BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume 19 Issue 1 Page 0227–0236
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss1pp0227-0236

A COMPARISON OF RANDOM FOREST AND DOUBLE RANDOM FOREST: DROPOUT RATES OF MADRASAH STUDENTS IN INDONESIA

Arie Purwanto [1*], Bagus Sartono[2], Khairil Anwar Notodiputro[3]

[1]*Mathematic Study Program, Faculty of Teacher Training and Education, Universitas Mercu Buana Yogyakarta*
Jln. Raya Wates 10, Bantul, DIY, 55752, Indonesia

[2,3]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University*
Jln. Raya Dramaga, Bogor, Jawa Barat, 16680, Indonesia

*Corresponding author's e-mail: * **arie@mercubuana-yogya.ac.id***

ABSTRACT

Random forest algorithm allows for building better CART models. However, the disadvantage of this method is often underfitting, especially for small node sizes. Therefore, the double random forest method was developed to overcome this problem. The research was conducted by utilising Education Management Information System (EMIS) data, which is related to the incidence of school dropout. The data used consists of 2 data, namely MTs and MA dropout data. The initial testing procedure was carried out using the random forest algorithm for each data set, then the data was evaluated using the double random forest method. From this study, the underfitting case can be overcome well using the double random forest algorithm, while in the fit case, the difference in the goodness-of-fit value of the model is relatively the same. The results obtained show that MTs prioritise school quality more than MA, although family factors are more important at the MA level. Although the total number of factors used is basically the same, it should be noted that the two school levels have different relevance variables. It should be noted that no forecasting was done in this study given that the methodology used two different types of data.

Article History:

Received: 22nd May 2024
Revised: 22nd November 2024
Accepted: 22nd November 2024
Published:13th January 2025

Keywords:

Random Forest;
Double Random Forest.

How to cite this article:

A. Purwanto, B. Sartono and K. A. Notodiputro., "A COMPARISON OF RANDOM FOREST AND DOUBLE RANDOM FOREST: DROPOUT RATES OF MADRASAH STUDENTS IN INDONESIA," *BAREKENG: J. Math. & App.*, vol. 19, iss. 1, pp. 0227-0236, March, 2025.

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

1. INTRODUCTION

Education is a basic need and an important aspect of human life. Not only for humans themselves, life plays a big role in the development of a nation. For the Indonesian people, this has become a strong philosophy in building the nation. The law of the Republic of Indonesia number 2 of 1989 article 5 states that every citizen has the same right to obtain an education. However, many factors cause of many the nation's sons and daughters to be unable or unable to participate in education. This can be seen from the large number of school dropout cases in Indonesia. According to [1] it is stated that in 2022 the incidence of dropping out of school in high school will be 1.38%, middle school will be 1.06%, and elementary school will be 0.13%. Not only at the primary and secondary education levels, similar cases also occur in higher education. There are several indications of the causes of teenagers or students in various regions in Indonesia, as mentioned in [2] **e**conomic conditions, academic satisfaction, academic achievement, and family economy are the most influential. In line with this, other influential indicators according to [3] are KIP/PIP ownership, number of household members, working children, poverty, and area of residence.

Indonesia which is rich in culture and ethnicity. Based on demographic data [4] noted that in 2020 Indonesia's Muslim population will currently reach 229.62 million people or around 87.2% of Indonesia's total population of 269.6 million people. In Indonesia, there are many Islamic-based schools. Madrasa is one type of education provider in Indonesia. According to [5] madrasa are schools that are characterized by Islam so madrasa curriculum has a greater burden than the curriculum in schools. This is because madrasa teaches all subjects at school plus religious subjects, which are more than the available lesson hours. Based on this problem, there is an interesting thing, namely the problems currently being faced by madrasa as education providers. According to [6] Islamic-based education is mired in decline, backwardness, helplessness, and poverty, as is also experienced by most Islamic countries and communities compared to those who are Non-Islamic. In line with this, [7] states that several solutions to problems that must be resolved are problems of facilities and infrastructure, increasing teacher qualifications, education costs, scientific perspectives, and community participation. Furthermore, the urgency faced also comes from students. So far, based on surveys conducted, on average there are around 1.02% of students dropping out of school at the Tsanawiyah or junior high school level and 1.17% of students dropping out of school at the Aliyah or senior high school level by district or city level in Indonesia. This can be said to be quite high considering that the number of students at that level can be said to be large.

In essence, a statistician tries to uncover and solve problems related to the data problems faced. One way to do this is to create a model that represents the data. One modelling technique that is still considered young is machine learning. In the last few decades machine learning has become a trending topic in the world of science. Machine learning and artificial intelligence have experienced a new wave of publicity fueled by enormous and ever-increasing amounts of data and computing power and the discovery of better learning algorithms [8]. One of the most popular machine learning algorithms is the Random Forest (RF) ensemble method. This method was introduced by Breiman in 2001 by defining a random forest as a combination of predictor trees such that each tree depends on random vector values that are sampled independently and with the same distribution for all trees in the forest [9]. The basic principle of random forest is the collection of decision trees that are built randomly [10]. The basic concept in Classification and Regression Tree (CART) which seeks to optimize a predictor is also the basis for building random forest methods. In the CART concept, a set of predictors are built in a decision tree, the predictors in question are not necessarily able to optimize the tree model that is built. Therefore, each tree may experience random disturbances. Each tree is then collected and known as the forest. It is further from this forest that a wider exploration is carried out of all possible tree predictions which in practice produce better prediction performance via the random forest method.

As science develops, new methods also develop to complement or evaluate existing methods to optimize existing methods. In the scope of machine learning, the Double Random Forest (DRF) method can be said to be new for evaluating special cases of random forests. The ensemble double random forest method can be said to be new because it was first released in 2020 by Han, Sunwoo., et., al. This new method uses a bootstrap technique on each node during the tree-building process rather than just bootstrapping once on the root node as in RF, furthermore this method, in turn, produces a more diverse set of trees, thus enabling more accurate predictions [11]. One of the goals to be achieved with the double random forest method is to overcome cases where the random forest method experiences underfitting [12]. However, questions arise when the random forest method does not experience underfitting. In this case, the researcher wants to know

the effectiveness of the double random forest method when experiencing underfitting and not experiencing underfitting. Therefore, case studies are needed, which in this case emphasize the topic of education.

Based on what has been explained previously, researchers want to conduct an assessment of machine learning-based statistical methods using education data. The education data comes from the Education Management Information System (EMIS) data, from the Directorate General of Islamic Education, Ministry of Religious Affairs of the Republic of Indonesia. The data used is related to the incidence of school dropouts and variables that are considered to affect the events studied in 2022/2023. In this study, two data are used which will be modelled using random forest and double random forest algorithms. The data used are study data on the percentage of school dropout rates at the Madrasah Aliyah (MA) and Madrasah Tsanawiyah (MTs) levels. Of the two data, it is expected that one can be said to be a fit model while the other experiences underfitting. The purpose of this study is to compare the performance of random forest and double random forest methods, especially if the model formed is underfitting. It is hoped that the evaluation carried out can provide an overview of the concept of using the two methods.

2. RESEARCH METHODS

This research was designed to determine the impact of models experiencing underfitting and fit by using two methods, namely the Random Forest (RF) and Double Random Forest (DRF) algorithms. The empirical case studied is the percentage of school dropouts at the Tsanawiyah (MTs) and Madrasah Aliyah (MA). The data used is secondary data from the Ministry of Religion of the Republic of Indonesia in 2023. Both data will be further modelled using the RF method and re-evaluated using the DRF method. The goal to be achieved is to be able to evaluate the performance of the two methods, namely RF and DRF. Furthermore, it is hoped that the procedures carried out can provide an equation model that can be predicted as material for evaluation studies related to similar events in the future.

2.1 Data

The data used is a type of secondary data obtained from the Education Management Information System (EMIS), Directorate General of Islamic Education, Ministry of Religious Affairs of the Republic of Indonesia. The data used relates to the incidence of dropout and variables that are considered to affect the incidence studied in 2022/2023. There are 950 observational data used in this study with 475 observations representing the dropout rate at the MTs level and 475 at the MA level. The amount of data represents city districts whose data is taken from all regions of Indonesia. In general, there are 6 variables used consisting of 1 response variable and 7 predictors. So far, the response and predictor variables used are the same at both the MTs and MA levels. To give unique characteristics to the data used, two variables were added at the MTs level for the district or city concerned. The additional data used is the percentage of poverty for each region and the percentage of average expenditure per capita for a week according to fish groups in 2022/2023 which can be accessed via the BPS Indonesia website (https://www.bps.go.id). Next, the data used will be modelled using two algorithms, namely random forest and double random forest. A more detailed description is presented in **Table 1**.

Table 1. Description of Research Variable

No	Education levels	Variable	Variable Description
1	MTs	Y_1	Percentage of students dropping out of school at MTs level
		X_{11}	The ratio of the number of students to the number of MTs
		X_{12}	The ratio of the number of students to MTs teachers
		X_{13}	Average number of family members of MTs students
		X_{14}	Percentage of fathers with junior high school education or less
		X_{15}	The percentage of MTs accredited at least B

No	Education levels	Variable	Variable Description
		X_{16}	Percentage of poor people
		X_{17}	Percentage of average expenditure per capita for a week by fish group
2	MA	Y_2	Percentage of students dropping out of school at MTs level
		X_{21}	The ratio of the number of students to the number of MTs
		X_{22}	The ratio of the number of students to MTs teachers
		X_{23}	Average number of family members of MTs students
		X_{24}	Percentage of fathers with junior high school education or less
		X_{25}	The percentage of MTs accredited at least B

2.2 Model

This research uses two modelling methods, namely the RF and DRF methods. These two methods have different models and modelling procedures. Below we will discuss the procedures for using RD and DRF.

2.2.1 Random Forest Algorithm

Random forest is a substantial modification of collecting many independent collections of trees and then making an average of these collections **[13]**. The algorithm used is as follows:

1) for $b = 1$ until B do bootstrap sample $\boldsymbol{Z}^*$of size N from train data
   (1). Build a random forest tree T_b on the bootstrapped data, by repeating the following steps for each terminal node in the tree, until the minimum node size is reached. Choose m variable randomly on p variable
   (2). Take the best variable/splitting point between m.
2) The output of the ensemble tree is$\{T_b\}_1^B$

Note: When making predictions at a new point, namely x, attend if:

$$\text{regression: } \hat{f}_{rf}^B = \frac{1}{B}\sum_{b=1}^{B} T_b x \quad (1)$$

2.2.2. Double Random Forest

One other algorithm that is based on the random forest concept is the double random forest. In contrast to the classic random forest, the DRF algorithm tries to modify the training data using the original data set to train the base so that it produces samples that are more unique exclusively than the classic random forest **[14]**. The algorithm used is as follows:

1) for $b = 1$ to B do bootstrap sample $\boldsymbol{Z}^*$of size N from training data $\boldsymbol{D}$, with $\boldsymbol{D}$ is the training data set with *n* sample, *p* predictor, and target variable.
   (1). For each node *t*
   if $n_t > n \times 0.1$, build a bootstrap sample $\boldsymbol{D}_t^*$ from $\boldsymbol{D}_t$ with $\boldsymbol{D}_t$ is training data set with n_t sample, *p* predictor, and target variable for nodes *t*.
   (2). Build bootstrap $\boldsymbol{D}_t^*$ from $\boldsymbol{D}_t$ and $\boldsymbol{D}_t^* = \boldsymbol{D}_t$ for the other
   (3). Select randomly from m variable of $\boldsymbol{D}_t^*$
   (4). Take the best variable/splitting point between *m*.
2) The output of the ensemble tree is $\{D_t^{*b}\}_1^B$

Note: When making predictions at a new point, namely x, attend if:

$$\text{regression: } \hat{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} D_t^{*b} x \quad (2)$$

2.3 Goodness of Fit Models

To evaluate the analysis procedure that will be carried out, the researcher uses a measure of model goodness. A goodness-of-fit measure is a characteristic used to examine a more complex relationship between actual conditions and the results obtained [15]. Many measures can be used as criteria for goodness-fit models. One that can be used is the Mean Absolute Percentage Error (MAPE). This measure is obtained by calculating the absolute error for each period divided by primary or actual data, then averaging the absolute error [16] The MAPE value is created and designed to calculate the percentage difference between the predicted and actual data or primary data.

$$\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \left|\frac{\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}}{\boldsymbol{y_i}}\right| \quad (\mathbf{3})$$

Where:

for $\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}$

n is the number of observations

y_i is the i-th actual value

$\hat{y}_i$ is the i-th predicted value

Furthermore, the lower the MAPE value, the better the ability of the model used in forecasting, then the value based on [17] the range of values that can be used as measurement material regarding the ability of a model in forecasting is presented as follows:

Table 2. Range and Value Description of MAPE

Range	Value Description
< 10%	The ability of the forecasting model is very good
10 – 20 %	Good forecasting model capability
20-50 %	Enough forecasting model capability
>50	Poor forecasting model capability

2.4 Variable Importance of Models

In predictive models, a measure called significant variables is used to determine how much influence each input variable (feature) has on model predictions. Based on [18] states that identifying the importance of variables for computational models or measured data is a very important task in many applications. Many disciplines create their significant variables. We are very aware of how important it is to collect all the good practices in each discipline and compare the relative merits of each method. The goal is to help practitioners choose the best method to meet various analysis objectives and to guide current research further, [19] describes the two main methods introduced: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). Meanwhile, decreased impurity [13] is used for decreased impurity AA (such as Gini impurity and entropy) to determine how important each feature is in the decision tree to separate the data.

2.5 Analysis Procedure

The data analysis procedures carried out in this research are described as follows:

1) Exploration of the required data to be able to see the general condition of the variables to be analyzed.
2) Divide each data, namely data on MTs and MA dropout rates, into two parts: training data 75% and testing data 25%. Furthermore, training data is generally used in modelling, while testing data is used for model evaluation.

3) Building an initial model and setting hyperparameters for the two data used. The performance of many machine learning methods is highly dependent on hyperparameter settings [20]. The stage aims to enable the best model to be obtained which can be built using the random forest algorithm as a reference in the analysis stage which will be carried out next using double random forest.
4) Model both data using the random forest algorithm with 100 repetitions. The data tested at this stage is data on the percentage of school dropouts at the MTs and MA levels.
5) Evaluate the results obtained whether they experience overfitting, fit, or underfitting based on MAPE. There are possibilities to occur in testing training and testing data where the data can experience overfitting, fit, and underfitting. Overfitting and underfitting incidents generally occur due to errors in training and testing. Overfitting occurs if the model fits the data very well whereas underfitting occurs whenever the model or algorithm is not applied to the data [21]. Underfitting events can be shown by looking at high MAPE values on training and testing data. In this study, underfitting was determined if the MAPE value was above 30%.
6) Model both data using the double random forest algorithm.
7) Evaluate all resulting models. In general, the evaluation is based on the MAPE values produced both for the dropout percentage model at the MTs level using the RF and DRF method and at the MA level using the RF and DRF method.
8) Consider the important variables that influence the response

3. RESULTS AND DISCUSSION

3.1 Data Exploration

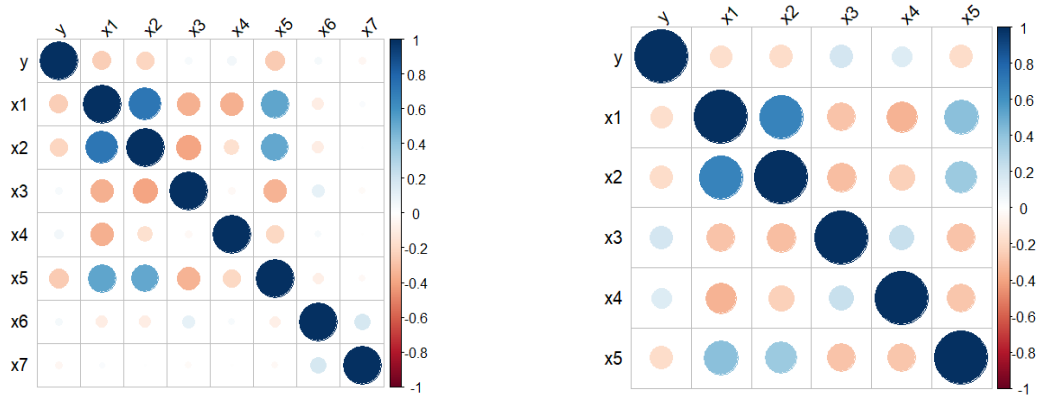
Based on what was explained in the previous chapter, there are two data used in this research. The first data is related to the percentage of students who have dropped out of school at the MTs level along with their predictor variables and the second data is related to the percentage of students who have dropped out of school at the MA level. The results of the data exploration carried out are presented in Table 3 as follows:

Table 3. Summary of statistics variables

No	Education Level	Variable	Statistic			
			min	median	Mean	max
1	MTs	Y_1	0.00	0.833	1.02	6.74
		X_{11}	21.57	164.42	177.81	577.33
		X_{12}	1.91	10.58	10.46	19.51
		X_{13}	4.06	5.04	5.11	6.99
		X_{14}	16.52	78.57	75.43	100.00
		X_{15}	15.06	82.31	78.59	100.00
		X_{16}	2.27	9.25	10.37	31.78
		X_{17}	0.00	0.07	0.07	0.25
2	MA	Y_2	0.00	0.88	1.19	14.94
		X_{21}	7.00	148.9	170.2	1229.0
		X_{22}	1.00	8.90	8.99	16.61
		X_{23}	3.00	5.33	5.47	9.52
		X_{24}	0.00	92.44	86.28	100.00
		X_{25}	0.00	64.29	60.79	100.00

Based on the data description that has been presented, the two data used generally have different characteristics. It can be seen that the maximum percentage of students dropping out of school based on data at the MA level is higher than MTS, namely 14.94% in North Lombok district, West Nusa Tenggara.

Furthermore, for the ratio variable between madrasahs and students, which can be interpreted as school availability to student capacity, there is a difference where the maximum ratio is 1229.0 for MA, namely Perada in Tegal City, Central Java, while at the MTs level, it is 577.33 in Empat Lawang district, South Sumatra. The ratio variable between teachers and students can be said to be relatively the same or stable. Furthermore, the average number of families based on the average is relatively the same, namely 5 family members. However, at the MA level, there is a fairly high figure, namely 9.52, in Bitung City, North Sulawesi. For the variable percentage of father's education at the junior high school level and below at the MA level, there are 87 districts or cities with a percentage of 100 and 10 city districts with a percentage of 100. Regarding the variable percentage of madrasahs accredited at least B, it can be seen that all MA have been accredited A and there are still 15.06 MTs that are accredited at least B. The poverty level variable used at the MTs level shows the highest poverty level, namely in Musi Banyuasin district, South Sumatra. On the other hand, the average percentage of expenditure for fish consumption is 0.25%, namely in Central Mamuju Regency, South Sulawesi. Data exploration was also carried out by looking at the correlation between variables to ensure that there was no linear correlation occurring in the data at either the MTs or MA level. In the data used, the highest correlation was obtained at 0.72 between the variable ratio of the number of students to madrasahs and the ratio of the number of students to teachers at the MTs level and 0.66 at the MA level for the same variable. However, both variables are still used in the model that will be built. The correlation graph between variables is presented in **Figure 1** as follows:



(a) **(b)**

Figure 1. The Correlation Graph Between Variables,
(a) Correlation in MTs level, (b) Correlation in MA level

3.2 Evaluation of Data Analysis Using Random Forest and Double Random Forest Algorithms

Data that has been confirmed to be used or passed the initial stage of data selection is then processed using the random forest algorithm. The procedures carried out start from the data division stage, hyperparameter tuning, to evaluation through MAPE calculations. The procedure is carried out to see the quality of the model on the two data used. To provide consistent results, 1000 repetitions were made in the random forest algorithm used. A summary of the procedures performed is presented in **Table 4** as follows:

Table 4. Summary of random forest analysis procedure

No	Education level	Type of data	Data percentage	hyperparameter				MAPE	Information
				mtry	Ntree	maxnode	Nodesize		
1	MTs	Train	75%	2	1000	75	9	29.82	Fit
		Test	25%	2	1000	75	9	24.65	
2	MA	Train	75%	2	500	75	7	58.58	Underfit
		Test	25%	2	500	75	7	64.38	

Based on **Table 4**, it can be seen that the treatment of the two data is different, which is based on hyperparameter settings. The parameters used in the analysis were selected based on hyperparameter tuning at the MTs' education level. This is because there are two criteria conditions used in the methodology, namely fit and underfit. The criteria in the procedure carried out provide hyperparameter values with mtry = 2,

Ntree = 1000, maxnode = 75, and Nodesize = 9 for MTs education level, as well as hyperparameter values of mtry = 2, Ntree = 500, maxnode = 1, and Nodesize = 1 for MA education level. The model at the MTs education level can be said to be fit considering the MAPE value obtained is relatively small, namely 29.82 on training data and 24.65 on testing data. While at the MA education level, the model is said to be less fit with a MAPE value of 58.58 on training data and 64.38 on test data. The next stage is to conduct an analysis using the double random forest algorithm to see the comparison between the two methods used. In this case, the hyperparameter value used is fixed as in the case of using the random forest algorithm. In addition, it is still repeated 1000 times with the same indicators, and the following results are obtained:

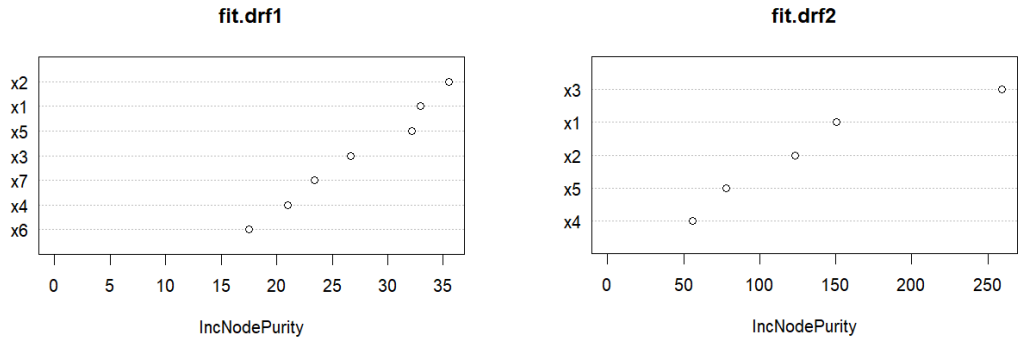
Table 5. Summary of double random forest analysis procedures

No	Education level	Type of data	Data percentage	Hyperparameter				MAPE	Information
				mtry	Ntree	Maxnode	Nodesize		
1	MTs	Train	75%	2	1000	75	9	24.85	Fit
		Test	25%	2	1000	75	9	22.13	
2	MA	Train	75%	2	500	75	7	27.14	Fit
		Test	25%	2	500	75	7	29.58	

Based on **Table 5**, it can be seen that the two levels of education, namely MTs and MA, have different information about where both fit. This is based on the two low MAPE value differences and not much difference between the two. It can be seen again in **Table 4** and **Table 5** where the education data at the MTs level is Fit, the MAPE values for both are relatively the same or there is no significant difference for both the test data and the test data. However, based on **Table 4** and **Table 5** where the education data is at the MA level, it can be seen that by applying the random forest algorithm this data is underfitted, but by using the double random forest algorithm the data is fit. From here we can also see that there is a difference in the MAPE value between the two algorithms used using both random forest and double random forest. This provides an understanding that based on the treatment that has been done, the double random forest algorithm will be better if implemented in conditions where the random forest algorithm is underfitted. On the other hand, in the case of the random forest algorithm in the fit condition, there is no significant difference between the two algorithms used. Underfitting happens when the model is too simplistic to capture patterns in the data, resulting in inferior prediction accuracy. Double Random Forest uses two levels of randomisation (for example, additional randomisation on input variables or tree structure), which improves the model's ability to capture data changes. As a result, DRF is less susceptible to underfitting, especially if the data includes DRF combines two layers of random sampling to reduce model bias and better capture non-linear patterns, which are frequently missed by ordinary Random Forest. DRF's added randomisation generates a more diverse ensemble, allowing the model to avoid underfitting. Simultaneously, the model decreases the risk of making overly broad assumptions based on little evidence.

3.3 Variabel Importance of Models

In this research, Mean Decrease Impurity (MDI) is a method used in predictive models such as random forest and double random forest to measure variable importance by assessing how much impurity (such as Gini impurity or entropy) decreases each time a feature is used to split data at nodes in a decision tree. Each time the decision tree uses a particular feature to split the data, the decrease in impurity is calculated and accumulated for that feature. The larger the impurity drop caused by a feature, the higher the importance of that feature. MDI provides insight into the relative contribution of each feature in making accurate predictions, allowing researchers to identify the most influential features in the model. With this information, researchers can perform dimensionality reduction by ignoring less important features, which not only improves model efficiency but also helps in the interpretation of model results. In addition, MDI helps in a deeper understanding of the relationship between variables in the context of the problem under study, providing a solid foundation for making more informed decisions based on the results of data analysis. The two plots used are based on the results of the analysis with the double random forest method for the whole data consisting of important variables in the case of dropouts at the MTS level and important variables in the case of dropouts at the MA level. This is because the resulting MAPE goodness value is relatively smaller than the random forest method. The MDI plot is presented in **Figure 2** as follows:



(a) **(b)**

Figure 2. Mean Decrease Impurity of Model,
(a) Important Variables in The Case of Dropouts at MTs Level,
(b) Important Variables in The Case of Dropouts at MA Level

Based on the resulting plot, it shows that at the MTS level the variables x2, x1, and x5, namely the ratio of the number of students to the number of MTs, the ratio of the number of students to MTs teachers, and the percentage of MTs accredited at least B are variables of importance in compiling the model. Furthermore, at the MA level, it shows that the average number of family members of MTs students is a variable of importance. In this case, it illustrates that different variables affect the dropout rate at each level. At the MTs level, it is more based on the quality of the school itself but at the MA level, it is more on family factors. It is quite interesting that the two levels provide different importance variables.

4. CONCLUSIONS

Based on the description that has been presented, it can be concluded based on the case study used that in the condition implementing the random forest algorithm in underfitting conditions, the implementation of the double random forest algorithm shows a significant difference. This provides an understanding that based on the treatment that has been done, the double random forest algorithm will be better if implemented in conditions where the random forest algorithm is underfitted. On the other hand, in the case of the random forest algorithm in the fit condition, there is no significant difference between the two algorithms used. Underfitting happens when the model is too simplistic to capture patterns in the data, resulting in inferior prediction accuracy. Double Random Forest uses two levels of randomisation (for example, additional randomisation on input variables or tree structure), which improves the model's ability to capture data changes. To identify the most influential features in the model the Mean Decrease Impurity (MDI) is used, this is because MDI provides insight into the relative contribution of each feature in making accurate predictions. The results of the MDI plot show that the percentage of MTs accredited at least B, the ratio of students to MTs, and the number of students to MTs teachers are significant factors in the model's compilation at both levels where there are various relevant variables for MTs. The average number of family members of students is also a significant variable at the MA level. At the MTs level, the quality of the school is a major factor in student dropout, while at the MA level, the average family member is significant. The fact that the two levels of education offer different variables means that the approach to dealing with them is also different. Improving school quality at the MTs level deserves more attention. On the other hand, at the MA level, it was found that the average family size had a significant effect on dropping out of school. Thus, the researcher provides several responses related to the possibilities that can be used as policies including free school fees, and an evaluation of the zoning system that allows a small number of public schools in the location of students. Another thing to be highlighted or studied in the future is whether students can continue their education at a higher level at a higher cost.

ACKNOWLEDGMENT

I would like to extend my heartfelt gratitude to the lecturer for the Advanced Data Analysis (STA17151) course IPB, for their invaluable guidance, insightful feedback, and unwavering support throughout this research. Your expertise and dedication to teaching have significantly enriched my understanding and application of advanced data analysis techniques. I am also immensely grateful to the Ministry of Education and Culture of the Republic of Indonesia (Kemendikbud) for awarding me the BPI scholarship, which has provided the financial assistance necessary to pursue my studies and conduct this research. This support has been crucial in enabling me to focus fully on my academic endeavours and professional growth. Lastly, I would like to acknowledge all those who have offered their encouragement and support during this research journey.

REFERENCES

[1] C. Indonesia, "Pembelajaran Daring Bisa Tekan Angka Putus Sekolah," CNBC Indonesia, Jakarta, 2022.

[2] e. a. Nurmalitasari, "Factors Influencing Dropout Students in Higher Education," *Education Research International*, vol. 2023, no. education, p. 13, 2023.

[3] A. Hakim, "Faktor Penyebab Anak Putus Sekolah," *Jurnal Pendidikan*, vol. 21, no. 2, pp. 122-132, 2020.

[4] K. A. R. Indonesia, "Menjadi Muslim, Menjadi Indonesia (Kilas Balik Indonesia Menjadi Bangsa Muslim Terbesar)," Kementrian Agama Republik Indonesia, Jakarta, 2020.

[5] A. Nurriqi, "Karakteristik Pendidikan Agama Islam di Madrasah Prespektif Kebijakan Pendidikan," *Bintang Jurnal Pendidikan dan Sains*, vol. 3, no. 1, pp. 124-141, 2021.

[6] I. Adelia and O. Mitra, "Permasalahan Pendidikan Islam di Lembaga Pendidikan Madrasah," *Islamika: Jurnal Ilmu-ilmu Keislaman*, vol. 21, no. 01, pp. 32-45, 2021.

[7] I. Turmidzi, "Pengelolaan Pendidikan Bermutu di Madrasah," *Tarbawi*, vol. 4, no. 2, pp. 165-181, 2021.

[8] S. e. Badillo, "An Introduction to Machine Learning," *Clinic Pharmacology and Therapeutics*, vol. 107, no. 04, pp. 871-885, 2020.

[9] L. Breiman, "Random Forests," *Kluwer Academic Publishers. Manufactured in The Netherland*, vol. 45, pp. 5-32, 2001.

[10] Genuer, Robin and J.-M. Poggi , Random Forest with R, Cham: Springer Cham, 2020.

[11] S. Han, H. Kim and Y.-S. Lee , "Double random forest," *Springer*, vol. 109, p. 1569–1586, 2020.

[12] A. N. A. Aldania, A. M. Soleh and K. A. Notodiputro, "A Comparative Study of CatBoost and Double Random Forest for Multi-class Classification," *Jurnal Resti : Rekayasa Sistem dan Teknologi Indformasi*, vol. 7, no. 1, pp. 129 - 137, 2023.

[13] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, New York: Springer, 2008.

[14] ,. M. e. a. Ganaie, "Heterogeneous Oblique Double Random Forest," *arXiv*, pp. 1-35, 2023.

[15] W. E. Hipson and D. G. Séguin, "Goodness of Fit Model," *Springer International Publishing*, no. Encyclopedia of Personality and Individual Differences,, 2017.

[16] Z. A. Sari and M. Andarwati, "PERAMALAN DOUBLE MOVING AVERAGEDAN DOUBLE EXPONENTIAL SMOOTHING JUMLAH PENUMPANGDI STASIUN KOTABARUMALANG," *JournalofInformationSystemsManagementandDigitalBusiness*, vol. 1, no. 2, pp. 263-272, 2024.

[17] A. H. e. a. Hutasuhut, "PEMBUATAN APLIKASI PENDUKUNG KEPUTUSAN UNTUK PERAMALAN PERSEDIAAN BAHAN BAKU PRODUKSI PLASTIK BLOWING DAN INJECT MENGGUNAKAN METODE ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE) DI CV. ASIA," *Jurnal Teknik ITS*, vol. 3, no. 2, 2014.

[18] P. Wei , Z. Lu and J. Song, "Variable importance analysis: A comprehensive review," *Science Direct*, vol. 142, pp. 399-432, 2015.

[19] L. Breiman , "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[20] F. Hutter, H. Hoos and K. L. Brown, "An Efficient Approach for Assessing Hyperparameter Importance," in *Proceedings of Machine Learning Research*, 2014.

[21] M. A. Salam and e. al., "The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem," *(IJACSA) International Journal of Advanced Computer Science and Applications,*, vol. 4, p. 12, 2021.