Jurnal Mantik, 6 (1) (2022) 351-356

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Analysis Using Random Tree And Random Forest With The Gini Index Algorithm On Data

Nuranisah

Science and Technology Universitas Pembangunan Panca Budi Medan

E-mail: nuranisahasriel123@gmail.com

ARTICLE INFO

Article history:
Received: 01 April 2022
Revised: 2 May 2022
Accepted: 14 May 2022

Keywords:
Analysis, Random, RFW, GIA

ABSTRACT

PPDB (Penerimaan Peserta Didik Baru) New Student Admission on the achievement path has been listed in the Minister of Education and Culture number 44 of 2019 in article 11 paragraphs 1 & 2, it is possible for new students who have achievements to not be able to enter the school they dream of being outside the zoning of residence. The limited quota for each school is only receive 5% according to the Permendikbud to new students. Achievement should also be prioritized as a driver and motivate the interest of new students. Classification of achievement participants is carried out. Where is the dataset of prospective students at school before becoming prospective participants for the next school level, as a reference for increasing track quotas achievement.Trying out the Random Tree and Random Forest methods with the Gini Index Algorithm , testing using cross validation as the initial classification model, the dataset is divided into a test and training ratio of 70: 30 then test and analysis to the Random Tree and Random Forest methods with the Gini Index Algorithm , obtained the results of an accuracy of 9 4.39% and using Random Tree with 93.48% accuracy results. But the need for further studies by utilizing more datasets and attributes and comparing to other methods because the test was carried out with 305 datasets, there were 5 attributes and 1 main attribute

1. Introduction

In Permendikbud number 44 of 2019, in article 11 paragraphs 1 and 2 it is stated that the process of accepting new students, one of which is carried out in a zoning manner, according to the statement from the Minister of Education and Culture, where the 50% quota is taken on the zoning route, 15% affirmation and 5% each achievement and moving. of school capacity. [1] The quota for student admissions based on zoning is very large compared to the quota based on merit. Achievement is highly prioritized in order to increase the motivation and interest in learning of students for the schools they dream of which are outside the zoning of their residence.

screening is to find out students who have achievements in each school, so that it becomes a reference so that later there will be changes to the increase in quota presentations for students who excel. In the process, we need a way to classify the data properly as a reference to improve the presentation on the achievement path. To classify the data, students try to use the Random Forest method with the Gini Index Algorithm which is closely related to data mining.

A technique that is used to analyze datasets and make predictions on the patterns contained in the data is data mining [2]. In data mining can be achieved by several techniques, one of which is Classification [3] . While the technique for collecting data is classification [2] . There are several mechanisms used in data mining, one of which is Random Forest (FR) combined with the Gini Index Algorithm and Random Tree combined with the Gini Index Algorithm .

There are several studies that raise about one of the classification models used in each of their studies, such as the research conducted using Random Forest and Multivariate Adaptive Regression Spline (MARS) binary response is the result of research from [4] , where the variable that has the highest dominance in HIV/AIDS status in Surabaya is age, after that type of work, having been detained for drug cases, marital

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

status, and the use of certain needles. Where the accuracy results obtained are MARS of 80.28 % , then RF MARS with 91.00% and the best accuracy results are the RF method with an accuracy of 97.80%.

Then use Random Tree to compare with the decision tree for pre-processing data [9] where the results concluded by preprocessing data handle missing value as category and missing value replenishment data from pre-processing results can be generated in the decision tree Decision tree, random tree and Random Forest. Using a random forest to predict the smoothness of credit carried out by researchers [5] , where the accuracy results obtained after carrying out several training scenario mechanisms produce an accuracy of 96.47%. Furthermore, researchers [6] using Random Forest Analysis, Multiple Regression and Backpropagation methods by predicting the apartment price index in Indonesia stated that the Backprogation method produces higher accuracy than Random Forest and Multiple Regression to minimize investment losses in buying and selling apartments during the Covid-19 pandemic. 19. So the researchers tried to do research by determining the Random Forest method with the Gini Index to improve and find out the accuracy results so that it can be used as a consideration to increase the number of quotas for new students for students who have achievements.

2. Method

The assessment for research requires steps that need to be taken to get the desired output using the Random Forest method with the Gini Index Algorithm, so it requires a flowchart to explain the entire series of processes carried out. Figure 1 is a series of Flowcharts of the system design made.

2.1 Dataset

The dataset used in this study was a dataset at SMP Negeri 22 Medan. In this dataset there are several attributes, but among the dataset only 5 attributes will be used, as shown in table 1.

TABLE 1
DATASET

Attribute	Variable
X1	Average value
X2	Category
X3	Attendance
X4	Behavior

2.2 Random Forest Method

The application of the random forest method can increase the accuracy of the results, where the child nodes for each node are carried out randomly. Generally, this method is used to build a decision tree consisting of root nodes, internal nodes and leaf nodes by taking attributes and data randomly according to applicable regulations. The root node is the topmost node, or commonly referred to as the root of the decision tree. Internal nodes are branching impulses, where these nodes have at least 1 or 2 inputs output. While the leaf node or terminal node is the final node that has one input and no output. The decision tree starts from the calculation of the entropy value as a determination of the level of attribute impurity and the value of information gain. To calculate the entropy value, the formula as in equation 1 [7] is used.

$$\text{Entropy (Y)} = -\Sigma_i \, p\,(c|Y) \log_2 p(c|Y) \qquad (1)$$

2.3. Random Tree Method

That is, learn about a decision tree in which this operator only uses a random subset of attributes for each split. So learn about decision trees, namely in nominal and numerical data. Decision tree is a powerful classification method so it can be easily understood. Collaboration with CART selects a random subset of attributes before they are applied. Where The size of the subset is determined by the section ratio parameter. [9].

2.4 Application Of The Gini Index Algorithm

Furthermore, the application of the Gini Index is the probability of two different data. The Gini Index was used by Breiman, Friedman and Olshen (1948) [8] in order to obtain results from the classification tree in the decision tree. Let S be 1 set of s number of data. This data has a number of m different classes (

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$C_i, = i = 1, \dots, m)$. Based on these classes , we can divide where S is processed into a number of m subsets (S_i, i= 1, …,m) eg Si is a dataset that is combined into a class C_i, $S_i$is the sum $S_i$of, therefore the Gini Index can be formulated as follows:

$$\text{Gini Index } (S) = 1 - \sum_{i=1}^{m} \left(\frac{S_i}{s}\right)^2 \quad (2)$$

2.5 Flowchart Flow

The initial step taken to process the data shown in Figure 1 is to preprocess the data first.

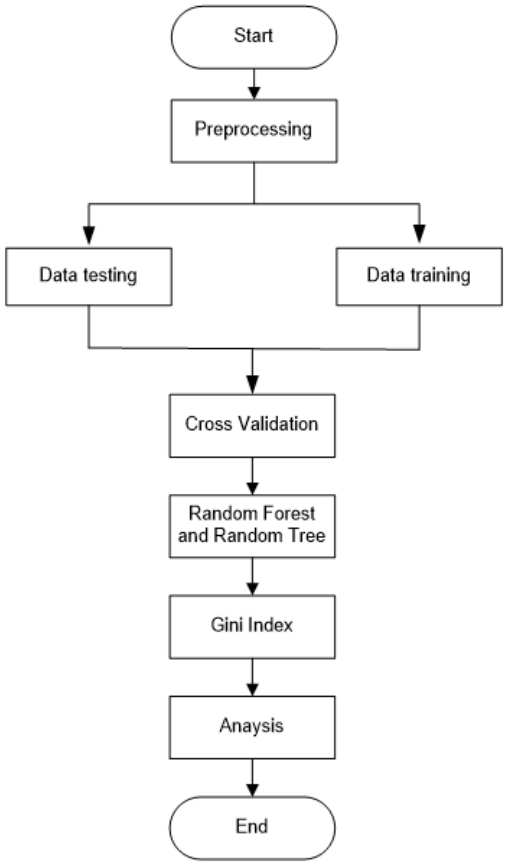

Figure 1. Predictive flow design

Explanation in Figure.1. explained that if the analysis design was carried out through several stages including reading the dataset first , the preprocessing stage then the data was divided into two ratios, namely a ratio of 70: 30 as testing data and training data after that before making predictions using Random Trees and Random Forests with algorithms. The Gini Index first goes through the Cross Validation stage which will be used as an evaluation of the predictive performance of the model.

3. Result and Disscussion

In this study the authors did several things including:

1. Look for nodes and leaves, as well as rules generated from Random Forest
2. Not all attributes in this case will be used, which only uses 5 attributes and 1 of them becomes the target attribute.
3. The input data test is carried out with a split data ratio of 70: 30

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

4. The process is carried out from the preprocessing stage, the testing stage, to the training stage, then performs a cross validation process before entering the process stage using a Random forest with the Gini Index.
5. Data accuracy.

The data used is data sourced from SMP Negeri 22 Medan. Data consisting of 8 attributes before processing the data for testing.

3.1 Random Forest with Gini Index

Split the data by separating the dataset into two parts with each ratio of 70: 30 randomly. Then the results of the comparison of the number of percentages obtained through Figure 2.

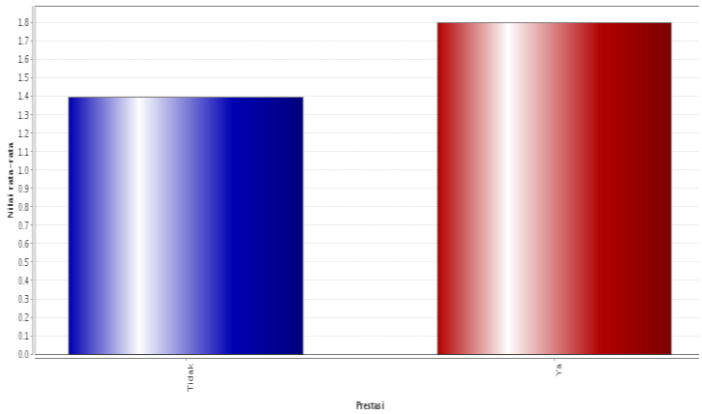

Figure 2. Chart of the results of the percentage of students.

Confusion Matrix is a method of measuring the performance of a classification method that contains information as a result of the classification carried out by the system with the classification results.

TABLE.2.
CONFUSION MATRIX

Classification Performance	Predicted Class	
Actual Class	Predicted.Yes	Predicted.No
Actual.Yes	59 (True Positive)	7 (False Negative)
Actual.No	5 (False Positive)	143 (True Negative)

Based on the table.2. then proceed with calculating the classification Accuracy value from the Random Forest classification model with the Gini Index using a dataset. Here are the results of the calculation:

$$\text{Accuracy} : \frac{TP+TN}{TP+TN+FP+FN} = \frac{59+143}{59+143+7+5} = \frac{202}{214} = 0.\ 9\ 439*100\% = 9\ 4\ .3\ 9\%$$

The knowledge generated by Random Forest with the Gini Index Algorithm is presented with a decision tree as shown in Figure.3.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

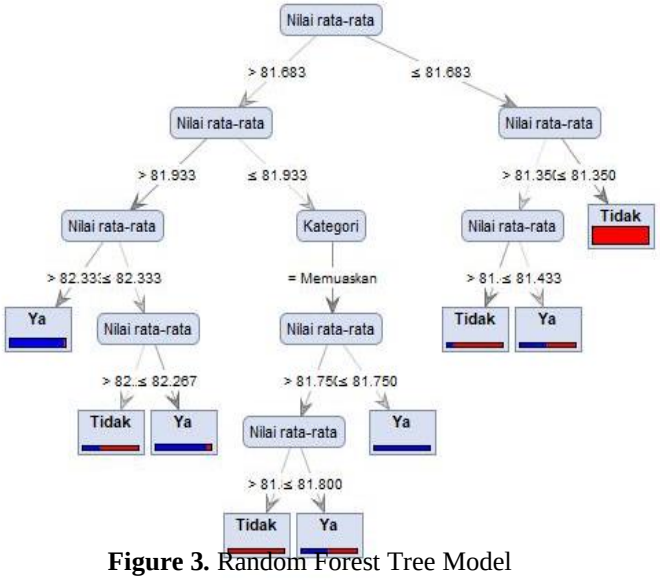

Figure 3. Random Forest Tree Model

The Tree Model can be read from top to bottom or from the root (the first node at the top) and then to the leaf (the outermost node that no longer has branches). Here's a little explanation of Figure 3.
If the value is > 81,083, because it states that the value is higher than 81,083, the achievement is immediately declared "YES" but if the value is less than 81,083 it goes to the next node with other attributes if the result is "Satisfactory" with a value > 81,933 then the result is an achievement "Yes ". And so on to the other attached nodes.

3.2 Random Tree with Gini Index

Split the data by separating the dataset into two parts with each ratio of 70: 30 randomly. Then the results of the comparison of the number of percentages obtained through Fig.4. So that it produces an accuracy of 93,46 % .

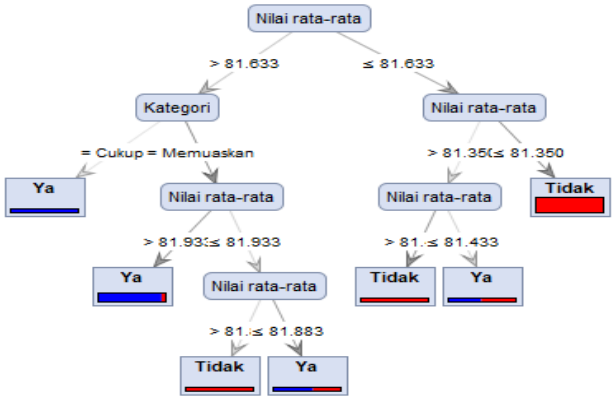

Figure 4. Random Tree Model

4. Conclusion

Based on the results of experiments conducted on Random Trees and Random Forests with the Gini Index Algorithm , the experimental results showed quite satisfactory results. Where there are 94.39 % accuracy results obtained with Random Forest and 93.46% with Random Tree . Models and rules generated by Random Tree and Random Forest with Algorithm The Gini Index is used as a basic reference for development. For further research, it can be done with a dataset with a larger number of records and make comparisons with other methods, because the author only uses 1 combined method to do the test.

5. Reference

[1]. Ministry of Education and Culture, "Regarding the Admission of New Students," in Permendikbud Number 44 , 2019.

[2]. P. Pumpuang, A. Srivihok and P. P, "Comparison of Classifier Algorithms : Bayesian Network, C4.5 Decision Forest and NBTree for Course Registration Planning Model of Under," 2008 IEEE International Conference on Systems, Man and Cybernetics, pp. IEEE. 3647-240, 2008.
[3]. D. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining," Jhon Willey & Sons, pp. Inc. 129-240, 2005.
[4]. Nidhomuddin and B. Otok, "Random Forest and Multivariate Adaptive Regression Spline (Mars) Binary Response for Classification of HIV/AIDS Patients in Surabaya," Statistika, vol. 3(1), December 2, 2015.
[5]. MI Putra, A. Yusuf and N. Yalina, "Classification of Smooth Credit by the Random Forest Method," Systemic : Information Systems and Informatics Journal, vol. No. 5, pp. 7-12, 2 December 2019.
[6]. NY Saputra, S. Saadah and PE Yunanto, "Analysis of Random Forest, Multiple Regresstion and Backprogation Methods in Predicting Apartment price index in Indonesia," Scientific Journal of Electrical Computer and Informatics Engineering (JITEK), vol. 7 No.2, No. ISSN : 2338-3070,001:10:26555/JITEKI.V712.20997, pp. 238-248, August 2021.
[7]. S. JK, F. F and R. Dekker, "An-Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," In International Conference on Application of Natural Language to Information Systems, pp. 48-59, 2016.
[8]. I. Breiman, J. Friedman, C. Shone and R. Oslen, "Classification And Regression Tress," 1984.
[9]. Saifullah, Muhammad Zarlis, Zakaria, Rahmat Widia Sembiring," Analysis of the Comparison of the Decision Tree Algorithm with the Random Tree Algorithm for Pre-Processing Data." Journal of Computer Science & Informatics (J-SAKTI), vol.1 no.2 Sepetember 2017 no. ISSN : 2548-9771/EISNN : 2549-7200.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.