Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of main characteristics of food insecurity using classification Tree and Random Forest

E Ramadhani[1*], B Sartono[2], A F Hadi[3], S 'Ufa[4], T Akhdansyah[5]
[1,4,5]Syiah Kuala University, Banda Aceh, Indonesia
[2]Bogor Agriculture University, Bogor, Indonesia
[3,]Jember University, Jember, Indonesia
* evi.ramadhani@unsyiah.ac.id

Submitted : Oct 14, 2022 | **Accepted** : Oct 17, 2022 | **Published** : Oct 19, 2022

Abstract. Since the emerging of big data era, the information and data are grown rapidly. It requires us to have ability to extract the knowledge and information that consisted in this explosion of the data. One of way that can be used for this purpose is by using machine learning method. One of purpose of machine learning implementation is to conduct classification analysis and to identify variable importance that contribute in the research. It's conducted the comparative study between two machine learning classification methods named classification tree and random forest method. This study is implemented on Indonesian Socioeconomic Survey (SUSENAS) 2020 in Aceh Province. The purpose of the study is to identify the optimum method between both and to identify the characteristics of food insecure household. The optimum method obtained by comparing the AUC value. The results obtained is random forest outperformed classification tree with the AUC value of random forest method is 0,718 and classification tree method is 0,668. The rank of variable importance of the optimum method is the type of cooking fuel used in the household, the area of house floor, education level of head of household, number of savers in a household, and the type of house floor.

Keywords: AUC, Classification Tree, Food Insecurity, Random Forest

INTRODUCTION

Machine learning (ML) is an algorithm that is used to increase the accuracy of data analysis automatically. In building the model, algorithms are used in machine learning to classify and make predictions based on training data. In the machine learning process, it works more efficiently by developing its algorithm independently, making it easier for humans to work (Ethem, 2020).

The way the machine learning algorithm works is known as computational learning. This is due to the limited training data set which is limited by the opportunities in other methods in general. One of the tools to measure error in the analysis is bias-variance decomposition. Machine learning is useful for solving world problems in a scalable way. Another advantage of machine learning is working on data by improving machine or system performance and able to create their own algorithm.

The division of machine learning algorithms is generally divided into three types, namely supervised learning, unsupervised learning and reinforcement learning (Ethem, 2020). The supervised learning

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

algorithm is formed by determining and labelling the desired input and output variables. Supervision is carried out through iterations that are carried out in optimizing the objective function which can monitor the output results that are affected by the new input. This algorithm can increase the level of accuracy of predictions on an ongoing basis in carrying out their duties (Mohri et al., 2012). Supervised learning algorithms are divided into two types, namely classification and regression. Regression can be used at the time of analysis for predicting or modelling purposes. While classification is more appropriate to be used to verify or classify objects into existing groups. Classification algorithms include nave bayes classifier, SVM, logistic regression, KNN, decision tree and random forest. Each method has its own advantages. Classification methods that are very often used are classification tree and random forest.

Classification Tree is a classification method with an exploratory technique that produces a decision tree. With a smaller error rate, each new tree is used to classify based on the response. Some of the advantages of tree classification include that the previous decision tree is very complex and general can be made simpler and more detailed, calculations are carried out based on needs so that it can summarize or shorten the operating process, test samples based on certain criteria

Random forest is a classification method that builds a lot of trees during training in its operation. In doing its job, the algorithm chooses a large tree-forming class. The random decision forest also makes corrections to the resulting tree in the training data set. The accuracy of the classification for the same hyper parameters and data produces different results with more than one operation even though they are carried out sequentially or at close times. This is one of the weaknesses of the random forest, the accuracy value is not stable. In this study, the classification of household indicators of food insecurity in Aceh Province will be carried out by comparing the random forest and classification tree methods based on the level of accuracy, as well as seeing the similarity of the importance variable generated from the two methods.

To conduct classification method, the class stability of response variable must first be inspected. The instability of responses variable class will make less accurate decision. Therefore, this problem should be controlled. There are a lot of ways to handle this problem, one of them is *random oversampling* method (Han et al., 2012). Random oversampling is a resampling method that works by resampling the data in so that the resulting training set contains an equal number of positive and negative, so the total of every class of the data would be equal.

This study is implemented in food insecurity data in Aceh Province 2020. The characteristics of the data can be seen on the figure below.

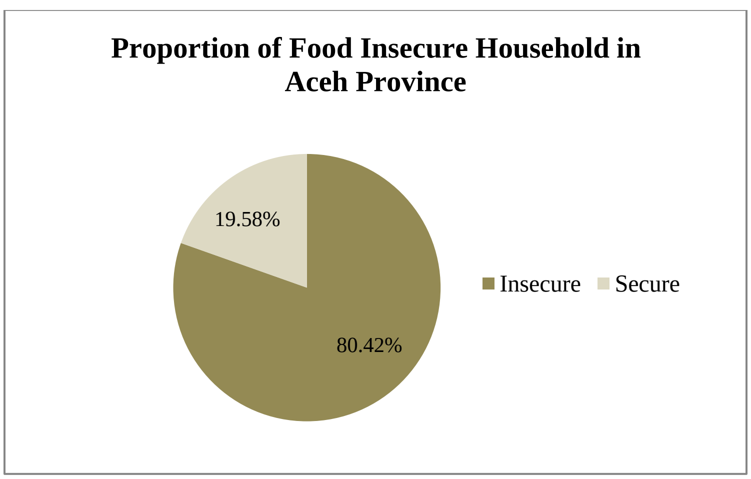

Figure 1. Percentage of food insecurity

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the socio-economy survey conducted by BPS, there are 20% of households in Aceh Province included food-insecure households. Meanwhile, 80% of households are included in food-secure households. It indicates data of food insecurity in Aceh Province 2020 is imbalanced. It will cause some issues for making decision in classification. Thus, before conducting the model, we should balance the data first using random oversampling method.

LITERATURE REVIEW

Classification is one machine learning methods built with the aim of decision making. There are two general steps in conducting classification analysis. Those are training and testing steps. The data is first split into training and testing data. Training step is a series of process for model building by using a set of training data with the known response variable category. While, testing step is used to evaluate the performance of classification model by using testing data.(Ghatak, 2017)

Classification Tree

Classification and Regression Tree (CART) is a non-parametric model that results a decision tree in a form of binary tree. Decision tree obtained is called classification tree when the response variable is categorical data. Otherwise, when the response variable is numerical data, the tree obtained is called regression tree.(L Breiman, 1993) Classification tree is binary recursive partitioning method because it always divides a group of data into two partitions. Each partitioned data is called as a node. The result of data partitioning is represented in a tree structure (L Breiman, 1993). This research (Lewis, 2000) named the origin node (root) by parent node. Parent node can be partitioned to children node. Children node can be partitioned to additional children node. Based on (L Breiman, 1993) classification tree is constructed by iterative selection of $\chi = root(t)$ and all $s \in sons(t)$, $t \in T$.

The algorithm for classification tree in general is divided into these steps as below (L Breiman et al., 1984):

1. Find each variable's best split. For each variable with K different values there exist K-1 possible splits. Find the split, which maximizes the splitting criterion. The resulting set of splits contains best splits (one for each variable).
2. Find the node's best split. Among the best splits from Step 1, find the one, which maximizes the splitting criterion.
3. Split the node using best node split from Step 2 and repeat from Step 1 until stopping criterion is satisfied.

As splitting criterion, we used Gini's impurity index, which is defined for node t as (L Breiman et al., 1984):

$$i(t) = \sum_{ij} C(i|j)p(i|t)p(j|t) \quad (1)$$

where $C(i|j)$ is cost of misclassifying a class j case as a class i case (in our case $C(i|j) = 1$, if $i \neq j$ and $C(i|j) = 0$ if $i=j$). $p(i|t)$ and $p(j|t)$ respectively) is probability of case in class $i(j)$ given that falls into node t.

The Gini impurity criterion is type of decrease of impurity, which is defined as follow:

$$\Delta i(s,t) = i(t) - pLi(t_L) - pRi(t_R) \quad (2)$$

where $\Delta i(s,t)$ is decrease of impurity at node t with split s, pL (pR) are probabilities of sending case to the left (right) child node tL (tR) and I (tL) $(i(tR))$ is Gini impurity measure for left (right) child node. In order to enhance generalization of decision tree we used pruning with combination of cross-validation error rate estimation. The algorithm for pruning works as follows (L Breiman et al., 1984).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

1. Randomly split training data into C folds.
2. Select pruning level of tree (level 0 equals to full decision tree).
3. Use C-1 folds for creation of C-1 new pruned trees and estimate error on last C-*th* fold.
4. Repeat from Step 2 until all pruning levels are used.
5. Find the smallest error and use the pruning level assigned to it.
6. Until pruning level is reached, remove all terminal nodes in the lowest tree level and assign decision class to parent node. Decision value is equal to class with higher number of cases covered by node.

Random Forest

Random Forest is a developed CART algorithm by implementing *bootstrap aggregating (bagging)* process and random feature selection. Random Forest is a classification model which is built from the aggregating of single classification tree. The algorithm of Radom Forest can be explained as below (Leo Breiman, 2001).

Let $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ denote the training data, with $x_1 = \left(x_{i,1}, x_{i,2}, \dots, x_{i,p}\right)^T$

For j = 1 to J:

1. Take bootstrap sample D of size N from D.
2. Using the bootstrap sample. D_j as the training data fit a tree.
   a) Start with all observation in single node.
   b) Repeat the following steps recursively for each node until the stopping criterion is met: (*i*) Select *m* predictor at random from the *p* available predictors.
3. Find the best binary split among all binary splits in the predictor from step 1.
4. Split the node into two descendant nodes using the split from step 2.

Imbalanced Data

Imbalanced data is the data distribution of each class differs substantially where, again, the main class or classes of interest are rare. The class imbalance problem is closely related to cost-sensitive learning, wherein the costs of errors, per class, are not equal (Han et al., 2012). An imbalanced training set may result in a classifier that is biased towards the majority class. When the trained model is applied to a test set that is similarly imbalanced, it would yield an optimistic accuracy estimate (Ghatak, 2017).

There are several ways to handle imbalanced data issue, one of them is *Random Oversampling* (ROS) (Burnaev et al., 2015). Oversampling can change the training data distribution so that the rare class is well represented. Oversampling works by resampling the positive so that the resulting training set contains an equal number of positive and negative (Han et al., 2012).

Receiver Operating Characteristics and Area Under the Curve

Receiver Operating Characteristics (ROC) curve are visual tool that can be used to compare two classification models. ROC curves show the trade-off between the *True Positive Rate (TPR)* and the *False Positive Rate (FPR).* Suppose there is a testing data and a model, TPR is the proportion of positive classes that correctly labelled by the model. While, FPR is the proportion of negative classes that are mislabelled as positive by the model. Given that *TP*, *FP*, *P*, and *N* are respectively the number of true positive, false positive, positive, and negative classes. we know that $TPR = \frac{TP}{P}$, which is sensitivity. Furthermore, $FPR = \frac{FP}{N}$, which is $1 - specificity$. For a two-class problem, an ROC curve allows us to visualize the trade-off between the rate at which the model can accurately recognize positive classes versus the rate at which it mistakenly identifies negative classes as positive for different portions of the test set. Any increase in *TPR* occurs at the cost of an increase in *FPR*. The area under the curve (AUC) is a measure of the accuracy of the model(Han et al., 2012).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The vertical axis of n ROC curve represents TPR. The horizontal axis represents FPR. To plot an ROC curve for a model, it can be started at the bottom left corner, where $TPR = FPR = 0$. The actual class label at the top of the list is then inspected. If there is a true positive (i.e., a positive class that was correctly classified), then TP and thus TPR increase. On the graph, the point is moved up and plotted. Otherwise, the model classifies a negative class as positive, there is a false positive, and so both FP and FPR increase. On the graph, the point is moved right and plotted. This process is repeated for each of the test classes in ranked order, each time moving up on the graph for a true positive or toward the right for a false positive(Han et al., 2012).

Variable Importance Score

A measurement that be used to determine the variable importance in classification analysis is *Mean Decrease Gini (MDG).* MDG is a ratio between the sum of reduced which is caused by the s-*th* predictor variables and the number of trees that formed. The enhancement of MDG means that the predictor variable has a great contribution for building a classification tree. MDG is formulated as below (Zani et al., 2005):

$$MDG_s = \frac{1}{k}\sum\nolimits_t [\Delta(s,t)I(s,t)] \tag{3}$$

With k is the number of trees that formed, $\Delta(s,t)$ is reduced impurity which is caused by the s-*th* predictor variables in the t-*th* node. $(s,)$ is the indicator function which is 1 if the s-*th* predictor variable is used in partitioning and otherwise is 0.

METHOD

Data

This study uses secondary data, namely the Aceh province SUSENAS data in 2020 which is sourced from BPS Aceh. The data observations used is 12,971 households consists of output indicators, namely the status of food insecurity and input data which are indicators of food insecurity. The variables used can be seen in the Table 1 as follows

Table 1. Variables Explanation

No	Type of Variable	Definition
1	Food Insecurity Status (Ballard & Cafiero, 2013)	A scale that can describe the inability of households or individuals to access the food they need on a regular basis. It's measured using FIES (Food Insecurity Experience Scale) with 2 categories; YES and NO.
Social Assistance Program Receiver Status		
2	BPJS (Wardani, 2018)	Social Health Insurance
3	BPNT (Wardani, 2018)	Food aid
4	PKH (Rahmansyah et al., 2020)	Government assistance
5	KKS (Sunarti, 2006)	Identity of non-welfare family
6	KIP (Pusat Penelitian Kebijakan Pendidikan dan Kebudayaan, 2017)	Identity of education assistance receiver
7	PIP (Pusat Penelitian Kebijakan Pendidikan dan Kebudayaan, 2017)	Education assistance receiver
8	Jamkesda (Wardani, 2018)	Health insurance provided by local government
9	Local Government Assistance (*PETUNJUK TEKNIS:*	Local Government Assistance

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	PENGENTASAN DAERAH RENTAN RAWAN PANGAN/PERTANIAN KELUARGA TAHUN 2021, 2021)	
Supporting Activities of Household		
10	Cooking Fuel (Lewis, 2000)	Type of cooking fuel used by the household
11	Sanitation Eligibility (Cecilia, 2015)	The eligibility of sanitation of the household
12	Source of Drinking Water (Cecilia, 2015)	The source of drinking water consumed by household member
13	Internet access (Pujilestari & Haryanto, 2020)	Internet accessibility
14	Drinking Water Eligibility (WFP & Pangan, 2015)	Drinking Water Eligibility
15	Electricity (Sulemana et al., 2019)	Electricity accessibility
Head/member of Household Characteristics		
16	Head of Household Education Level (Irawan et al., 2019)	The level of Education of Head of Households
17	Illiteracy Number (Hapsari & Rudiarto, 2017)	Illiteracy number in the household
18	Vulnerability of Household's Head (Wardani, 2018)	Head of household is woman with the member of household 0-14 years old
19	Non-outpatient with sick status (BPS, 2006); (Damayanti et al., 2018)	Non-outpatient with sick status
Socio/economy of Household		
20	Number of Saver (Irawan et al., 2019)	The number of household member that have saving
21	Land Asset (Million & Muche, 2020)	Land asset status
22	Transfer Receiver (Magaña-Lemus et al., 2016)	Transfer receiver status
Physical Condition of Household		
23	Type of Floor (Umniyyah, 2018)	Type of floor of household
24	Floor Area (Damayanti et al., 2018); (BPS, 2007)	Area of house, discretization with CHI-MERGE
25	Type of Wall (Umniyyah, 2018)	Type of wall of the house
26	Type of Roof (Umniyyah, 2018)	Type of roof of household

Research Procedure

Research procedure to conduct this study is shown below:

1. Conducting data exploration in descriptive form. It aims to see the characteristics of the data used.
2. Split the data into training data and testing data. 70% training data to form the model and 30% testing data to check the accuracy of the formed model.
3. Perform data resampling. Data resampling aims to overcome training data and full data which have different amounts of data between classes. Data resampling was done using random oversampling method.
4. Creating Classification Tree

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

a. Tuning hyper-parameters
   In the Classification Tree, the hyper-parameters used are minsplit and cp. The minsplit used starts from 50 to 1000, while the Cp value used is 0.
b. Seeing the most optimum model
   The most optimum model is seen from the resulting AUC value. The higher the AUC value, the better the model formed.

5. Creating Random Forest
   a. Tunning hyper-parameter
      In the Random Forest, the hyper-parameters used are mtry and ntree. The mtry used starts from 500 to 2000, the Ntree used starts from 2 to 5.
   b. Seeing the most optimum model
      The most optimum model is seen from the resulting AUC value. The higher the AUC value, the better the model formed.
6. Comparing Classification Tree and random forest methods. The comparison between Classification Tree and random forest method is seen based on the resulting AUC value. The method with the highest AUC value is the best method.
7. Get the importance variable score. The importance variable score can show how big the contribution of the predictor variables in explaining the criteria for food insecurity in a household.
8.

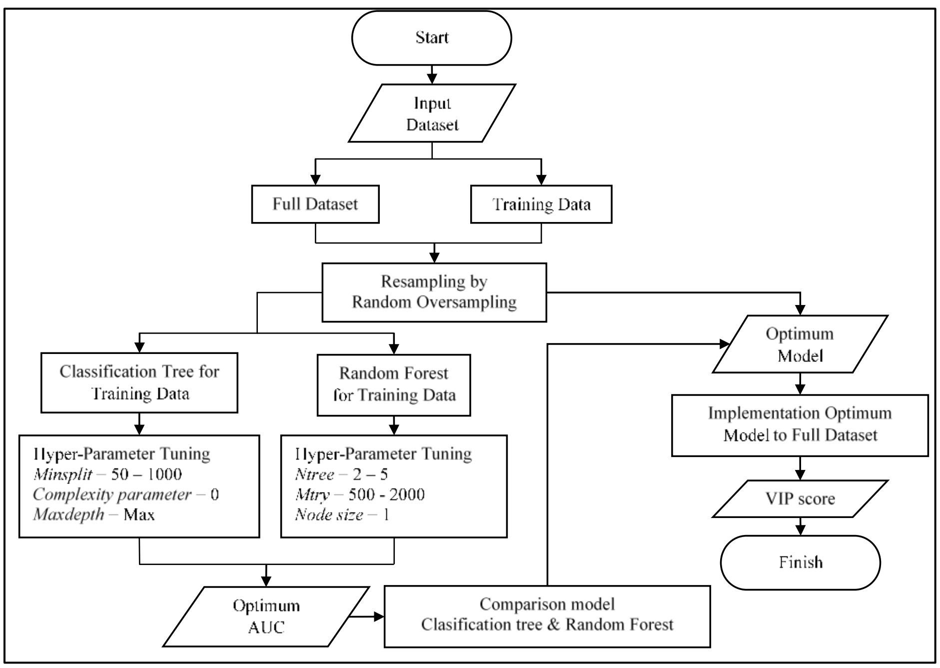

Figure 1. Flowchart of Random Forest Aceh Province 2020

RESULT

Data Overview

To conduct classification analysis, the dataset is split into training and testing data. In this study, the training data used are 70% of all surveyed households. Training data is used to build the model and determine the optimum model. While, the testing data is used to extract the information from the model. In 2020, the socio-economy survey was conducted in 12.971 households in Aceh Province. Based on the survey, it's obtained there are 2.540 households with food insecurity and the other 10.431 are secure. From the splitting data, training data obtained consists of 1.778 households with food insecurity and other

*name of corresponding author

7.303 households are secure. This shows the unstable proportion between response variable categories which indicate a problem called imbalanced data. The issue of imbalanced data can lead to less accurate model obtained(Ghatak, 2017). Therefore, this issue should be addressed first. The Table. 2 below shows the food insecurity overview of original data and the *random oversampling* data.

Table 2. Food insecurity overview

Data/Status	Original		Resample	
	Insecure	Secure	Insecure	Secure
Full dataset	2540	10431	10431	10431
Training	1778	7302	7302	7302

One method that can be used to deal with imbalanced data is by applying *resampling* using *random oversampling. Random oversampling* is implemented in both training data and full dataset. It's aimed to result a balance data which have stable response variable categories proportion(Ghatak, 2017). For training data, it's obtained 7.302 observations for each category of response variable; food insecure and secure households. While for full data, it's obtained 10.431 observations for each category of response variable.

DISCUSSIONS

Comparison Classification Tree and Random Forest

Both in classification and random forest method, hyper-parameter tuning must be done first in order to obtain the fit model of the algorithm. In classification tree, the hyper parameter that was tuned is *minsplit* which is a multiple of 50 started from 50 to 400 and continued with a multiple of 100 started from 400 to 1000. Parameter Cp is set to 0. This is done with the aim of obtaining classification tree results that are easy to interpret. In this research, goodness of model is measured by AUC. The greater AUC obtained indicates the better model. Hyper-parameter tuning for classification tree is shown in Table 3 below.

Table 3. Hyper-Parameter Tuning in Classification Tree

Minsplit	AUC Aceh 2020
50	0.647365
100	0.6699939
200	**0.6778106**
250	0.6757721
300	0.6758495
350	0.6762406
400	0.6748726
500	0.6779809
600	0.6766103
700	0.6759526
800	0.6758941
900	0.6757161
1000	0.6757161

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3 shows that AUC value for every *minsplit* changed started from 50-1000 tend to be constant. However, the highest AUC among those is found when *minsplit* is 200, which is 0,678. Thus, classification tree at a *minsplit* of 200 is determined as the optimum classification tree model.

The hyper parameter in random forest method that tuned are *ntree* and *mtry*. The *ntree* used are 500, 750, 1000, 1250, 1500, 1650, 1750, and 2000. While, *mtry* used are 2, 3, 4, and 5. The AUC values obtained are various as is shown in Table 4 below

Table 4. Hyper-Parameter Tuning in Random Forest

Ntree	*Mtry*			
	2	3	4	5
500	0.7165	0.7161	0.7081	0.7017
750	0.7173	0.7148	0.7102	0.7019
1000	**0.7184**	0.7154	0.7075	0.7021
1250	0.7154	0.7156	0.7090	0.7014
1500	0.7161	0.7163	0.7087	0.7014
1650	0.7160	0.7161	0.7100	0.7013
1750	0.7161	0.7156	0.7094	0.7005
2000	0.7154	0.7160	0.7093	0.7010

As the AUC values in classification tree, hyper parameter tuning parameter in random forest with various *ntree* and *mtry* also tend to be constant. Table 4 indicates the optimum AUC value is 0,718 at the *ntree* of 1000 and *mtry* is 2. This indicates that the random forest method is optimum when *ntree* is 1000 and *mtry* is 2.

Based on the comparison of both methods, the optimum AUC value is obtained from random forest method with the AUC value is 0,718. This indicates that random forest can improve the level of accuracy more than classification tree does.

4.1. VIP Score of Optimum Method

Variable Importance (VI) score is a value that indicates how much the contribution of predictor variables towards the response variable. Hence, based on the VI score obtained, it can be identified how capable the predictor variables are in explaining the criteria of food insecurity case. VI score used in this study is the VI score obtained from the optimum method, which is random forest that implemented in the over-sampling full dataset. It's obtained the VI score for food insecurity case in Aceh Province in 2020. The VI score can be seen in Figure 1 below.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

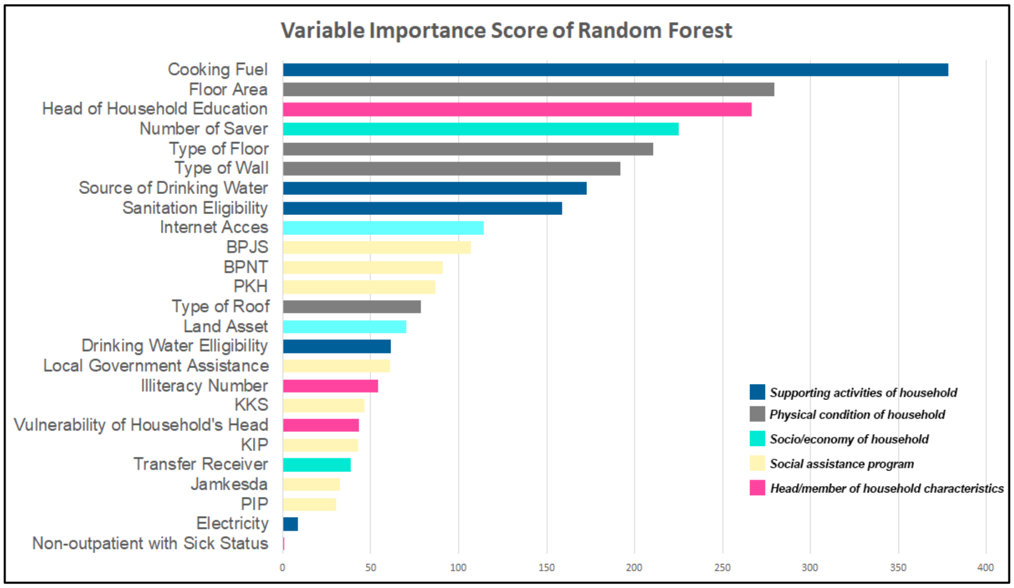

Figure 2. Variable Importance Score Food Insecurity Aceh 2020

Figure 1 shows VI score of food insecurity case in Aceh Province 2020. It seems the majority at the top 10 indicators of food insecurity are from supporting activities of household and physical condition of household. Supporting activities of household category is consists of cooking fuel, head of household education, source of drinking water, and sanitation eligibility. Meanwhile, the physical condition of household which consists of floor area, type of floor, and type of wall. Therefore, those variables become determinant of food insecurity case in Aceh Province 2020.

CONCLUSION

This research purposed to compare both classification method which are classification tree and random forest. Besides, this research also purposed to identify indicators of food insecurity in Aceh Province 2020. It is obtained that the optimum method based on the optimum value of AUC is Random Forest, which is 0.7184 with *ntree* is 1000 and *mtry* is 2 for training data. The optimum model obtained is then implemented to full dataset for identifying indicators of food insecurity Aceh Province 2020 based on the rank of VI score. It seems the majority at the top 10 indicators of food insecurity are from supporting activities of household and physical condition of household. Supporting activities of household category is consists of cooking fuel, head of household education, source of drinking water, and sanitation eligibility. Meanwhile, the physical condition of household which consists of floor area, type of floor, and type of wall. Therefore, those variables become determinant of food insecurity case in Aceh Province 2020.

ACKNOWLEDGMENTS

Authors would like to express gratitude to the BPS Aceh for data support and Ministry of Research and Technology/National Agency of Research and Innovation - Republic of Indonesia (BRIN) for funding and supporting this research. The content is solely the responsibility of the authors and does not represent the official views of the National Agency of Research and Innovation.

REFERENCES

Ballard, T. J., & Cafiero, A. W. K. and C. (2013). *The Food Insecurity Experience Scale: Developing s Global Standard for Monitoring Hunger Worldwide*. *October*, 1–58.

BPS. (2006). *ingkat Kemiskinan di Indonesia.*

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11852
e-ISSN : 2541-2019
p-ISSN : 2541-044X

BPS. (2007). *Analisis dan Penghitungan Tingkat Kemiskinan Tahun 2007*.

Breiman, L. (1993). *Classification and Regression Tree*. Chapman and Hall.

Breiman, L, Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

Breiman, Leo. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of Resampling on Accuracy of Imbalanced Classification. *Eight International Conference on Machine Vision*.

Cecilia, A. (2015). Population food security assessment: A methodological approach. Agricultural Economics and Rural Development—Realities and perspectives for Romania. *6th Edition of the International Symposium, November 2015, Bucharest, The Research Institute for Agricultural Economy and Rural Development (ICEADR)*, 2–8.

Damayanti, H. O., Perencanaan, B., Daerah, P., & Pati, K. (2018). *Jurnal Litbang Vol. XIV, No.1, Juni 2018: 15-26*. *XIV*(1), 15–26.

Ethem, A. (2020). *Introduction to Machine Learning* (fourth edi). MIT Press.

Ghatak, A. (2017). Machine Learning with R. In *Machine Learning with R*. https://doi.org/10.1007/978-981-10-6808-9

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013* (Third Edit). Morgan Kaufmann. https://doi.org/10.1109/ICMIRA.2013.45

Hapsari, N. I., & Rudiarto, I. (2017). Faktor-Faktor yang Mempengaruhi Kerawanan dan Ketahanan Pangan dan Implikasi Kebijakannya di Kabupaten Rembang. *Jurnal Wilayah dan Lingkungan*, *5*(2), 125. https://doi.org/10.14710/jwl.5.2.125-140

Irawan, H., Sartono, B., & Erfiani. (2019). *Faktor-faktor Rumah Tangga Yang Mencirikan Tingkat Kerawanan Pangan*.

Lewis, J. . (2000). *An introduction to Classification and Regression Analysis (CART)*.

Magaña-Lemus, D., Ishdorj, A., Rosson, C. P., & Lara-Álvarez, J. (2016). Determinants of Household Food Insecurity in Mexico. *Agricultural and Food Economics*, *4*(1). https://doi.org/10.1186/s40100-016-0054-9

Million, M., & Muche, M. (2020). *FACTORS DETERMING THE FOOD INSECURITY STATUS OF RURAL HOUSEHOLD IN MANNA WOREDA OF JIMMA ZONE , ETHIOPIA . FACTORES QUE DETERMINAN EL ESTADO DE INSEGURIDAD ALIMENTARIA DEL HOGAR RURAL EN LA ZONA MANA WOREDA*. *8*(X), 1–15.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

PETUNJUK TEKNIS: PENGENTASAN DAERAH RENTAN RAWAN PANGAN/PERTANIAN KELUARGA TAHUN 2021. (2021). BADAN KETAHANAN PANGAN KEMENTERIAN PERTANIAN.

Pujilestari, T., & Haryanto, T. (2020). Peran Perempuan dalam Meningkatkan Ketahanan Pangan Rumah Tangga di Provinsi Nusa Tenggara Barat Triana. *Media Trend*, *15*(2), 319–332.

Pusat Penelitian Kebijakan Pendidikan dan Kebudayaan. (2017). *Kajian Implementasi Program Indonesia Pintar*. 17.

Rahmansyah, Z., Senjawati, N. Z., & Juarini. (2020). *ANALISIS KETAHANAN PANGAN RUMAH TANGGAMISKIN BERDASARKAN PANGSA PENGELUARAN PANGAN DANKONSUMSI ENERGI DI DESA GIRIREJO KECAMATAN IMOGIRIKABUPATEN BANTUL*. *21*(Juni), 68–78.

Sulemana, I., Anarfo, E. B., & Quartey, P. (2019). International remittances and household food security in Sub-Saharan Africa. *Migration and Development*, *8*(2), 264–280. https://doi.org/10.1080/21632324.2018.1560926

Sunarti, E. (2006). *No Title*. Institut Pertanian Bogor.

*name of corresponding author

Umniyyah, T. (2018). *Analisis Faktor-faktor yang Mempengaruhi Ketahanan Pangan Rumah Tangga Penderita Tuberkulosis di Wilayah Non Pesisir Surabaya*. Institut Teknologi Sepuluh Nopember.

Wardani, A. S. (2018). Determinan Ketahanan Pangan Dan Gizi Rumah Tangga Petani Indonesia Di Kawasan Pedesaan. *Jurnal Ekonomi*, *January*, 153. https://doi.org/10.13140/RG.2.2.14278.98881

WFP, & Pangan, D. K. (2015). *Peta Ketahanan dan Kerentanan Pangan Indonesia 2015: Versi Rangkuman*.

Zani, S., Marco, R., & Maurizio, V. (2005). *Data Analysis, Classification and The Forward Search: Proceedings of The Meeting of The Classification and Data Analysis Group (CLADAG) of The Italian Statistical Society* (eds (Ed.)). Springer Science & Business Media.

*name of corresponding author