International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Multivariate Random Forest to Identify the Importance Variable of 8 National Education Standards toward National Examination of Student High School in Indonesia

Ardiana Alifatus Sa'adah[a]*, Indahwati[b], Budi Susetyo[c]

[a,b,c]*Department of Statistics, IPB University, Jl. Raya Dramaga, 16680 Bogor, Indonesia*

[a]*Email: ardianaalifatus@gmail.com*

Abstract

Quality of human resources is one of the important aspect in terms of national development. One way that can be used to improve the quality of human resources in Indonesia is by improving the quality of the education. Therefore, the quality of education in Indonesia needs to be considered. The quality of education is the level of conformity between education implementers with the National Education Standards (SNP) in schools. One of the factors that is used to measure the level of success of SNP can be evaluated from National Examination (UN). Therefore it is necessary to do an analysis to find out the important factors of 8 SNP indicators which have a high influence on the UN results. The response variable is the average of national exam scores of the three main subjects tested. The response variables are numerical and multivariate and also have a high correlation between the scores of the three subjects. Based on these considerations, the Multivariate Random Forest (MRF) analysis method was applied. The results of the analysis that can be taken in this study are that the MRF method is able to identify the model stable even though it only uses training data with a cut off of 5%. The results of the analysis of importance variable from 8 variables of the national education standard toward variables of national examination scores, obtained 3 standards with the highest level of importance that are the competency standard of graduates (SKL), content standards (SI) and management standards (SPL).

Keywords: multivariate random forest; national education standards; national exam.

---

* Corresponding author.

1. Introduction

Education is an important factor in improving the quality of human resources in Indonesia. The quality of education is the level of conformity between the implementation of education and the National Education Standards (SNP) in schools. SNP is a minimum standard by the government in the field of education. SNP consists of eight standards namely content standards (SI), process standards (SPR), graduate competency standards (SKL), educator and staff standards (SPT), facilities and infrastructure standards (SSP), management standards (SPL), financing standards (SB), and education assessment standards (SPN). SNP itself is a benchmark for various aspects related to the implementation of the national education system. The results of the SNP fulfillment are explained in the form of accreditation, the assessment of which is carried out by the National Accreditation Board (BAN). To see the success of the quality of education, certainly can not be separated from how the results of evaluating the teaching and learning process in Indonesia. Indicators of the success of the learning process can be seen through the results of the National Examination (UN). Some educational theories that explain the causality of the eight SNPs were published in the Ministry of National Education and Ministry of Religion in 2010, the Ministry of Education and Culture in 2012, and the Ministry of Education and Culture in 2017[7,8]. Several studies on the relationship of causality of SNP to academic achievement have also been carried out, for example Setiawan and his colleagues applied the GSCA method to compare the relationship of accreditation results with the national exam for junior high schools [13], Wahyuni analyzed the relationship between 8 SNPs and UNBK at the junior high schools level using the fuzzy clusterwise GSCA method [15] and Ramadhan used random forest classification modeling to identify important factors in improving the quality of high school education [11]. The focus of modeling lately has shifted toward prediction with an emphasis on deeper descriptions and explanations. Classification and Regression Tree (CART) is one of the classification techniques for constructing prediction models by exploring data. CART was first proposed by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone in the 1980s [9]. CART produces a Classification Tree (CT) if the response variable is categorical, and a regression tree (RT) if the response variable is numeric [2]. CART is a non-parametric classification method so no assumptions are needed to be fulfilled. But Berk explained the weakness of the CART method is that it is unstable if an example of training data from a similar population is used, it is very likely that the results of the classification tree will be different [1]. To overcome the weaknesses of CART, the Random Forest (RF) method was developed by Leo Breiman in 2001. The RF method is one of the combined tree development methods from the CART method by applying the bootstrap aggregating (bagging) method and random feature selection [3]. According to Miller and his colleagues a model with a combined tree shows high accuracy and powerful prediction ability in various fields of application [10]. CART and RF methods so far have mostly been applied to single response variables (univariate), while the academic achievement variable as a response variable is a UN score from several subjects so that it is numerical and multivariate. In this study, academic achievement was measured through the average of the UN in 3 subjects tested that is Mathematics, Indonesian and English per school. In a study conducted by Setiawan and his colleagues (2018) there is a strong correlation between the average scores of subjects tested on the national exam. Based on this, we need an analytical method that is able to accommodate multivariate response variables that have a high correlation. De'ath proposes the development of the RT method, the Multivariate Regression Tree (MRT) where this method can be used to accommodate multivariate response

variables [4]. Furthermore, for the purpose of increasing the accuracy and prediction of MRT, a method of combining MRT with RF was developed by Segal & Xiao in 2011 namely Multivariate Random Forest (MRF) [12]. MRF is able to accommodate multivariate response variables by combining the MRT method with bootstrap resampling and predictor subsampling from tradisional random forest [10]. In addition, the MRF method can be used to determine important factors that influence the multivariate response variables. The aim of this study is to apply the MRF method to find out the important factors of 8 SNPs that influence the results of the average scores of the National Examination (UN) for high school students in 2018 which is a multivariate variable.

2. Materials and Method

2.1. Materials

The data used in this study are secondary data which is the data of accreditation results and data on the results of computer-based national exams (UNBK) for SMA / MA in Indonesia. Accreditation data was obtained from BAN-S/M while UN results was obtained from the Ministry of Education and Development (Balitbang) of the Ministry of Education and Culture. The data used were 6,771 high schools in 2018. Accreditation data consists of 8 indicators with accreditation years 2017 and 2018. The UN score results consist of 3 indicators used in the study which are average scores for 3 main test subjects that is Indonesian, English and Mathematics. The following is the description of the variables used in this study:

Table 1: List of variables used

Variable	Description
Y1-Y3 (respon variable)	The average scores for 3 main test subjects (Indonesian, English and Mathematics)
X1	standard of content
X2	standard of process
X3	standard of competency
X4	standard of educator and staff
X5	standard of facilities and infrastructures
X6	standard of management
X7	standard of financing
X8	standard of assessment

2.2. Method

The steps of data analysis carried out in this study are as follows:

- Pre-processing data
  - Perform data cleaning and merging of the data obtained.
- Data exploration
  - Exploring data with descriptive statistics

- See the correlation between the variables to be analyzed.
- Perform k-fold cross validation techniques on data
- Data divided into 5 group (k = 5) and than 4 data group were obtained as training data and 1 data group as testing data.
- Modeling data with the percentage of training data cut-off by 1% to 15%.
- Modeling the training data group according to the method used (MRF)
- Applying the MRF algorithm to model 8 scores of SNP toward the average value of the UN results with the following stages:
- Random sample collection of observations with returns from observational data sets. This stage is called bootsraping.
- The classification tree formation is based on the bootstrap method in step 1. The tree construction is carried out by applying random feature selection to each selection process. For each node, the optimal node splitting feature is selected from a set of m features that are picked randomly from the total M features (m <M).
- Splitting data groups with a series of binary splitting until the child node is generated. Each splitter depends only on the value of a predictor variable [2]. For continuous predictor variables, the binary questions are all questions in the form of "is $x \leq z$?", with $z \in \eta_p$ and z are the intermediate values between the two observed values of the x variable in sequence. Every observation on η_p that answers "yes" is sent to node η_L, while those who answer "no" are sent to node η_R. So if x has n different values, there will be n-1 splitting.
- Choosing the best splitting node. At any node η_p we aim to select a feature j_s from a random set of *m* features and a threshold *z* to partition the node into two child nodes η_L (left node) and η_R (right node). The partition that maximizes the node cost for all possible partitions is selected for node η_p [5].
- Steps 1 to 4 are repeated *k* times to form a group of trees or forest. The response of observation is predicted by aggregating the predicted results from *k* trees. MRF prediction results are based on the average output of a set of *k* trees formed.
- Test the goodness of the model using the average RMSEP value using testing data.
- Comparing the results of MRF method with RF regression method.
- Analyze the variable importance.
- Make conclusions.

3. Results and Discussion

3.1. Data exploration

Accreditation data consists of 8 indicators with accreditation years of 2017 and 2018. The data used were 6,771 high schools in 2018 consisting of 1894 (SMAN), 215 (MAN), 2359 (SMAS) and 2303 (MAS). The overall percentage of schools accredited A is 35.2%, accredited B is 42.4%, accredited C is 20% and not accredited (TT) is 2.5%.

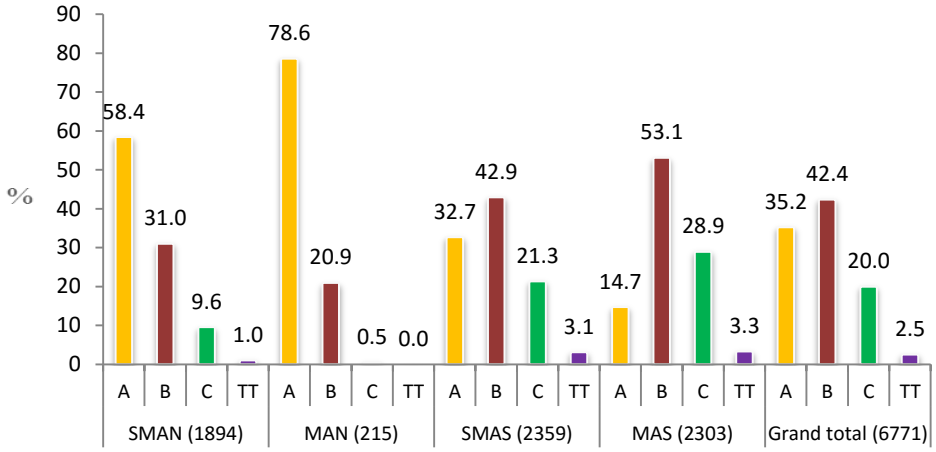

Figure 1: percentage of accreditation status based on type of school

The percentage of school accreditation status by type of school can be seen in Figure 1. The percentage of SMAN and MAN tends to get accreditation A which is 58.4% from 1894 schools and 78.6% of 215 schools. SMAS and MAS tend to get accreditation B with a percentage of 42.9% from 2359 schools and 53.1% of 2303 schools. Figure 2 shows the average UNBK based on accreditation status. Schools with accreditation status A have the highest average UNBK scores in all fields of study when compared to other accreditation status. The figure also shows that there is a relationship between accreditation status and UNBK values which can be seen from the decline in the average UNBK followed by a decrease in accreditation status.

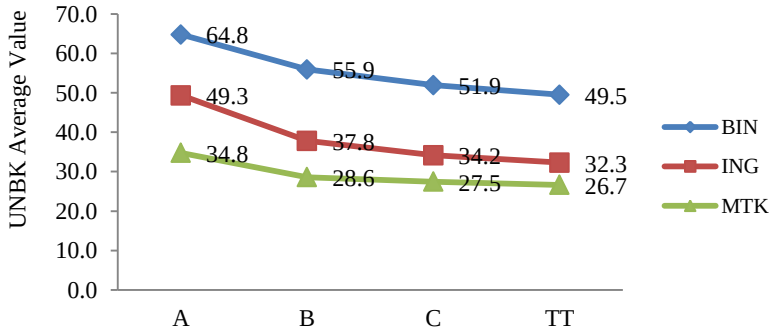

Figure 2: UNBK score is based on accreditation status

Table 2 explains the correlation values between the UNBK scores with score of eight SNPs based on 2017 and 2018 accreditation results. Table 2 shows the correlation value of UNBK scores with score of eight SNP have a positive correlation. It can be said that the greater the SNP value, the UNBK value will also be greater. Correlation values between the average values of the 3 main test subjects namely Indonesian, English, and Mathematics are quite high. The correlation between Indonesian and English is 0.81, Indonesian with Math is 0.67 and English with Math is 0.81. Correlation between the average values of the 3 main subjects show the correlation value with the direction of the positive correlation.

Table 2: SNP score correlation matrix with UNBK

	BIN	ING	MATH	SI	SPR	SKL	SPT	SSP	SPL	SB
ING	0.81									
MATH	0.67	0.81								
SI	0.38	0.34	0.22							
SPR	0.39	0.37	0.25	0.86						
SKL	0.42	0.40	0.27	0.80	0.85					
SPT	0.40	0.39	0.30	0.66	0.74	0.72				
SSP	0.50	0.49	0.36	0.67	0.74	0.74	0.81			
SPL	0.39	0.37	0.26	0.80	0.83	0.81	0.73	0.75		
SB	0.31	0.28	0.18	0.71	0.69	0.68	0.59	0.62	0.74	
SPN	0.36	0.35	0.23	0.82	0.83	0.80	0.65	0.68	0.82	0.70

3.2. Application of multivariate random forest

The implementation of multivariate random forest was analyzed using the "MultivariateRandomForest" package using the R program. The MultivariateRandomForest package has an algorithm where to do the modeling it is necessary to set the parameters first. Some modeling parameter settings are:

- Number of single trees built as many as 100 trees. Sutton states that the number of trees ≥ 100 tends to produce low levels of misclassification [14].
- The number of predictor variables used as splitting variable is 3 predictor variables. Intake of 3 predictor variables was based on the default regression tree that the calculation of the number of predictor variables used as splitting variable is obtained from the formula M / 3 so that a number of 3 predictor variables is obtained.
- The minimum number of samples at the leaf node is 5.
- Use 5-fold cross validation.

Model was analyzed by determining the training data cut-off for each fold. The cut-off for training data used starts from 1% to 15% of the total 6771 data. Modeling using several cut-offs is done due to the large amount of data while the algorithm used requires iteration that is long enough so that it requires a long duration to run the program. The determination of several cut-offs is also done to see how sensitive the performance of the MRF method is in classifying the quality of education in order to obtain an optimal evaluation value. Table 3 shows the results of calculating the accuracy of prediction by calculating the average Root Mean Square Error of Prediction (RMSEP). The average value of RMSEP is obtained from the average of the five fold RMSEP. The MRF model evaluation results showed that the smallest average RMSEP was obtained in the model with a 15% cut off of training data which was 8,427. The predicted RMSEP results in Table 3 also show that the bigger the percentage of training data to used, the smaller the average RMSEP value obtained.

Table 3: The results of model evaluation using multivariate random forest (N = 6771)

Training data (%)	n	RMSEP	Training data (%)	n	RMSEP	Training data (%)	N	RMSEP
1	68	8.727	6	406	8.518	11	745	8.437
2	135	8.677	7	474	8.503	12	813	8.464
3	203	8.693	8	542	8.499	13	880	8.443
4	271	8.617	9	609	8.484	14	948	8.416
5	339	8.545	10	677	8.450	15	1016	8.427

The average difference of RMSEP can be seen in Figure 3. Figure 3 shows that the results of the prediction began to stabilize in the training data with a 5% cut off. The results of the stable analysis can be seen from the decrease in the average value of RMSEP which is no longer significant in the training data cut-off from 5% to 6% and so on.

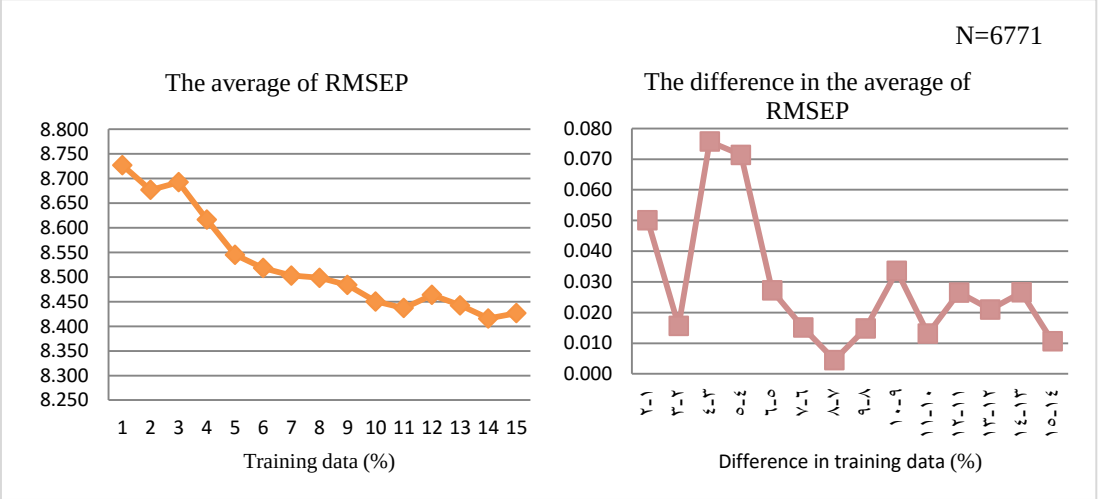

Figure 3: The average of RMSEP and the difference in the average of RMSEP

3.3. The comparison of MRF to RF regression

MRF is used as a method for analyzing data in this study, one of the reasons is because there is a strong correlation between the three response variables. The researcher then tries to compare the MRF analysis with RF regression in order to find out whether the MRF method is the right method to use based on the existing data conditions. RF regression method is applied to the data by estimating each response variable separately so that the RMSEP value of each variable is obtained which is then averaged.

Table 4: Model evaluation results of MRF and RF regression

Data latih (%)	RMSEP (MRF)	RMSEP (RFregression)
5	8.545	8.568

Table 4 presents the estimated results of the average RMSEP using the RF regression and MRF methods. The training data used are training data with a 5% cut-off refer to Figure 3 where the data starts to stabilize at the 5% cut-off training data. The analysis results obtained are that the accuracy of the model produced by the MRF method is better when compared to using the RF regression method. The MRF method is considered better seen from the results of the smaller average RMSEP value of 8,545.

3.4. Importance of Predictor Variables

Modeling using the MRF method is able to produce information about important variables used in building models. Scores of the importance variables are obtained from the number of times the variable is used as splitting node variable in building the model. The more often a variable is used as a splitting node variable, the higher the importance of the variable in constructing the model. Table 5 is a table of importance of 8 SNPs in classifying education quality based on the results of the analysis of the MRF and RF regression methods.

Table 5: Level of importance 8 SNP

MRF			RFregression					
			BIN		ING		MATH	
Standard	frequency	Rank	frequency	Rank	frequency	Rank	frequency	Rank
SKL	1757	1	1825	1	1785	1	1793	1
SI	1666	2	1768	2	1784	2	1776	2
SPL	1593	3	1537	3	1575	3	1555	3
SSP	1503	4	1452	4	1499	4	1444	4
SPR	1315	5	1404	5	1370	5	1376	5
SPT	1264	6	1256	6	1268	6	1272	6
SB	1072	7	1057	7	1070	7	1086	7
SPN	1039	8	1003	8	990	8	1022	8

The analysis of importance of 8 SNPs variables using MRF method was found that SKL variable had the highest chosen frequency level in classifying education quality that was equal to 1757. Variable with the second highest level of importance was SI with a variable importance score 1666 and the third was SPL with a variable importance score 1593 and then followed by 5 other variables. The importance level analysis of 8 SNP variables using RF regression method also gave the same result where the highest rank 3 was obtained by SKL, SI and SNP. Referring to the previous research conducted by Ramadhan on the modeling of the random forest classification to identify important factors in improving the quality of education, it was found that the 3 highest ranks of variable importance were occupied by the SSP, SPT and SKL [11]. The highest rank obtained is

different because, in addition to the different analysis methods, the variables used are also different. The fetaure variable used in the study was the score of 129 items of SNP. The response variable used was the average of the three subjects tested (Mathematics, Indonesian and English) so that they become one response variable which is then categorized. In line with the research conducted by Ramadhan, even though they don't have the exact same rank, in this study SKL was ranked as the top 3 most important variables out of the 8 SNP variables tested. The overall analysis of the importance variables shows that SKL always places the first rank as the variable that has the highest level of importance. SKL in the concept of interaction between SNP and UN is one of the references in developing curriculum where the output of SNP is the national exam. SKL is considered right if it is said to be a variable that has an important contribution in fulfilling the quality of education because SKL or graduate competency standards are related to the qualifications of graduates' abilities in high school education institutions which include the attitudes, knowledge, and skills of their graduates. Variable with the second highest level of importance in this study is SI and the third highest level of importance is SPL. SKL and SI are used as a reference for curriculum development in the interaction of 8 SNPs and UN, while the SPL is part of the standard for supporting curriculum implementation.

4. Conclussion

The conclusion that can be drawn in this study is that the MRF method is able to identify a stable model even though it only uses training data with a 5% cut off. MRF method is able to give better results for multivariate response variables when compared to using the RF regression method. The results of the analysis of importance variables from 8 national education standard variables toward national exam score variables, obtained 3 standards with the highest level of importance variables namely graduate competency standards, content standards and management standards. The three national education standards are the three most important standards in improving the quality of education of high school students in Indonesia.

5. Recommendation

Recommendations that can be given relating to the analysis that has been done is that for the next research can develop existing classification and regression tree methods especially MRF by developing a programming algorithm that is able to accommodate all data and then compared with the results of the analysis in this study.

References

[1] Berk RA. Statistical Learning from a Regression Perspective. New York (US): Springer Science + Business Media. 2008.

[2] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. New York (US): Chapman and Hall. 1984.

[3] Breiman L. Random Forests. Machine Learning 45(1), 5–32. 2001.

[4] De’ath G. Multivariate regression trees: a new technique for modeling species–environment relationships.

Ecology 83:1105–1117. 2002.

[5] Haider S, Rahman R, Ghosh G, Pal R. A Copula Based Approach for Design of Multivariate [6] Random Forests for Drug Sensitivity Prediction. PLoS One. 2015.

[7] Kementerian Pendidikan dan Kebudayaan. Indikator Mutu dalam Penjaminan Mutu Pendidikan Dasar dan Menengah. Jakarta (ID): Kemendikbud. 2017.

[8] Kementerian Pendidikan Nasional dan Kementrian Agama. Sistem Penjaminan Mutu Pendidikan : Panduan Teknis Evaluasi Diri Sekolah. 2010.

[9] Larose DT. Discovering knowledge in data : an introduction to data mining, Jhon Wiley & Sons Inc. 2005.

[10] Miller K, Huettman F, Norcross B, Lorenz M. Multivariate Random Forest Models of Estuarine-Associated Fish and Invertebrate Communities. Marine Ecology Progress Series Vol. 500; 159-174. 2014.

[11] Ramadhan A. Pemodelan Klasifikasi Random Forest Untuk Mengidentifikasi Faktor Penting Dalam Meningkatkan Mutu Pendidikan [tesis]. Bogor (ID): Institut Pertanian Bogor. 2019.

[12] Segal M, Xiao Y. Multivariate random forest. WIREs Data Mining and Knowledge Discovery Vol. 1: 80-87. 2011.

[13] Setiawan IA, Susetyo B, Fitrianto A. Application of Generalized Structural Component Analysis to Identify Relation between Accreditation and National Assessment. International Journal of Scientific Research in Science, Engineering and Technology Vol. 4(10):93–97. 2018.

[14] Sutton CD. Classification and regression trees, Bagging, and Boosting. Handbook of statistics Vol. 24:303-329. Elsevier. 2005.

[15] Wahyuni R. Evaluasi Hubungan antara Akreditasi dan Ujian Nasional Dengan Fuzzy Clusterwise Generalized Structured Component Analysis [tesis]. Bogor (ID): Institut Pertanian Bogor. 2019.