy value is in the SMOTE RF method with a value of 0.9. This shows that the model can correctly predict all observations with an accuracy percentage of 90%. Meanwhile, other comparison methods have lower accuracy values compared to the SMOTE RF method. On the other hand, the CART method has a very low accuracy value, so the model is less able to predict accurately all observations.

F1 score is the harmonic mean of precision and recall (sensitivity). The use of the F1 score is to see the performance of classification methods in predicting family program recipients from the perspective of precision and sensitivity. Of all the methods, the highest F1 score value is SMOTE RF at 0.81 or 81%.

It can be seen in Figure 6 that the F1 score becomes smaller in the SMOTE CART, RF, and Bagging methods. This is due to the large number of prediction errors received when predicting family program recipients, shown in the low precision values of the four models. However, this low precision value is not a problem if the aim of predicting recipients is as a preventative measure against households indicated as not receiving assistance by the model. As a result, North Aceh can pay more attention to households that are indicated to have not received the Family Program in the past year as a form of reducing poverty rates in the following year among communities in North Aceh.



	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
RF	0,80	0,33	0,91	0,50	0,40	0,62
Smote RF	0,90	1,00	0,87	0,68	0,81	0,94
Bagging	0,75	0,33	0,87	0,42	0,37	0,60
Smote Bagging	0,86	1,00	0,81	0,60	0,75	0,90
CART	0,25	0,18	0,58	0,67	0,28	0,40
Smote CART	0,27	0,37	0,00	0,50	0,43	0,75

Figure 6. The Proportion of Classes Receiving Family Programs After Implementing the SMOTE Method

3.5 Importance Variables

In Figure 6, the best performance of the classification model is obtained compared to other classification models, namely the RF Model with SMOTE for the case of Family Program Recipients in North Aceh. Determining variable importance for family program recipients refers to the SMOTE RF model. This aim is to see the level of importance of each variable used in modeling. Of the seventeen predictor variables, three predictor variables have the highest contribution in classifying PKH recipient households, namely the floor area of the house, the number of household members aged 10 years and over, and the type of occupation of the head of the household. The relationship between these variables and households receiving family program assistance can be seen in Figure 7. The relationship between the floor area of the house in households receiving assistance has a negative relationship, which means that

the smaller the house occupied, the greater the chance that the household will be categorized as an aid recipient. Meanwhile, the number of household members aged 10 years and over has a positive relationship, namely the more household members aged 10 years and over, the greater the chance that the household will be included in the aid recipient category.

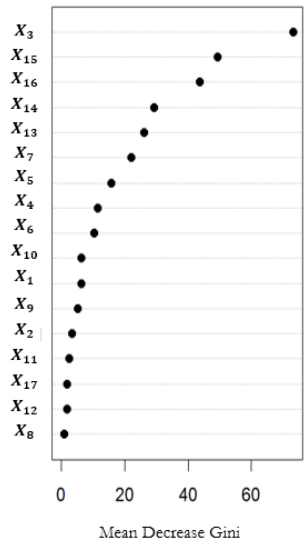


Figure 7. The Proportion of Classes Receiving Family Programs After Implementing the SMOTE Method

4. Conclusion

This study compares the performance of Random Forest (RF), Bagging, and CART classification methods in predicting households receiving assistance from the Family Program in North Aceh. The results of the three classification methods, both before and after applying the SMOTE technique to address data imbalance, show that the application of SMOTE can improve model performance. Among the three classification methods that use SMOTE, the RF method shows the best performance in predicting beneficiary households in North Aceh. Overall, the SMOTE-RF method produces the highest accuracy value of 0.9, indicating that the model is able to classify 90% of observations accurately. Conversely, the CART method shows the lowest performance and is less able to predict accurately. In the SMOTE-RF model, the three predictor variables with the highest contribution to the classification of households receiving assistance are: floor area of the house, number of household members aged 10 years and above, and type of occupation of the head of the household.

This study makes an important contribution in the context of applying data mining to social policy, demonstrating the effectiveness of the SMOTE technique in improving classification performance on imbalanced data. The results of this study also confirm the superiority of the Random Forest algorithm when combined with balancing techniques and hyperparameter optimization, particularly in identifying vulnerable households targeted for assistance. Additionally, these findings can be utilized by policymakers and social program managers in designing more targeted and data-driven assistance distribution systems. The identification of key variables also provides valuable insights to support more accurate and efficient social intervention planning.

Acknowledgement

The authors would like to express their gratitude for the Indonesian Education Scholarship (BPI) which has supported the education fund during the doctoral study process at the Department of Statistics, IPB University.

Reference

[1] L. Breiman and A. Cutler, "Manual on Setting Up, Using and Understanding Random Forest V4.0," [Online].

[2] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

[3] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, pp. 123–140, 1996.

[4] K. Sukarna, K. A. Notodiputro, and B. Sartono, "Comparison of Logistic Model Tree and Random Forest on Classification for Poverty in Indonesia," Media Statistika, vol. 16, no. 2, pp. 112–123, Dec. 2023.

[5] S. Kidd and E. Wylde, "Targeting the poorest: An assessment of the proxy means test methodology," Overseas Development Institute, London, 2011.

[6] A. Aribowo, R. Kuswandhie, and Y. Primadansa, "Implementasi Algoritma CART dalam Penentuan Kelayakan Penerima Bantuan PKH di Desa Ngadirejo," Cogito Smart Journal, vol. 7, no. 1, pp. 40–51, 2021.

[7] A. A. Aziz, R. Royani, and S. Syukriati, "The Implementation of Family Hope Program in Social Protection and Welfare in West Lombok," Journal of The Community Development in Asia, vol. 4, no. 3, pp. 1–11, 2021.

[8] Herianto, "Bansos PKH Untuk Aceh Tahun 2022 Senilai Rp 825 M, Disalurkan Minggu Pertama Februari 2022," Tribun News, 2022. [Online]. Available: https://aceh.tribunnews.com/2022/01/31/bansos-pkh-untuk-aceh-tahun-2022-senilai-rp-825-m-disalurkan-minggu-pertama-februari-2022

[9] F. Pebrianto and D. R. Cahyani, "Mensos Mengakui Penyaluran PKH Masih Bermasalah," Tempo.co, 2019. [Online]. Available: https://bisnis.tempo.co/read/1282593/mensos-mengakuipenyaluran-pkh-masih-bermasalah

[10] Kementerian Sosial Republik Indonesia, Laporan Evaluasi Program Keluarga Harapan, Jakarta, 2021.

[11] BPS, Aceh Utara dalam Angka 2022, Badan Pusat Statistik Kabupaten Aceh Utara, 2022.

[12] 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.

[13] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor," Jurnal ISD, vol. 3, no. 1, pp. 44–49, 2018.

[14] M. Cakmak, "Heart Disease Classification Using Random Forest Machine Learning," All Sciences Proceedings, 2024.

[15] A. Accardo et al., "Toward a Diagnostic CART Model for Ischemic Heart Disease and Idiopathic Dilated Cardiomyopathy Based on Heart Rate Total Variability," Medical & Biological Engineering & Computing, vol. 60, pp. 2655–2663, 2022.

[16] M. A. Yaman, R. Rattay, and A. Subasi, "Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Face Recognition," Procedia Computer Science, vol. 194, pp. 202–209, 2021.

[17] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," International Journal of Computer Applications (IJCA), vol. 163, no. 8, pp. 15–19, 2017.

[18] A. Aulia, D. Jeong, I. M. Saaid, D. Kania, M. T. Shuker, and N. A. El-Khatib, "A Random Forests-based sensitivity analysis framework for assisted history matching," *Journal of Petroleum Science and Engineering*, vol. 181, Article 106237, May 2019, doi: 10.1016/j.petrol.2019.106237.

[19] Z.-H. Zhou, "Ensemble Methods: Foundations and Algorithms". London: CRC Press, 2012

[20] A. Riandhoko, N. Amalita, D. Vionanda, and A. Salma, "Handling unbalanced data with SMOTE algorithm for unemployment classification in Lima Puluh Kota Regency using CART method," *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 2, pp. 166–177, 2024, doi: 10.29244/ijsa.v8i2p166-177.

[21] P. Webb, and Yohannes, "Classification and Regression Trees, CART", International Food Policy Research Institute, Washington D.C, 1999.

[22] J. Han, J. Pei, and M. Kamber, "Data Mining: Concepts and Techniques". Elsevier, 2012.

[23] D. M. W. Powers, "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness, And Correlation", Journal of Machine Learning Technologies. 2(1):37-63, 2011.

[24] J Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences*, vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.