The 3rd International Seminar of Science and Technology
ISST 2023 Vol 3 (2024) 03 002
Trends in Science and Technology for Sustainable Living
Faculty of Science and Technology
Universitas Terbuka

Classification of the human development index in Kalimantan using random forest method

Siti Mahmuda*, Darnah Andi Nohe, and Aji Muhammad Leonardo

Mulawarman University, Statistics Department, Samarinda, East Kalimantan, Indonesia, 75123

Abstract - Random Forest is an ensemble classifier in a machine-learning algorithm. The ensemble classifier aimed to improve model accuracy and classification performance. Based on accuracy measures, Random Forest shows the best performance with existing ensemble classifiers like Support Vector Machine (SVM) and AdaBoost. Hence, this research will classify Human Development Index in Kalimantan using a Random Forest classifier. The predictor variables of classification are the average length of schooling, adjusted per capita additions, life expectancy, and length of school expectations. The Random Forest showed that the number of trees selected was 500 and the *m* being tried was 2. Adjusted per capita additions were the most influential variable in the increase of The Human Development Index with an importance of 10,61%. The Accuracy of classification was 58,33%.

Keywords: accuracy,ensemble classifier, human development index, Random forest

1 Introduction

Research on the Human Development Index (HDI) has been done using several methods, including Smooth Support Vector Machine [1], Decision Tree C4.5 [2], and many more. HDI data is category data, so classify using the classification method. One classification method that uses decision tree techniques is Classification And Regression Trees (CART). CART calculates faster and can solve cases that have large amounts of data. This method still has weaknesses, namely producing less stable trees where small changes in the learning data can cause significant changes to the trees formed [3] and tend to be overfitting [4]. So, to increase stability and avoid overfitting, the ensemble method is applied [3].

The ensemble classifier is a method that combines several classification algorithms to increase model strength and improve classification performance [5]. Ensemble classifiers are considered more effective than single classifiers due to the following reasons: (1) the training set does not always provide sufficient information to choose an accurate hypothesis, (2) the learning process of a classifier is weak under certain conditions, (3) the hypothesis space searched may not contain the actual target function while an ensemble classifier may provide a good approximation [6]. Ensemble classifiers are considered more resistant to noise and able to minimize bias and variance than single-learning.

Several methods apply ensemble classifiers, including Random Forest, Support Vector Machine (SVM), and AdaBoost. The average accuracy produced by the three algorithms is quite good and has

*Corresponding author: sitimahmuda@fmipa.unmul.ac.id

almost equal performance in classification, but the Random Forest method has the highest metric value compared to other methods [7]. Another research shows that Random Forest has the best conduct for all accuracy criteria [8]. This research compares the performance of several ensemble classifier methods, including Bagging, Random Forest, AdaBoost, XGBoost, and LightGBM.

Based on the discussion above, this research aims to classify the Human Development Index (HDI) in Kalimantan using the random forest method. The result of this research could be able to provide information about HDI classification, the level of classification accuracy, and the most influential variable in HDI.

2 Methodology

2.1 Random Forest

Breiman introduced The Random Forest method in 2001. Random Forest has two functions for solving problems: classification and prediction. The basic technique used by Random Forest is a decision tree. In other words, the Random Forest is a collection of decision trees for the classification and prediction of data by entering input into the root at the top and then down to the leaves at the bottom [9]. The results of Random Forest analysis for classification are the mode of each tree from the built forest, while the prediction results are obtained from the average value of each tree [10].

Random Forest is a result of the development of Classification and Regression Tree (CART) method, which applies the bagging or bootstrap aggregating method and random feature selection. Bagging is a method that can improve the results of classification algorithms. The basis of this bagging method is the ensemble method [11]. According to [12], The Random Forest algorithm:

1 Take *n* sample data from the initial dataset using the bootstrap resampling technique with replacement.
2 Compile a classification tree from each dataset resulting from bootstrap resampling, determining the best classifier based on attribute variables taken at random. The count of variables taken at random could be determined by calculating $\frac{1}{2}\sqrt{m}$ or $\sqrt{m}$ or $2\sqrt{m}$, where *m* is the number of predictor variables.
3 Predict the classification of sample data based on the classification tree formed.
4 Repeat steps 1-3 until the desired number of classification trees obtained. Then, do the repetition for *K* times.
5 Predict the classification of the final sample data by combining the results of the classification tree predictions obtained.

The Random Forest algorithm has an *m* value that could be different. The *m* value is the number of predictor variables used as separators in tree formation. A higher *m* value will cause a higher correlation.

2.2 Information Gain

The calculation used when building a decision tree using the CART method is gain information. The gain information describes the size of the selection of variables used for classification by each node in a tree. For example, N is a node to separate each class based on variables from the dataset denoted *D*. Node separation is done based on the highest gain information of the variable. The formula for getting gain information is as follows:

$$\text{Information Gain (IG)} = \text{entropy(D)} - \text{entropy}_A\text{(D)} \quad (1)$$

The *entropy*(D) value can be obtained by using fomula (2) and formula (3) to get the entropy_A(D) value:

$$\text{Entropy(D)} = -\sum_{i=1}^{c} p_i^2 \log_{f0}(p_i) \quad (2)$$

c : the number of target class
p_i : the proportion of class i based on partition of D

$$\text{Entropy}_A\text{(D)} = \sum_{j=1}^{v} \frac{n_j}{n} \times \text{entptopy}(D_j) \quad (3)$$

v : the number of partition
n : the number of observation
n_j : the number of observation j
D_j : the partition j

The partition process in each iteration in the classification tree algorithm is basically looking for a partition method that provides the highest information gain [13].

2.3 Accuracy

If a classification model has been produced based on the dataset, it is necessary to know how well the prediction performance of the classification model is. The most intuitive way to see the goodness of the prediction results from a model is to look at the accuracy of the predictions, namely comparing the class of predicted results with the actual [13].

According to Carlo Vercellis in research [14], there are two reasons for measuring the accuracy value of a prediction model. First, the accuracy value is needed to determine the prediction model with the most accurate results. Second, the accuracy value can detect deficiencies in the model formed. The prediction model produces performance based on analysis using several methods that can be measured based on the errors from the prediction results. The error value can be obtained using the Mean Absolute Percentage Error (MAPE) calculation. MAPE is the average percentage of all differences between actual data and data from predicted results [10]. The MAPE formula is:

$$MAPE = \frac{1}{n} \sum_{I=1}^{n} \frac{|X_i - F_i|}{X_i} \times 100 \quad (4)$$

X_i : the actual value i
F_i : the prediction value i
n : the number of observation (data)

Based on the MAPE formula (4), which describes the error value of the prediction model, the accuracy of the model can be obtained using the following calculation for:

$$Accuracy = 100\% - MAPE \quad (5)$$

2.4 Variable Importance

Analysis of the rate of variable importance in modeling can be found in various terms, such as variable importance analysis and dominance. In predictive modeling, important predictor variables are variables when not included in the model, the quality of the prediction results will decrease significantly.

Based on the definition above, intuitively, it is used to measure the importance of a predictor variable, which is the comparison between the quality of model predictions involving that variable and the model predictions not involving that variable.

Assuming there are p predictor variables and l is one of the variables in it, then it can be formulated:

$$VI_l = \text{accuracy}(M_w) - \text{accuracy}(M_s) \quad (6)$$

Note:
VI_l : decrease in the rate of model accuracy without variable l
M_w : the model without variable l
M_s : the model with variable l

Variable l will be necessary when the variable is removed from the model, so there is a high decrease in accuracy, and vice versa.

The calculation process of VI_l generally depends on the accuracy of the testing dataset. Because the testing dataset is taken randomly from the complete dataset, if the VI_l process is calculated many times, different values may be obtained due to the random processing. Encourages the calculation of VI_l depends on the average difference in accuracy from several repetitions. The statistic obtained is known as the average accuracy drop.

Apart from using changes in accuracy values, in classification tree-based models, the importance measure can also be based on changes in gini values. The Gini values from the nodes of the M_w model compare to the Gini values from the final covers of the M_s model. This measure is known as the mean decrease gini [13].

2.5 Variables and Data Structure

The dataset in this research is Human Development Index (HDI) data and its attribute variables from 56 districts/cities in Kalimantan. The dataset is secondary data obtained from Badan Pusat Statistik (BPS) in 2022. There are five variables available in the dataset presented in Table 1.

Table 1. Variables of Dataset

No	Variable	Description
1	Human Development Index (HDI)	Achievement of human development based on a number of basic components of quality of life.
2	Average length of schooling (ALS)	The average length of time used by residents in undergoing formal education.
3	The adjusted per capita expenditure figures (ACF)	The purchasing power of the community and better describes the income of the people during a certain period.
4	Life expectancy (LE)	Estimated average length of time a person has lived in years.
5	Expected length of school (ELS)	Estimate of the length of schooling (in years)

The response variable is data in category form, namely the Human Development Index (HDI). The HDI contains four categories, namely low, moderate, high, and very high. The predictor variables

are average length of schooling, the adjusted per capita expenditure figures, life expectancy, and expected length of school.

3 Results and discussion

3.1 Statistic Descriptive

HDI categories are the classes in this classification analysis. The HDI classification is in a low class if the value is < 60, moderate class if the value is 60 ≤ HDI < 70, high class if the value is 70 ≤ HDI < 80, and very high class if the value is ≥ 80. Figure 1 shows a comparison of the proportion of HDI classes in Kalimantan in 2022.

Figure 1 shows the proportion of HDI in Kalimantan in 2022. The very high class has a proportion value of 9%, the high class has a proportion value of 45%, the moderate has a proportion value of 46%, and the low class has a proportion value of 0%. The low class is not included in the HDI class, because it does not have observation.

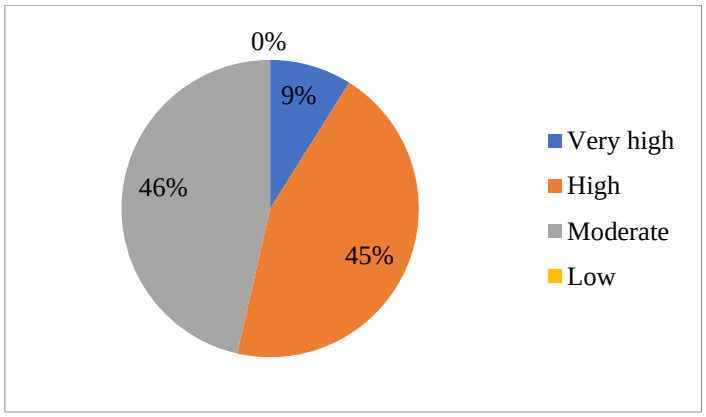

Fig. 1. Proportion of HDI in Kalimantan

3.2 Testing and Training Data

Before using the Random Forest method, the dataset is divided into two, namely (1) training data and (2) testing data. The training data is data for training an algorithm to form a model, but the testing data is data for evaluating the performance or accuracy of the model.

In this research, the proportion of training data and testing data is 80% and 20%. So, the amount of training data is 45 observations, and the amount of testing data is 11 observations. The Random Forest algorithm takes training and testing data randomly from the dataset.

3.3 Random Forest Algorithm

3.3.1 Selecting mtry value and number of trees based on misclassification

The Random Forest algorithm uses mtry values to determine the number of randomly selected predictor variables. Using the mtry as a separator candidate in a tree formation. Determination of the mtry value used in this research is [12]:

$$\text{mtry}_1 = \frac{\sqrt{\text{number of predictor variables}}}{2} = \frac{\sqrt{4}}{2} = 1$$
$$\text{mtry}_2 = \sqrt{\text{number of predictor variables}} = \sqrt{4} = 2$$
$$\text{mtry}_3 = \sqrt{\text{number of predictor variables}} \times 2 = \sqrt{4} \times 2 = 4$$

The calculation above shows the number of predictor variables that will be evaluated to determine the number of variables used in partitioning the classification tree nodes.

Table 2. Misclassification each *mtry*

Mtry	Misclassification
1	6,84%
2	6,84%
4	10,26%

The selection of mtry values is chosen by looking at the average size of misclassification from the Random Forest model. The smaller the misclassification value, the more optimal the number of mtry used will be. Table 2 shows that the optimum mtry is 1 or 2 because it has the smallest misclassification value. Therefore, the classification in this study uses an entry value of 2.

Table 3. Misclassification each number of trees

Number of trees	Misclassification
100	7,70%
500	7,69%
1000	8,55%

The accuracy of Random Forest classification is also determined by the number of trees. Breiman (2001) stated that the Random Forest misclassification rate will converge towards a certain value when the trees of Random Forest size become larger [15]. This research uses several trees of 100, 500, and 1000. Because classification analysis using this number of trees is expected to reduce misclassification [16].

Based on Table 3, the Random Forest classification sets the number of trees at 500, because it has the smallest misclassification value, namely 7.69%. It means that the average misclassification with several trees built of 500 is the minimum error rate.

3.3.2 Random Forest Prediction

The Random Forest classification model with mtry value of 2 and the number of trees of 500 obtained the prediction results from the testing data in Table 4. Table 4 shows the predictions for class 12 testing data. The correct prediction results for the moderate class are 7 observations and not any incorrect predictions. In the high class, there are 4 incorrect predictions. Meanwhile, predictions that are incorrect in the very high class are 1 observation. The conclusion from the prediction results of 12 data testing observations is that 5 observations are misclassification, and 7 observations are predicted correctly.

Table 4. Prediction of Testing Data

Prediction	Testing data		
	Very high	High	Moderate
Very high	0	0	1
High	0	0	4
Moderate	0	0	7

3.4 Accuracy of Classification

The accuracy of the model classification Random Forest can be obtained using formula (4) and (5):

$$\text{MAPE} = \frac{1}{n}\sum_{I=1}^{n}\frac{|X_i - F_i|}{X_i} \times 100$$

$$\text{MAPE} = \frac{1+4+0}{12} \times 100 = 41{,}67$$

$$\text{Accuracy} = 100\% - \text{MAPE} = 100\% - 41{,}67 = 58{,}33\%$$

Based on the calculations above, it was found that the accuracy of the testing data was 58.33%. The accuracy value shows that the Random Forest classification model can classify the Human Development Index in Kalimantan correctly at 58.33%.

3.5 Features Importance

Figure 2 and Table 5 show feature importance is useful for describing data and understanding variables that are more important in building models and determining predictions. Thus, in this research, it is known that the adjusted per capita expenditure figure is the variable with the high importance which can influence the determination of used camera price predictions with a value of 10.61. Then the level of importance is followed by the average length of schooling variable of 5.73, the expected length of school variable of 4.69, and the life expectancy variable of 4.44 which are also variables with a low level of importance. Variables with a high level of importance are a driving force in the accuracy of the model formed because the prediction results will be close to the actual value.

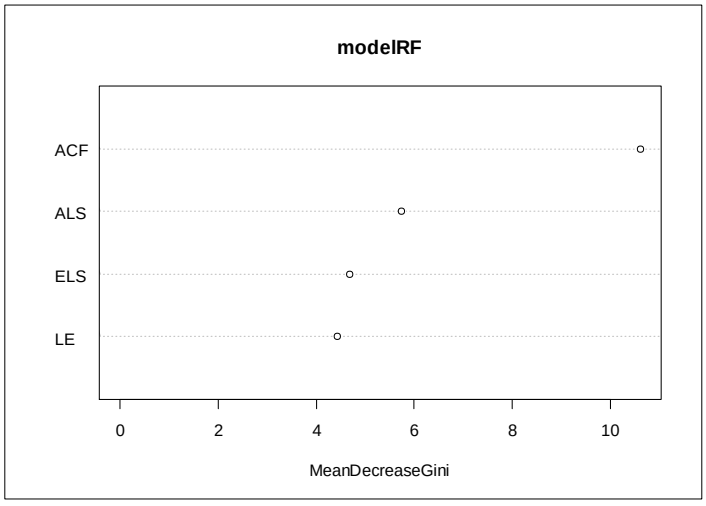

Fig. 2. Mean decrease gini of predictor variables.

Table 5. Mean decrease gini of predictor variables

Predictor variables	Mean decrease gini
Average length of schooling (ALS)	5,73
The adjusted per capita expenditure figures (ACF)	10,61
Life expectancy (LE)	4,44
Expected length of school (ELS)	4,69

4 Conclusion

The results of the analysis from this research obtained several conclusions:

1 Descriptive statistics show that the Human Development Index (HDI) in Kalimantan in 2022 has three classes, namely the very high class at 9%, the high class at 45%, and the moderate class at 46%.
2 Classification using Random Forest with a value of mtry = 2 and the number of trees = 500 produces an accuracy rate of 58.33%, which means that the accuracy of the Random Forest classification method for this data is very high. This Random Forest method test uses 80% training and 20% testing data.
3 Importance The variable that has the most influence in increasing the Human Development Index by Regency/City in Kalimantan in 2022 based on the mean decrease Gini value is the adjusted per capita expenditure figures (ACF) variable.

References

[1] F. Fatkhurohman, "Smooth Support Vector Machine (SSVM) untuk Pengklasifikasian Indeks Pembangunan Manusia se-Indonesia, [Smooth Support Vector Machine (SSVM) for Human Development Index Classification in Indonesia]," Indonesian Journal of Mathematics and Natural Sciences, vol. 40(2), 2017.

[2] Putra, R. Miftahul, A.Elvira "Prediksi Indeks Pembangunan Manusia Menggunakan Decision Tree C4.5 di Kabupaten Kampar, [Prediction of Human Development Index Using Decision Tree C4.5 in Kampar Regency]", Journal of Chemical Information and Modeling, vol. 53(9), 2013.

[3] S. Dian, et al., "Klasifikasi Rumah tangga Sangat Miskin di Kabupaten Jombang Menurut Paket Bantuan Rumah Tangga yang Diharapkan dengan Pendekatan Random Forest Classification and Regression Trees (RF-CART), [Classification of Very Poor Households in Jombang Regency According to Expected Household Assistance Packages Using the Random Forest Classification and Regression Trees (RF-CART) Approach]", Kementerian Riset, Teknologi, dan Pendidikan Tinggi,pp 1-7, 2014.

[4] M. Dhawangkara and E. Riksakomora, "Prediksi Intensitas Hujan Kota Surabaya dengan Matlab Menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya), [Prediction of Rain Intensity in Surabaya City with Matlab Using Random Forest and CART Techniques (Case Study of Surabaya City)]", Jurnal Teknik ITS, vol.6(1), 2014.

[5] Syarif *et al*, "Application of bagging, Boosting, and Stacking to Intrusion Detection", Springer-Verlag Berlin, vol.7376, pp.593-602, 2012.

[6] T.G. Diettarich, "Ensemble Method in Machine Learning", Springer-Verlag Berlin, vol.1857, pp.1-15, 2000.

[7] R.I. Arumnisa and A.W. Wijayanto, "Perbandingan Metode Ensemble Learning: Random Forest, SVM, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM), [Comparison of Ensemble Learning Methods: Random Forest, SVM, AdaBoost in Human Development Index (HDI) Classification]", Jurnal Sistem Informasi , vol.12(1), 2023.

[8] Y. Li and C.Weidong, "A Comparative Perfomance Assesment of Ensemble Learning for Credit Scoring", Mathematics, vol.8(10), 2020.

[9] M.J. Wulansari, "Analisis Faktor-faktor yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest, [Analysis of Factors that Influence Someone to Get Diabetes Mellitus Using Random Forest Regression]", Thesis. Yogyakarta. 2018.

[10] P.R.D. Lingga, "Deteksi Gempa Berdasarkan Data Twitter menggunakan Decision Tree, Random Forest, dan SVM, [Earthquake Detection Based on Twitter Data using Decision Tree, Random Forest, and SVM]", Jurnal Teknik ITS, vol.160, 2011.

[11] A.Y. Samudra, "Pendekatan Random Forest untuk Model Peramalan Harga Tembakau Rajangan di Kabupaten Temanggung, [Random Forest Approach for Forecasting Models of Rajangan Tobacco Prices in Temanggung Regency]", thesis. Yogyakarta. 2019.

[12] Rahmawati, "Klasifikasi Faktor-faktor yang Mempengaruhi Korban Kecelakaan Lalu Lintas di Surabaya dengan Pendekatan Regresi Logistik Multinomial dan Random Forest, [Classification of Factors that Influence Traffic Accident Victims in Surabaya using Multinomial Logistic Regression and Random Forest Approaches]", thesis. Surabaya. 2015.

[13] B. Sartono and H. Dharmawan, "Pemodelan Prediksi Berbasis Pohon Klasifikasi, [Classification Tree Based Prediction Modeling]", IPB Press Bogor, 2023.

[14] B. Lareno, "Analisa dan Perbandingan Akurasi Model Prediksi Rentet Waktu Arus Lintas Jangka Pendek, [Analysis and Comparison of the Accuracy of Short-Term Cross-Flow Time Series Prediction Models]", CSRID Journal, vol. 6(3), 2014.

[15] L. Breiman, "Random Forest" Machine Learning Journal, vol. 45(5), 2001.

[16] N. K. Dewi, U. D. Syafitri, and S. Y. Mulyadi, "Penerapan Metode Random Forest dalam Driver Analysis, [Application of the Random Forest Method in Driver Analysis]", Statistics and Its Application, vol.16(1), 2011.