Modeling of Quality of Education in Junior High Schoolusing Multivariate Adaptive Regression Splines (MARS) Method

Urwawuska Ladini[1], Budi Susetyo[2], Indahwati[3]
{ladinidini@gmail.com[1], buset008@yahoo.com[2],indah.stk@gmail.com}

Student of Department of Statistics, IPB University, Indonesia[1]
Lecturer of Department of Statistics, IPB University, Indonesia[23]

Abstract. The accreditation of education system in Indonesia was established based on national education standards, while the national examinations were conducted to measure student's academic achievement. The relationship between accreditation and national examinations is still under debate, considering that both are important in measuring education quality. Multivariate adaptive regression spline is the implementation of regression problem solving to predict the variable response that continuously based on several independent variables arranged in a set of several basic function coefficients, which overall controlled in the regression data. This research conducted to model national examinations based on national education standard scores at the junior high school level and can accommodate interaction of independent variables. Facilities and infrastructure standards is the most important, which is 100% has a large influence on the goodness of the model, and then graduate competency standards with an importance of 45.34%. Process and management standards didn't significantly influence national examinations.

Keywords: accreditation, generalized cross validation, multivariate adaptive regression spline, national examination, national education standards.

1 Introduction

Education quality system expected managed evenly in all regions of Indonesia. The quality of primary and secondary education is the level of compatibility between the provision of primary and secondary education with the national education standard (SNP) in schools [1]. One of the government programs in education quality assurance is conducting accreditation carried out by BAN-S/M in assessing the feasibility of education unit program that refers to the SNP [2]. SNP is a minimum criterion of the education system in Indonesia. SNP consists of content standards (SI), process standards (SPR), and graduate competency standards (SKL), educator standards and education personnel (SPT), facilities and infrastructure standards (SSP), management standards (SPL), financing standards (SB) and education assessment standards (SPN) [3]. Besides, to measure and monitor the achievement of SNPs in education units, evaluation through the national examination (UN) conducted. UN is an assessment of learning outcomes to measure graduate competency achievement nationally in certain subjects

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290480

[4]. UN results can describe the academic achievements of students according to the SNP. Two types of UN that have been conducted in the past few years until now are the paper-based national exam (UNKP) and the computer-based national exam (UNBK). UNBK implemented as an effort to improve the efficiency and credibility of the UN.

Several studies have been conducted related to accreditation and the UN, including the application of structural equation modeling (SEM) to see SNP relations. The results of the study using SEM-GSCA concluded that at the senior high school level, SKL is influenced by other SNP variables, both directly and indirectly. The direct influence on SKL is the variable SB, SPL, and SPN, while SPT, SI, SPR, and SSP have an indirect effect on SKL through SPN. 46.7% of the diversity of all variables explained in the model [5]. The results of the study using SEM-GSCA concluded that at the junior high school level, SNP affected academic achievement, namely SKL, SPR, and SPN [6]. Researched using PLS-PM at the vocational school level illustrates that SKL influenced directly by the SPN and SPR variables. The R^2 value of structural models except SI and SPL is more than 0.75, which indicates that this model is very good, with the highest R^2 value in the model, namely the SPR variable of 86.3% that explained by the SSP, SPT and SI variables [7].

To see the pattern of the relationship between response variables and independent variables, it usually done using regression analysis [8] [9]. The regression curve approach that is often used is the parametric regression approach, where the pattern of the relationship between the response variable and the independent variable is known whether it is linear, quadratic, etc. However, when the pattern of the relationship between the response variable and the independent variable is not known and there is no past information about the shape of the relationship between the data patterns, so nonparametric regression is used. In nonparametric regression, the data left to search for itself regarding its data pattern, thus providing high flexibility for the nonparametric regression model. One of nonparametric regression method is the spline regression. The spline first developed by Whittaker in 1921 and developed its use in a regression analysis by Schoenberg in 1942 which resolved nonparametric regression problems [10]. The spline is a piecewise order q polynomial and has a derivative that is continuous with knots to order (q-1) [11]. At q = 0, 1, 2, and 3 respectively, the basis of truncated power is constant, linear, quadratic, and cubic [12]. In addition, there is also a method of recursive partitioning regression (RPR). RPR is a nonparametric method, which can be seen as a stepwise regression procedure. However, RPR has the disadvantage of not being continuous on knots. However, the base function has similarities to the base function of truncated power in order q=0. Therefore, multivariate adaptive regression spline (MARS) is a result of Friedman's modification of the RPR algorithm to overcome the shortcomings of RPR [11].

MARS, introduced by Friedman in 1991, is the implementation of regression problem solving to predict the variable response value that continuously based on several independent variables arranged in a set of several basic function coefficients, which overall controlled in the regression data [13]. MARS is useful for overcoming high dimensional data problems and producing accurate response variable predictions, as well as overcoming the disadvantages of recursive partitioning (RPR) which is producing a continuous model on knots, which based on minimum generalized cross-validation [14].

Some studies that apply the MARS method include research by [15] which identified optimal quantitative methods for mapping vulnerability of landslides in the city of Mizunami, Japan using logistic regression (LR), bivariate statistical analysis (BS), and MARS models. As a result, the method produced a better predictive rate (79%) when compared with LR (75%) and BS (77%) [15]. The research by [16] conducted to develop nonparametric regression

using MARS in continuous response data, which resulted in modeling the percentage of poor people and HDI giving a better level of accuracy compared to the estimation results obtained by the parametric regression method [16]. Based on the description above, the researcher will apply the MARS method to model UNBK based on the accreditation score of junior high school level.

2 Materials and Methods

The data used in this study are secondary data, which is the score of eight SNPs on accreditation as an independent variables, and average variable of UNBK scores as a response variable at the level of junior high school in 2017 and 2018. Accreditation data obtained from Ministry of Education's Research and Development Agency, while UNBK obtained from Center for Educational Research. The nine variables used in the study have numerical data types with 0-100 interval scales.

The stages of data analysis used are as follows.

1. Exploring data descriptively using plots and diagrams
2. Test the linearity relationship using the Ramsey reset test.
   Nonparametric modeling emphasized because this method moves on a model that does not meet the linearity assumption. One test to determine the relationship of linearity is the regression of the specification error test (reset test) which was first introduced by [17]. If the test results show p-value $<\alpha$, it can be said that the independent variable relationship model and response variable are not linear.
3. Partitioning the data into training data and test data with a proportion of 80%: 20%.
4. Conduct MARS analysis.
   a. MARS modeling is performed on training data using a combination of base functions (BF) 2-4 times independent variables, maximum interactions (MI) = 1, 2, and 3, and minimum observations between knots (MO) = 0, 1, 2, and 3 assuming> 3 will increase GCV. In this case, a combination of 36 models was be obtained.
   b. Choose the best model among 36 models based on the smallest GCV value.
   c. Determine the equation of the UNBK model based on the best model.
   d. Interpret the model to find out the SNPs that affect UNBK.
   e. Identify the importance of each SNP that influences UNBK.
   f. Calculates the RMSEP and R2-adj values for model validation in predicting test data.
5. Make conclusions.

3 Results and Discussion

3.1 Data Exploration

Data used as many as 6434 junior high school levels throughout Indonesia, which implemented the UNBK system consisting of 1928 of state junior high school (SMPN), 2542 of private junior high school (SMPS), 231 of state madrasa (MTSN), and 1733 of private madrasa (MTSS).

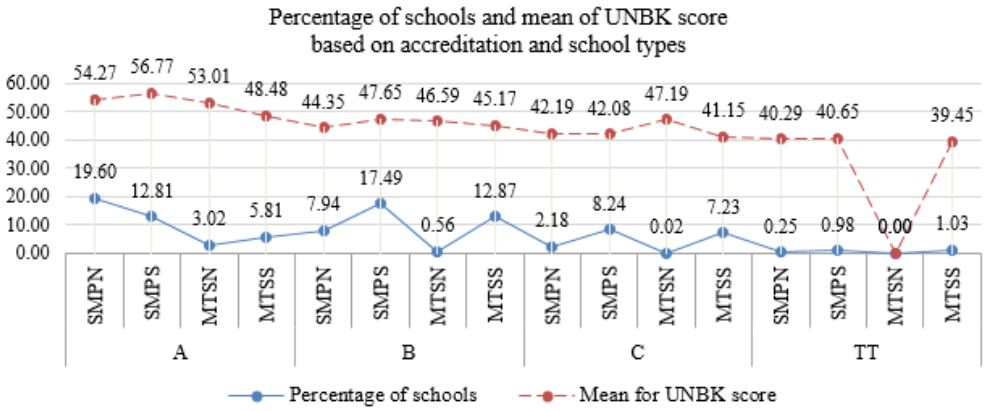

Fig. 1. Percentage of schools and mean of UNBK score based on accreditation and school types

Fig. 1.shows that from 6434 schools, the highest percentage of schools is 19.60% in SMPN that A-accredited with a mean of UNBK score of 54.27. The highest mean of UNBK score is 56.77 in SMPS that A-accredited. There is no MTSN that TT-accredited who also held UNBK.

The correlation value of the scores of eight SNPs with UNBK in **Table 1.** shows a positive correlation value but not too large. **Fig. 2.** shows the form of correlation between the eight SNPs and UNBK.

Table 1. Correlation value between SNP scores and UNBK

	UNBK
SI	0.347
SPR	0.367
SKL	0.398
SPT	0.388
SSP	0.477
SPL	0.370
SB	0.300
SPN	0.356

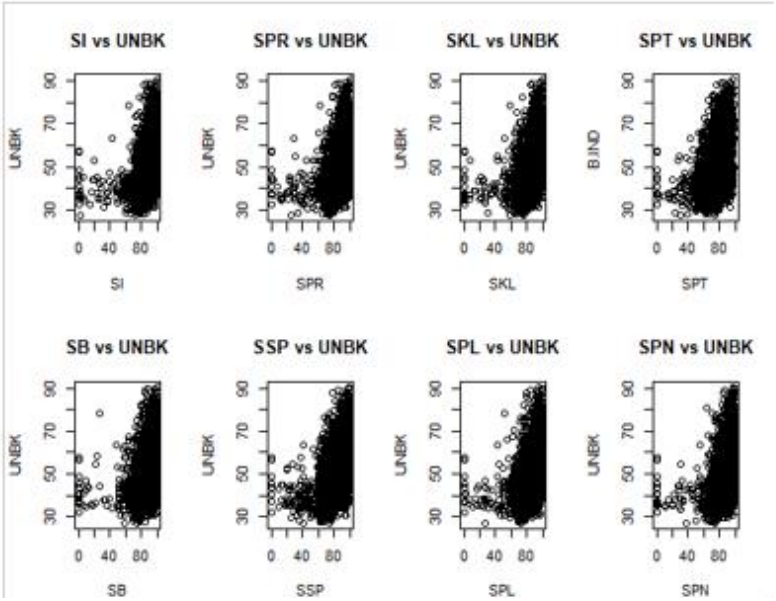

Fig. 2. SNP relationship plot with UNBK

Based on **Fig. 2.**, in general, the eight SNPs positively correlated with UNBK, but the correlation does not have a clear pattern, or the tendency to plot data has a change in behavior at a certain interval. This means that not always the greater the SNP value, the greater the UNBK score.

The limited information regarding the form of the pattern of relationships between variables is a consideration for the use of nonparametric regression approaches, one of which is the MARS method in data modeling. This is reinforced from the results of the specification error test on the linear model, namely the Ramsey reset test as in **Table 2.**, where p-value <0.05 means that there is a specification error in the linear model so that the pattern of the relationship between SNP and UNBK is not linear.

Table 2. The Ramsey reset test on SNP and UNBK data

Relationship	*F*	*df*1	*df*2	*p-value*	Conclusion
SI, SPR, SKL, SPT, SSP, SPL, SB, and SPN towards UNBK	48.009	16	6409	< 2.2e-16	Nonlinear

3.2 MARS Analysis

The stages of MARS modeling begin by finding optimal basic functions based on minimum GCV values using a combination of BF, MI, and MO in **Table 3.** Based on this process, UNBK modeling obtained 16 optimal base functions that require three interactions between variables in the model with GCV values of 72.3224with the test criteria for each response as in **Table 4.**

Table 3. GCV for MARS modelling

Model	BF	MI	MO	GCV	Model	BF	MI	MO	GCV
1	16	1	0	72.8993	19	24	2	2	72.3963
2	16	1	1	72.8993	20	24	2	3	72.3963
3	16	1	2	72.8993	21	24	3	0	72.3224
4	16	1	3	72.8993	22	24	3	1	72.3224
5	16	2	0	72.3963	23	24	3	2	72.3224
6	16	2	1	72.3963	24	24	3	3	72.3224
7	16	2	2	72.3963	25	32	1	0	72.8993
8	16	2	3	72.3963	26	32	1	1	72.8993
9	16	3	0	72.3224	27	32	1	2	72.8993
10	16	3	1	72.3224	28	32	1	3	72.8993
11	16	3	2	72.3224	29	32	2	0	72.3963
12	16	3	3	72.3224	30	32	2	1	72.3963
13	24	1	0	72.8993	31	32	2	2	72.3963
14	24	1	1	72.8993	32	32	2	3	72.3963
15	24	1	2	72.8993	33	32	3	0	72.3224
16	24	1	3	72.8993	34	32	3	1	72.3224
17	24	2	0	72.3963	35	32	3	2	72.3224
18	24	2	1	72.3963	36	32	3	3	72.3224

Table 4. Goodness of the train and test data model using MARS

Criteria	Model (BF=16; MI=3; MO=0)	
	Data training	Data testing
R^2-adj	0.3396	0.3441
RMSE	8.4406	8.2109

The SNP model equation for each response is as follows.

UNBK = 45.883 + 0.283 BF1 - 0.125 BF2 + 0.333 BF3 + 0.126 BF4 + 0.058 BF5 + 0.237 BF6 - 0.009 BF7 - 1.155 BF8 - 0.169 BF9 + 0.004 BF10 + 0.187 BF11 + 0.0004 BF12 + 0.070 BF13 + 0.332 BF14 - 0.002 BF15

which:

Base Function	Information
BF_1 = H(SSP-86)	SSP > 86
BF_2 = H(86-SSP)	SSP < 86
BF_3 = H(SKL-76)	SKL > 76
BF_4 = H(76-SKL)	SKL < 76
BF_5 = H(SKL-76)* H(SSP-90)	SKL > 76; SSP > 90
BF_6 = H(SKL-76)* H(SPT-97)	SKL > 76; SPT > 97
BF_7 = H(SKL-76)* H(97-SPT)	SKL > 76; SPT < 97
BF_8 = H(SI-95)	SI > 95
BF_9 = H(95-SI)	SI < 95
BF_{10} = H(SKL-76)* H(SSP-90) * H(SPN-95)	SKL > 76; SSP > 90; SPN > 95
BF_{11} = H(70-SB)	SB < 70
BF_{12} = H(97-SI)* H(SKL-76)* H(97-SPT)	SI < 97; SKL > 76; SPT < 97
BF_{13} = H(SI-95)* H(SPN-82)	SI > 95; SPN > 82
BF_{14} = H(SI-95)* H(82-SPN)	SI > 95; SPN < 82
BF_{15} = H(SKL-76)* H(SSP-90)* H(SB-84)	SKL > 76; SSP > 90; SB < 84

The interpretation of the above model, for example in the first base function (BF1) explained that this base function would contribute to the model of 0.283 without involving interaction

with other independent variables. If the SSP value > 86, and the other SNP values were fixed, it would increase the UNBK mean value for the school by 0.283. Besides, the base function contains interactions between independent variables, namely in BF14 which contains the interaction of two independent variables. In this BF14, it explained that if the SI value > 95, SPN < 82, and other SNP values remained, it would increase the UNBK mean value in the school by 0.332. The above equation shows that eight SNPs have an effect on UNBK.

Table 5. shows the ranking of importance of variables in modeling UNBK, where the SSP is the most important which has a large influence on the goodness of the model. The next relatively important is SKL, which is 45.34%, etc. SPR and SPL no effect on UNBK.

Table 5. SNP interest rates for UNBK data

Rating	Variable	Interest (%)
1	SSP	100.00
2	SKL	45.34
3	SPT	22.22
4	SB	17.86
5	SI	15.41
6	SPN	12.70
7	SPR	0.00
8	SPL	0.00

3.3 Conclusion

Based on the explanation, it can be concluded that the MARS method is effective in modelling data, which a pattern of relationships between several independent variables and the response variable is not linear. This method can also accommodate interaction of independent variables with up to three levels. In the case of this study, it can be concluded that the three most important components that influence the achievement of learning outcomes in the education system are SSP, SKL, and SPT, so that these components are expected to be maximized in improving the quality of education.

References

[1]Direktorat Jenderal Pendidikan Dasar dan Menengah: Indikator Mutu dalam Penjaminan Mutu Pendidikan Dasar dan Menengah. Direktorat Jenderal Pendidikan Dasar dan Menengah Kementerian Pendidikan dan Kebudayaan, ID (2017)

[2]Peraturan Menteri Pendidikan Nasional Republik Indonesia: Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 29 Tahun 2005 Tentang Badan Akreditasi Nasional Sekolah/Madrasah. Menteri Pendidikan Nasional, ID (2006)

[3] Badan Akreditasi Nasional Sekolah/Madrasah: Perangkat Akreditasi SMP/MTs. Badan Akreditasi Nasional Sekolah/ Madrasah, ID (2017)

[4] Peraturan Pemerintah: Peraturan Pemerintah Republik Indonesia Nomor 13 Tahun 2015 tentang Perubahan Kedua atas Peraturan Pemerintah Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan. ID (2015)

[5] Vita D, Susetyo B, Indriyanto B: Generalized Structured Component Analysis (GSCA) For National Education Standard (NES) Of Secondary School In Indonesia. Global Journal of Pure and Aplied Mathematics. Vol. 11(4), pp. 2441-2449 (2015)

[6] Setiawan IA, Susetyo B, Fitrianto A: Application of Generalized Structural Component Analysis to Identify Relation between Accreditation and National Assessment. Int. J. Sci. Res. Sci. Eng. Technol. Vol. 4(10), pp.93–97 (2018)

[7] Hijrah M, Susetyo B, Sartono B: Structural Equation Modeling Of National Standard Education Of Vocational High School Using Partial Least Square Path Modeling. Int. J. Sci. Res. Sci. Eng. Technol. Vol. 4(4), pp. 1418-1422 (2018)
[8] Walpole RE, Myers RH, Myers SL, Ye K: Probability & Statistics for Engineers & Scientists Eighth Edition. Pearson Education Inc, NJ (2007)
[9] Montgomery DC, Peck EA, Vining GG: Introduction to Linear Regression Analysis Fifth Edition. John Wiley & Sons Inc, NJ (2012)
[10] Budiantara IN: Pemodelan Regresi Nonparametrik dan Semiparametrik Spline (Konsep, Metode, dan Aplikasinya). Pros. Semin. Nas. Matematika. ISSN: 2406-9869, pp. 1-16 (2014)
[11] Friedman J: Multivariate adaptive regression splines (with discussion). Ann. Stat. Vol. 19(1), pp.1-67 (1991)
[12] Wu H, Zhang JT: Nonparametric Regression Methods for Longitudinal Data Analysis. John Wiley & Sons Inc, NJ (2006)
[13] Otok BW: Pemilihan Model Terbaik pada Mars Respon Kontinu. Statistika Vol. 8(1), pp. 19–29 (2008).
[14] Hastie T, Tibshirani R, Friedman J: The Elements Of Statistical Learning Data Mining, Inference, And Prediction. Springer Series in Statistics, CA (2008)
[15] Wang LJ, Guo M, Sawada K, Lin J, Zhang J: Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. Catena, Vol.135, pp. 271-282 (2015)
[16] Ampulembang AP: Pengembangan Model Regresi Nonparametrik Birespon Kontinu menggunakan Metode MARS. Institut Teknologi Sepuluh November, ID (2017)
[17] Gujarati DN: Basic Econometrics Fourth Edition. McGraw-Hill, US (2004)