Jurnal Matematika MANTIK
Vol. 7, No. 1, May 2021, pp. 51-58

ISSN: 2527-3159 (*print*) 2527-3167 (*online*)

Multivariate Adaptive Regression Splines (MARS) for Modeling The Student Status at Universitas Terbuka

Siti Hadijah Hasanah

Department of Statistics, Universitas Terbuka, Indonesia

Article history:
Received Dec 23, 2020
Revised May 10, 2021
Accepted May 30, 2021

Kata Kunci:
Fungsi basis, GCV, Multivariate, Recursive, Splines

Abstrak. Multivariate Adaptive Regression Splines (MARS) digunakan untuk memodelkan status mahasiswa aktif program studi Statistika Universitas Terbuka serta untuk mengetahui faktor-faktor yang mempengaruhi variabel respon. Penelitian ini terdiri dari 9 variabel yaitu jenis kelamin, umur, pendidikan, status pernikahan, pekerjaan, tahun registrasi awal, jumlah registrasi, SKS, dan IPK, tetapi setelah dilakukan pemodelan menggunakan metode MARS maka variabel penjelas yang dapat mempengaruhi variabel respon adalah tahun registrasi awal, jumlah registrasi, IPK, dan SKS. Berdasarkan hasil output *R* dan menggunakan selang kepercayaan 95%, maka masing-masing fungsi basis 1 sampai 10 adalah signifikan secara parsial dengan nilai-p dari fungsi basis 1-10 lebih kecil dari 0,05 dan secara simultan dengan nilai nilai-p lebih kecil dari 0,05, sehingga model di atas memiliki pengaruh yang signifikan secara parsial maupun simultan terhadap variabel respon. Dari hasil tersebut maka disimpulkan bahwa model MARS layak digunakan untuk menentukan faktor-faktor yang mempengaruhi status aktif siswa.

Keywords:
Basis function, GCV, Multivariate, Recursive, Splines

Abstract. Multivariate Adaptive Regression Splines (MARS) used to model the active student's status in the Department of Statistics at Universitas Terbuka and determine the factors that influence the response variable. This study consists of 9 variables, namely gender, age, education, marital status, job, initial registration year, number of registrations, credits, and GPA, but after modeling using the MARS method, the explanatory variable can affect the response variable is the initial registration year. Several registrations, GPA, and credits. Based on the results of the R output and using a 95% confidence interval, each base 1 to 10 function is partially significant with the p-value of the base 1-10 function being smaller than 0.05 and simultaneously with a smaller p-value. of 0.05, so that the above model has a significant effect partially or simultaneously on the response variable. From these results, it is concluded that the MARS model is suitable for determining the factors that affect the active status of students.

How to cite:
S. H. Hasanah, *"Multivariate Adaptive Regression Splines (MARS) For Modeling The Student Status at Universitas Terbuka", J. Mat. Mantik*, vol. 7, no. 1, *pp. 51-58*, May 2021.

CONTACT:
Siti Hadijah Hasanah *sitihadijah@ecampus.ut.ac.id* Statistics Department, Faculty of Science and Technology, Universitas Terbuka, Tangerang Selatan 15437, Indonesia

The article can be accessed here. *https://doi.org/10.15642/mantik.2021.7.1.51-58*

1. Introduction

Universitas Terbuka has 39 service offices spread from Sabang to Merauke. The advantages of Universitas Terbuka are that there are no restrictions on the period of completion of the study, no implementation of the dropout system, no restrictions on the year of graduation or age, and registration time throughout the year [1]. Based on the advantages and ease of studying at Universitas Terbuka there is a problem that must be faced one of them is the active status of UT students, students who have been registered at Universitas Terbuka are given the facility to register at any time so many students with inactive status in a semester and do not know when status as an active student again. This study was conducted by analyzing the factors that affect the status of student active, especially in the Department of Statistics based on several indicators, namely age, gender, education, marital status, employment status, year of initial registration, number of registrations, credits, and GPA.

Several statistical methods used to determine the effect of explanatory variables on categorical response variables include the Binary Logistic Regression method and the Multivariate Adaptive Regression Spline (MARS) [2]. Both methods are included in regression analysis, which is a statistical method that studies the mathematical relationship pattern between one or more explanatory variables and the response variable [3]. Logistic regression has the advantage that it does not need to fulfill assumptions [4], can be used on non-linear data, and is easy to interpret. However, logistic regression has a weakness, namely that there is multicollinearity between the explanatory variables which is one of the problems that make the parameter estimation results unstable [5]. The MARS method is able to accommodate interactions between variables even for high-dimensional data [6], [7].

MARS is a relatively flexible classification method used to determine the pattern of relationships between explanatory variables and response variables without using initial assumptions about the form of functional relationships. This method is a complex combination of spline and recursive partitioning and involves a high data dimension, namely the number of observations and a large number of variables [8]. In addition, MARS can effectively explore the hidden non-linear relationship between response variables and predictor variables as well as the interaction effects on complex data structures [9]. MARS can solve the problem of high and non-continuous dimensions in nodes. This method can analyze 50-1000 amounts of data with 3-20 explanatory variables [10]. Dynamic non-linear patterns and interactions can be explained by this method [11]. MARS can also be applied in credit assessment [12], [13], transportation [9], and software engineering [14]. So the method we use to determine the active status of Universitas Terbuka students in this study is Multivariate Adaptive Regression Splines (MARS). This research is an alternative study program to determine the best step in overcoming the many inactive student statuses, especially in the Statistics Study Program.

2. Multivariate Adaptive Regression Splines (MARS)

Suppose y a single response variable depends on n predictor variables x, where $x' = (x_1, x_2, \dots, x_n)$ then the regression model is :

$$y = f(x_1, x_2, \dots, x_n) + \varepsilon \tag{1}$$

Assume f is a linear combination of the base functions of $B_m(x)$, with the base of each subregion $m = 1,2, \dots, M$.

$$f(x) = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{2}$$

$a_0, a_1, a_2, \dots, a_M$ = regression coefficient
$B_m(x)$ = m - base function
M = maximum base function

Each base function is a truncated power splines function. Univariate truncated power basis can be described as an indicator function. The functions of the base of univariate splines base from right and left are as follows :

$$b_q^+(x,c) = [+(x-c)]_+^q, \;\; b_q^-(x,c) = [-(x-c)]_+^q$$

A single base function can be specified:

$$B_m(x) = \prod_{h=1}^{H_m} \left[s_{hm}(x_{i(h,m)} - c_{hm})\right]_+^q \tag{3}$$

H_m = the number of interactions on the M based function
s_{hm} = value ±1, if the knot is located to the right or left of the subregion
$x_{i(h,m)}$ = predictor variables (1,2, …, n), h-interactions (1,2, …, M) and base of each m subregion (1,2,…, M)
c_{hm} = knot point position
q = order of splines

The MARS model can be stated as follows :

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{h=1}^{H_m} \left[s_{hm}(x_{i(h,m)} - c_{hm})\right] \tag{4}$$

Location selection and the number of knots on MARS using forward stepwise and backward stepwise steps:

1. Forward Stepwise

   This stage is to determine the location of the knot and the maximum base function based on the data by minimizing the average sum of square residual (ASR) [15]. The addition of the base function is continued until it reaches the maximum base function.

2. Backward Stepwise

   This stage is to determine the size of the appropriate base function. At this stage, the removal of the base function that contributes to the estimated value of the small response until a balance between bias and variety and a suitable model is obtained, that is, by minimizing the value of generalized cross-validation (GCV) [16].

   The minimum GCV functions are as follows :

$$GCV(M) = \frac{ASR}{\left[1 - \frac{C(\widehat{M})}{n}\right]^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_M(x_i)\right]^2}{\left[1 - \frac{C(\widehat{M})}{n}\right]^2} \tag{5}$$

y_i = response variable
$\hat{f}_M(x_i)$ = the value of the response variable estimates on the M base function
N = many observations
$C(\widehat{M})$ = $C(M) + dM$
$C(M)$ = $Trace\,[B(B_M^T B_M)^{-1} B_M^T] + 1$
d = value when each base function achieves optimization ($2 \leq d \leq 4$)

2.1 Decomposition ANOVA

Interpretation of the MARS model through decomposition ANOVA [17] :

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \left[s_{1m}\left(x_{i(1,m)} - c_{1m}\right)\right] + \sum_{m=1}^{M} a_m \left[s_{1m}\left(x_{i(1,m)} - c_{1m}\right)\right]\left[s_{2m}\left(x_{i(2,m)} - c_{2m}\right)\right] + \sum_{m=1}^{M} a_m \left[s_{1m}\left(x_{i(1,m)} - c_{1m}\right)\right]\left[s_{2m}\left(x_{i(2,m)} - c_{2m}\right)\right]\left[s_{3m}\left(x_{i(3,m)} - c_{3m}\right)\right] + \cdots \quad (6)$$

Can be written as follows: In general, it can be written as follows:

$$\hat{f}(x) = a_0 + \sum_{Hm=1} f_i(x_i) + \sum_{Hm=2} f_{ij}(x_i, x_j) + \sum_{Hm=3} f_{ijk}(x_i, x_j, x_k) + \cdots \quad (7)$$

2.2 Estimation of MARS Model

Parameters Suppose the base function $B_m(x),\ m = 0,1,2,\ldots,M$, to estimate the regression coefficient a_m using the smallest square method.

$$\hat{\alpha}_{MKT} = (B^T B)^{-1} B^T Y, \quad (8)$$

$$a = (a_0, a_1, \ldots, a_M)^T \quad (9)$$

$$Y = (y_1, y_2, \ldots, y_n)^T \quad (10)$$

$$B = \begin{pmatrix} 1 & \prod_{h=1}^{H_1} s_{1m}\left(x_{1(1,m)} - c_{1m}\right) \ldots \prod_{h=1}^{H_M} s_{Mm}\left(x_{1(M,m)} - c_{Mm}\right) \\ 1 & \prod_{h=1}^{H_1} s_{1m}\left(x_{2(1,m)} - c_{1m}\right) \ldots \prod_{h=1}^{H_M} s_{Mm}\left(x_{2(M,m)} - c_{Mm}\right) \\ & \ldots \\ & \ldots \\ 1 & \prod_{h=1}^{H_1} s_{1m}\left(x_{n(1,m)} - c_{1m}\right) \ldots \prod_{h=1}^{H_M} s_{Mm}\left(x_{n(M,m)} - c_{Mm}\right) \end{pmatrix} \quad (11)$$

2.3 MARS Model Significance Test

The significance test of the MARS Model is on the base function which includes simultaneous and partial tests, Simultaneous tests with the following hypotheses:

H_0: $a_1 = a_2 = \cdots = a_m = 0$
H_1: there is at least one $a_m \neq 0; m = 1,2,\ldots,M$
F test :

$$F = \frac{SSR}{SSE} \sim F_{(M,N-M-1)} \quad (12)$$

Partial test with the following hypotheses:

$$H_0: a_m = 0$$

$$H_1: a_m \neq 0\ ; m = 1,2,\ldots,M$$

t test :

$$t = \frac{\hat{a}_m}{s_e(\hat{a}_m)} \sim t_{(\frac{a}{2},N-M-1)} \quad (13)$$

3. Data and Methods

The data used in this study is secondary data, which is the data of the student characteristics of the Department of Statistics, Universitas Terbuka from year 2009.1 to 2019.2. The data is divided into two parts, namely training data 1046 and testing data 447 which consists of 9 explanatory variables and one response variable :

Table 1. Characteristics Students at Department of Statistics, Universitas Terbuka

Variable	Information	Scale	Category
X1	Gender	Nominal	1 = Man 2 = Woman
X2	Age	Interval	
X3	Education	Ordinal	1 = Senior High School 2 = Associate Degree 3 = Bachelor Degree 4 = Master Degree 5 = Doctoral Degree
X4	Marital Status	Nominal	1 = Single 2 = Married
X5	Job	Nominal	1 = Unemployment 2 = Private Employees 3 = Entrepreneur 4 = Civil Servants 5 = Army/Police
X6	Initial Registration Year	Interval	
X7	Number of Registrations	Interval	
X8	Credits	Interval	
X9	GPA	Interval	
Y	Student Status	Nominal	0 = Not Active 1 = Active

The steps for applying the MARS method are as follows:

1. Determining the number of base functions
   Base function limit between 2 - 4 times the number of explanatory variables.
2. Determination of maximum interaction (MI)
   The MI used are 1, 2, and 3.
3. Determination of minimum observations in each knot (MO)
4. Perform a trial and error process
   Combine base, MI, and MO functions to get minimum GCV value.
5. Test the significance of the regression coefficient
   Partial test statistics with *t* and *F* test statistics to test the significance of the regression coefficient simultaneously.

4. Result

By using the R software and simulating the number of basis functions, MI, MO, and GCV, the MARS model is obtained as follows:

$$Y = -0{,}003BF_1 - 0{,}082BF_2 + 0{,}029BF_3 - 0{,}027BF_4 - 0{,}233BF_5 - 0{,}272BF_6 + 0{,}254BF_7 - 0{,}004BF_8 - 0{,}008BF_9 + 0{,}006BF_{10} \quad (14)$$

$BF_1 = h(20152 - Year\_Early\_Registration)$
$BF_2 = h(10 - Total\_Registration)$
$BF_3 = h(Year\_Registration\_Early - 20121)$
$BF_4 = h(Year\_Early\_Registration - 20142)$
$BF_5 = h(GPA - 0{,}33)$
$BF_6 = h(0{,}33 - GPA)$
$BF_7 = h(GPA - 1{,}63)$
$BF_8 = h(Credits - 21)$
$BF_9 = h(21 - Credits)$
$BF_{10} = h(Credits - 70)$

Table 2. MARS Model Interpretation

Basis Function (BF)	Interpretation
1	Base 1 function contributes to the model of -0.003, when students register early in 2015 semester 2 then the student status is active.
2	Base 2 function contributes to the model by -0.082, if the student register 10 times then the student status is active.
3	Base 3 function contributes to the model of 0.029, when the student registers early in 2012 semester 1 then the student status is active.
4	Base 4 function contributes to the model of -0.027, if the student registers early in 2014 semester 2 then the student status is active.
5	Base 5 function contributes to the model by -0.233, if the student has a GPA of 0.33 then the student status is active.
6	Base 6 function contributes to the model by -0.272, if the student has a GPA of 0.33 then the student status is active.
7	Base 7 function contributes to the 0,254 models. if the student has a GPA of 1.63 then the student status is active.
8	The base 8 functions contribute to the model of -0.004, when a student takes 21 credits then the student status is active.
9	The base 9 functions contribute to the model of -0.008, when a student takes 21 credits then the student status is active.
10	The base 10 function contributes to the model of 0.006, when a student takes 70 credits then the student status is active.

The significance test is then performed on each of the above Basis functions as follows:

Table 3. Significance Test

Basis Function	T	P-value	F	P-value
BF_1	-2,252	0,024537	297,7	< 2,2e-16
BF_2	-16,277	< 2e-16		
BF_3	12,425	< 2e-16		
BF_4	-10,58	< 2e-16		
BF_5	-7,577	7,86e-14		
BF_6	-2,327	0,02014		
BF_7	4,945	8,90e-07		
BF_8	-4,684	3,18e-06		
BF_9	-3,873	0,000114		
BF_{10}	4,976	7,60e-07		

Previously, based on research data, there were 9 explanatory variables used to determine the effect of student activeness status, but after modeling using the MARS

method. These explanatory variables could affect the active status of students majoring in Statistics were initial registration year, number of registrations, GPA, and credits. Based on the output of the R software in table 3 and using a 95% confidence interval, each of the base 1 to 10 functions is partially significant (t-statistic) with the p-value of the base function 1-10 being smaller than 0,05 and automatically simultaneous (f-statistic) with p-value less than 0,05 so that the above model has a partially and simultaneously significant influence on the response variable. From these results, it is concluded that the MARS model is suitable for determining the factors.

5. Conclusions

There are nine explanatory variables used to determine the effect of student active status, including gender, age, education, marital status, job, initial registration year, number of registrations, credits, GPA, but after modeling using the MARS method, the explanatory variables can affect Statistics student active status is the initial registration year, number of registrations, GPA, and credits. Based on the results of the R application output and using a 95% confidence interval, each of the base 1 to 10 functions is partially significant (t-statistic) with the p-value of the base function 1-10 less than 0,05 and simultaneously (f-statistic) with a p-value less than 0,05, so that the above model has a significant effect partially or simultaneously on the response variable. From these results, it is concluded that the MARS model is suitable for determining the factors that affect student active status. The results of this study are as a way to find solutions for study programs in overcoming the many inactive student statuses, especially in the Statistics Study Program, namely by paying attention to these four factors, the study program will provide treatment in further research so that students can graduate on time so that can reduce the number of inactive students in each semester.

References

[1] S. Hasanah and S. Permatasari, "METODE KLASIFIKASI JARINGAN SYARAF TIRUAN BACKPROPAGATION PADA MAHASISWA STATISTIKA UNIVERSITAS TERBUKA," vol. 14, no. 2, pp. 243–252, 2020, doi: 10.30598/barekengvol14iss2pp249-258.

[2] A. Wibowo and M. R. Ridha, "Comparison of Logistic Regression Model and MARS Using Multicollinearity Data Simulation," *JTAM | J. Teor. dan Apl. Mat.*, vol. 4, no. 1, p. 39, 2020, doi: 10.31764/jtam.v4i1.1801.

[3] S. Park, S. Y. Hamm, H. T. Jeon, and J. Kim, "Evaluation of logistic regression and multivariate adaptive regression spline models for groundwater potential mapping using R and GIS," *Sustain.*, vol. 9, no. 7, pp. 1–20, 2017, doi: 10.3390/su9071157.

[4] U. Wagschal, *Regression analysis*, no. March 2014. 2016.

[5] H. Midi, S. K. Sarkar, and S. Rana, "Collinearity diagnostics of binary logistic regression model," *J. Interdiscip. Math.*, vol. 13, no. 3, pp. 253–267, 2010, doi: 10.1080/09720502.2010.10700699.

[6] C. Mina and E. Barrios, "Profiling Poverty with Multivariate Adaptive Regression Splines," *Development*, no. September, 2009, [Online]. Available: http://publication.pids.gov.ph/pubdetails.phtml?code=DP 2009-29.

[7] M. Rosenblatt, "Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to The Annals of Statistics. ® www.jstor.org," *Ann. Stat.*, vol. 19, no. 3, pp. 1403–1433, 1991.

[8] G. W. Weber, I. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya-Özkurt, "CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization," *Inverse Probl. Sci. Eng.*, vol. 20, no. 3, pp. 371–400, 2012, doi: 10.1080/17415977.2011.624770.

[9] L. Y. Chang, "Analysis of bilateral air passenger flows: A non-parametric multivariate adaptive regression spline approach," *J. Air Transp. Manag.*, vol. 34, pp. 123–130, 2014, doi: 10.1016/j.jairtraman.2013.09.003.

[10] T. Harju, "Derivation of Aircraft Performance Parameters Applying Machine Learning Principles," 2017, [Online]. Available: www.aalto.fi.

[11] M. Y. Cheng and M. T. Cao, "Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines," *Appl. Soft Comput. J.*, vol. 22, pp. 178–188, 2014, doi: 10.1016/j.asoc.2014.05.015.

[12] T. S. Lee, C. C. Chiu, Y. C. Chou, and C. J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," *Comput. Stat. Data Anal.*, vol. 50, no. 4, pp. 1113–1130, 2006, doi: 10.1016/j.csda.2004.11.006.

[13] S. Hasanah, "Islamic Countries Society of Statistical Sciences," in *Comparison Of Method Classification Artificial Neural Network Back Propagation, Logistic Regression, And Multivariate Adaptive Regression Splines (Mars) (Case Study Data Of Unsecured Loan)*, 2014, pp. 477–486, [Online]. Available: www.isoss.net.

[14] Y. Zhou and H. Leung, "Predicting object-oriented software maintainability using multivariate adaptive regression splines," *J. Syst. Softw.*, vol. 80, no. 8, pp. 1349–1361, 2007, doi: 10.1016/j.jss.2006.10.049.

[15] R. Biswas, B. Rai, P. Samui, and S. S. Roy, "Estimating concrete compressive strength using MARS, LSSVM and GP," *Eng. J.*, vol. 24, no. 2, pp. 41–52, 2020, doi: 10.4186/ej.2020.24.2.41.

[16] S. Sekulic and B. R. Kowalski, "MARS : A TUTORIAL," vol. 6, no. April, pp. 199–216, 1992.

[17] E. Kartal Koc and H. Bozdogan, "Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function," *Mach. Learn.*, vol. 101, no. 1–3, pp. 35–58, 2015, doi: 10.1007/s10994-014-5440-5.