Bootstrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) to Analyse the Lecturer Research Performance in Private University

Maylita Hasyim*, Dwi Shinta Rahayu, Novita Eka Muliawati, Diesty Hayuhantika, Ratih Puspasari, Dewi Anggreini, Ratri Candra Hastari, Setyo Hartanto and Fajar Hendro Utomo

Department of Mathematics Education, STKIP PGRI Tulungagung, East Java, Indonesia

* maylita@stkippgritulungagung.ac.id

Abstract. This study explores the use of a fairly simple nonparametric regression algorithm known as multivariate adaptive regression splines (MARS) which has the ability to approximate the relationship between the predictor and response variables. The main advantages of MARS are its capacity to produce simple, easy-to-interpret models, its ability to estimate the contributions of the predictor variables, and its computational efficiency. In this research, MARS is combined with one of nonparametric approach bootstrap aggregating (bagging). Bagging is used to improve the classification accuracy of the MARS method. This study is aimed to analyse lecturer research performance in private university using Bagging MARS algorithm. In modelling bagging MARS for lecturer research performance in a private university that there are three dominant influencing factors: 1) amount of research with cost an internal college granted by internal college with interest level of 86.20%; 2) number of publications of research results in national journals with interest level of 69.83%; and 3) number of speakers within national scientific meeting/ seminar with interest level of 63.34%. The accuracy of the MARS model classification is 84.615% with the classification error rate (APER) of 15.385%.

1. Introduction
Multivariate Adaptive Regression Splines (MARS) is one of nonparametric and semi parametric regression that takes into accounts the covariates with a multivariate approach [1]. Nonparametric regression using MARS method does not depend on the assumption of a certain curve shape, has flexibility in high dimensional data and modelling involves a lot of interaction with a few variables [2]. MARS is a hybrid of nonparametric methods, like recursive partitioning regression and additive model. The data are left to reveal the variable knot locations, while the users need not input any specification into the model [3]. The basis functions in MARS (which serve as independent variables) are truncated linear functions, which address the problem of discontinuity of recursive partitioning algorithms. In contrast to additive models, MARS allows interactions up to an order specified by the user and trades off the interaction order and complexity of the additive functions and interactions [4]. Certainly, MARS is a highly commended predictive modelling technique in almost all applications [5].
Currently MARS is often combined with a nonparametric approach to modelling flexibility. In this research, MARS is combined with one of nonparametric approach bootstrap aggregating (bagging). Bagging method is used as a tool to form a more stable classifier. Bagging predictor is a method to generate multiple versions of predictors and use them for aggregate predictors [6]. In this case, bagging is used to improve the classification accuracy of the MARS method. Therefore, this study is expected to obtain better modelling and classification functions through bagging MARS method.

The main objective of this research is to analyse lecturer research performance in a private university (case study at STKIP PGRI Tulungagung) using Bagging MARS method. Research performance is one component in the determination of the lecturer performance index or Indeks Kinerja Dosen (IKD). In a previous study by [7], IKD is modelled by using Survival MARS, where the response variables studied were time intervals. Lecturers are professional educators and scientists with the main task of transforming, developing and disseminating science, technology and the arts through education, research and dedication to society [8,9]. Lecturer performance is defined as the ability to carry out the work or tasks that lecturers have in completing a job [10]. Performance can be interpreted as work presentation, job performance, work achievement, or work result [11]. It also refers to the result or output of a process. Education is aimed to: (1) improve the performance, capability and output of education; (2) facilitate communication and information exchange about the best practice of education with various types of educational institutions; and (3) be a tool to improve institutional performance education and guidelines in strategic planning.

Through the use of the MARS Bagging approach, it is expected to obtain the dominant factors that influence the lecturer’s research performance more accurately. The classification produced by the MARS bagging model is expected to be optimal. The results of this study can be used as input for the leadership of private universities for policy making improvement of lecturer performance quality.

2. Literature Review
2.1. Multivariate Adaptive Regression Splines (MARS)
In 1991, Jerome H. Friedman introduced the Multivariate Adaptive Regression Spline (MARS) method as a new method that automates the development of accurate predictive models for continuous and binary response variables. MARS is one of the flexible methods for modelling high-dimensional regression data. MARS is a form of extension of the Basis Spline Functions where the number of basis function is the parameters of the model [4]. Some terms that need to be considered in MARS is as follows [5],

a. Knots, the point of a regression line to form a region of a regression function,
b. Basis Function (BF), collection of some functions that are used to describe the relationship between the response variable and the predictor,
c. Interaction, a correlation between variables and the maximum number of interaction (MI) 1, 2, and 3.

MARS model is formulated as follows:

\[
f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} S_{km}(x_{v(k,m)} - t_{km}) + \epsilon
\]

where,
\[
\alpha_0 = \text{main of basis function}
\alpha_m = \text{coefficient of basis function } -m
M = \text{maximum of basis function (nonconstant basis fungsi)}
K_m = \text{degree of interaction}
s_{km} = \text{its value 1 or -1 if the data is to the right of the knot point or left of the knot point}
x_{v(k,m)} = \text{predictor variable}
t_{km} = \text{knot point of predictor variable } x_{v(k,m)}
The best model produces smallest GCV which is denoted as:

\[ GCV(m) = \frac{ASR}{1 - \frac{C(M)}{n}} = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - f_M(x_i) \right]^2 \]

(2)

where,

\[ C(M) = \text{trace}(B(B^T B)^{-1} B^T) + 1 \]

\[ y \quad = \text{response variable} \]

\[ n \quad = \text{amount of observation} \]

To calculate the accuracy of classification, Apparent Error Rate (APER) is used. The APER value represents the proportion representation of the wrong sample classified by the classification function [21]

\[ APER(\%) = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \]

(3)

2.2. **Bootstrap Aggregating (Bagging)**

The first bagging method used by [18] as a tool to form a more stable classifier. Bagging predictor is a method to generate multiple versions of predictors and use them for aggregate predictors. Multiple versions are formed by a bootstrap replication of a data set. In some cases, bagging on real data sets can improve accuracy. If the changes in the data sets lead to significant changes then bagging can improve accuracy. The basic idea of bagging is to use bootstrap resampling to generate predictors with many versions, which when the combination should be better than a single predictor built to solve the same problem.

Determination of the amount of replication \( B \) is very varied, because the size of \( B \) can give different results at each stage of analysis. Based on the References [13] it recommends 25 or fifty times replications, but [14] suggests that increased accuracy will occur if the number of replications is increased from fifty to one hundred times, whereas if the number is increased to more than a hundred times it will produce no greater accuracy of the replication accuracy of a hundred times. Meanwhile, References [15] recommend small values of \( B \), such as 25 times.

3. **Research Methodology**

This research used a quantitative approach because the process of this research is deductive. Response variable of this study is lecturer research performance, that measured for both lecturer as civil servant (DPK) and lecturer with permanent status (DTY). Categorical response variable has goal to be suitable with MARS binary response. The predictor variable in this study is a research activity consisting of eleven indicators. The sample used are 35 lecturers at STKIP PGRI Tulungagung. Technique of collecting data in this research is survey with questionnaire instrument.

Steps of analyses in this study are as follows:
a. Establishing a MARS model for initial data sets: (1) specifying BF; (2) determining MI; (3) determine MO between knots;
b. Getting the best MARS model for a single dataset based on the smallest GCV value;
c. Getting a significant variable influencing from the best MARS model for a single dataset;
d. Conducting bagging of the response variable pairs and significant predictor variables of the best MARS model for single data sets with fifteen, twenty, 25, and thirty bootstrap replications;
e. Modelling MARS of each sampling of B bootstrap replication with the amount of Basis Function (BF), Maximum Interaction (MI) and Minimum Observation (MO) among knots equal to the sum of BF, MI and MO among the knots on the best MARS model for a single data set;
f. Obtaining a misclassified level value on each sampling of bootstrap B replication;
g. Obtaining the value of the baggage misclassification level of the mean misclassification on each sampling up to B.

4. Analysis and Discussion

Determining the best MARS model is based on the minimum GCV value obtained by trial and error in combining BF, MI, and MO criteria to get the best model. Formation of MARS model by combining: 1) BF with a value of two to four times the number of predictor variables which are 44, 66 and 88; 2) MI with values of 1, 2, and 3; and 3) MO with values 0, 1, 2 and 3. Determination of MARS model uses MARS 2.0 software, by entering combination of BF, MI and MO.

The MARS modelling obtained from the best combination of BF, MI and MO was selected based on minimum GCV values. When GCV is the same, the next consideration is to choose the model with the largest R² value, if it turns out that some of the models have the same R², then the consideration switches to the greatest classification (total correct). From the overall MARS model that has been obtained based on the minimum GCV value, the best model of MARS selected is the model with the value of BF = 66, MI = 3, and MO = 2 and the GCV value of 0.018; the value of R² is 93.2%; and the value of classification accuracy (total correct) of 88.524%.

The best MARS model is:

\[
\hat{f}(x) = 51,203 + 0.173BF_5 + 0.041BF_{13} + 0.003BF_{19} + 0.046BF_{21} + 0.049BF_{35} + 0.351BF_{43} + 0.206BF_{54} + 0.002BF_{62}
\]

The best MARS model in Equation (4) is shaped by seven variables that significantly contribute to the model. The importance of each variable to the model is shown in Table 2.

| Variable | Importance | -GCV |
|----------|------------|------|
| X₃       | 86.20%     | 0.016|
| X₈       | 69.83%     | 0.011|
| X₆       | 63.34%     | 0.008|

Table 2 shows the importance of predictor variables in the grouping function, which is estimated by the increase of GCV values due to the removal of the variables considered from the model. In modelling bagging MARS for lecturer research performance in private university that there are three dominant influencing factors: 1) amount of research with cost by internal college grant with interest
level of 86.20%; 2) number of publications of research results in national journals with interest level of 69.83%; and 3) number of speakers within national scientific meeting/seminar with interest level of 63.34%.

The next step is to evaluate the classification accuracy of the MARS model. Evaluate the classification of private high school quality by calculating the value of APER obtained through the following classification Table 3.

| Actual class | Estimation class | Total of actual |
|--------------|-----------------|-----------------|
|              | 1 (DPK)         | 2 (DTY)         |                  |
| 1 (DPK)      | 12              | 3               | 15              |
| 2 (DTY)      | 3               | 17              | 20              |
| Total of prediction | 15  | 20  | 35  |

Based on Table 3 the precision of classification and APER value can be calculated as follows:

\[
\text{Classification accuracy} = \frac{12 + 17}{35} = \frac{29}{35} = 0.82857
\]

\[
APER = \frac{3 + 3}{35} = \frac{6}{35} = 0.17142
\]

(5)

Based on the above calculation, it can be seen that the accuracy of MARS model classification is 82.857% with the classification error rate (APER) of 17.142%.

Furthermore, a bagging approach is used to find out the accuracy of the classification of the MARS model. Bagging with thirty replications is proved to demonstrate the accuracy of MARS model classification with single datasets, comparison of classification results for each \(E^{(B)}\) with the result of MARS model classification of single dataset along with the decrease of classification error level shown in Table 4.

| \(E^{(B)}\)      | Mean of misclassification \(\mu\) | Mean of misclassification rate MARS model of single dataset |
|------------------|----------------------------------|-----------------------------------------------------------|
| \(E^{(25)}\)     | 0.21192                          | 0.24601                                                   |
| \(E^{(30)}\)     | 0.19006                          | 0.24601                                                   |
| \(E^{(35)}\)*    | 0.19723                          | 0.24601                                                   |
| \(E^{(45)}\)     | 0.24601                          | 0.24601                                                   |

Table 4 shows that \(E^{(25)}\), \(E^{(30)}\) dan \(E^{(35)}\) have not succeeded in providing better classification accuracy than the model classification MARS with single datasets. Whereas \(E^{(45)}\) can provide classification accuracy, by providing a convergent classification error rate. Based on the application of baggage algorithm, it can be concluded that bagging with 45 datasets replication successfully shows the same classification accuracy with MARS model with single dataset.
5. Conclusion

Based on the empirical results using Bagging MARS approach, the best model of MARS of lecturer research performance selected is the model with the value of BF = 66, MI = 3, and MO = 2 and the GCV value of 0.018; the value of $R^2$ is 93.2%; and the value of classification accuracy (total correct) of 88.524%. The obtained significant influencing predictors of the lecturer research performance are 1) amount of research with cost an internal college grant with interest level of 86.20%; 2) number of publications of research results in national journals with interest level of 69.83%; and 3) number of speakers within national scientific meeting/ seminar with interest level of 63.34%. The classification accuracy of MARS model is 82.857% with the classification error rate (APER) of 17.142%.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from research partner for any advice and idea. Special thanks to Indonesian ministry for research and higher education for the financial support under PEKERTI scheme.

References

[1] Friedman J H 1991 Estimating Function of Mixed Ordinal and Categorical Variables Using Multivariate Adaptive Regression Splines. Technical Report LCS (107) Statistics Department Stanford University
[2] Friedman J H 1991 Multivariate Adaptive Regression Spline, the Annals of Statistics (19: 1-141)
[3] De Veaux R D. A L Gordon, J C Comiso, and N E Bacherer 1993a Modeling of topographic effects on Antarctic Sea ice using Multivariate Adaptive Regression Splines Journal of Geophysical Research 98 (20): 307–320
[4] Frank I E 1995 Tutorial: modern nonlinear regression methods Chemometrics and Intelligent Laboratory Systems 27: 1–19
[5] Mina C D and Barrios E B 2010 Profiling Poverty with Multivariate Adaptive Regression Splines Philippine Journal of Development (69): 57-59.
[6] Breiman L 1994 Bagging Predictor Technical Report No. 421 Department of Statistics University of California.
[7] Hasyim M and Prastyo D D 2018 Modelling Lecturer Performance Index of Private University in Tulungagung by Using Survival Analysis with Multivariate Adaptive Regression Spline J. Phys.: Conf. Ser. 974 012065: 1-7
[8] UU No. 14 2005 about Teacher and Lecturer
[9] Peraturan Pemerintah No. 37 of 2009 about Lecturer
[10] Depdiknas 2005 Undang-Undang Republik Indonesia No 20 tahun 2003 about Sistem Pendidikan Nasional Jakarta: Departemen Pendidikan Nasional
[11] Lembaga Administrasi Negara 2000 Implementasi Good Governance Jakarta
[12] Johnson R A and Wichern D W 1992 Applied Multivariate Statistical Analysis Prentice Hall New Jersey
[13] Sutton C D 2005 Classification and Regression Trees, Bagging, and Boosting Handbook of Statistics Vol. 24: 303-329
[14] Hastie T, Tibshirani R and Friedman J 2001 The Elements of Statistical Learning: Data Mining, Inference and Prediction Springer-Verlag New York
[15] Efron B and Tibshirani R J 1993 An Introduction to the Bootstrap Chapman Hall New York