Sys Rev Pharm 2020;11(8):406-413
A multifaceted review journal in the field of pharmacy

BOOTSTRAP AGGREGATING MULTIVARIATE ADAPTIVE REGRESSION SPLINE FOR OBSERVATIONAL STUDIES IN DIABETES CASES

***Bambang W. Otok[1], Romy Y. Putra[1], Sutikno[1], and Septia D. P. Yasmirullah[1]**

[1]Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
***Corresponding Author's E-mail**: bambang_wo@statistika.its.ac.id , dr.otok.bw@gmail.com

ABSTRACT

Background: Diabetes is a serious chronic disease that occurs either when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Based on reports from the ministry of health, East Java Province has a Diabetes Melitus (DM) prevalence rate of 2.1%. This figure is greater than the prevalence rate in Indonesia, which is 1.5%.
Materials and Methods: This study using Bootstrap Aggregating (Bagging) Multivariate Adaptive Regression Spline (MARS) method to analyze observational studies in diabetes cases. This study is aimed to analyze the factors that influence the complications type II diabetes and compare the level of accuracy between MARS and Bagging MARS.
Results: The results showed that the probability of not occurring disease complications in patients is 0.708 and the occurrence of complications is 0.202. The variable that has the greatest influence is diabetes gymnastics. Type 2 DM patients with attending to diabetes gymnastics tend to not get disease complications as 1.857 times compared to patients with not attend to diabetes gymnastics.
Conclusion: The accuracy of the classification between the MARS method and bagging MARS with 50, 100, 150, and 200 replications obtained the same results. This shows that bagging MARS cannot always improve accuracy.

Keywords: Accuracy, Bagging MARS, Diabetes Melitus, and Replication

Correspondence:
Bambang W. Otok
[1]Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
***Corresponding Author's E-mail**: bambang_wo@statistika.its.ac.id , dr.otok.bw@gmail.com

INTRODUCTION

Diabetes is a serious chronic disease that occurs either when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. The prevalence of diabetes has been steadily increasing over the past few decades. The global prevalence (age-standardized) of diabetes has nearly doubled since 1980, rising from 4.7% to 8.5% in the adult population. Diabetes caused 1.5 million deaths in 2012. Forty-three percent of these 3.7 million deaths occur before the age of 70 years old [1].

In Indonesia, diabetes is one of the biggest causes of death with a percentage of 6.7%, after stroke (21.9%) and coronary heart disease (12.9%). If not addressed, this condition can cause a decrease in productivity, disability, and premature death [2]. One of the provinces in Indonesia which has a high prevalence of DM sufferers is East Java [3]. DM Sufferers in East Java reach 2.1%. This figure is greater than the prevalence rate in Indonesia, which is 1.5% [4]. Previous studies in cases of type 2 diabetes complications have been carried out by [5] and [6].

In this research, Bootstrap Aggregating Multivariate Adaptive Regression Spline is used to analyze observational studies in diabetes cases. Bootstrap aggregating (bagging) is one form of bootstrap. Bagging is a method that can be used in statistical classification and regression that can reduce variance. Bagging is designed to improve stability, improve classification accuracy and predictive power [7]. Previous research on bootstrap has been done by [8].

Multivariate Adaptive Regression Spline (MARS) is one of the nonparametric regression methods. MARS method can be used to solve the problem of regression and classification to predict the response variable that is continuous or binary. MARS is a combination of Recursive Partition Regression (RPR) with a truncated spline. The MARS method is useful for handling high dimension data with a sample size of 50-1000 samples and predictor variables 3-20 variables. MARS does not depend on assumption that function must be linear so as to produce accurate response variable prediction and be able to overcome the weaknesses of RPR and spline truncated by producing a continuous model on knots, which is based on the minimum Generalized Cross Validation (GCV) value [9]. Previous research on MARS has been done by [10].

Literature Review

Multivariate Adaptive Regression Spline

Multivariate Adaptive Regression Spline (MARS) introduced by Jeromy H. Friedman in 1991 [9]. This method is a complex combination of spline and recursive partitioning regression. The MARS method can be used in responses that are categorical and continuous [11]. MARS estimator according as follows.

$$\hat{f}(\mathbf{x}_i) = \hat{\gamma}_0 + \sum_{m=1}^{M} \hat{\gamma}_m \prod_{d=1}^{D_m} \left[s_{dm}(x_{v(d,m)i} - r_{dm}) \right]_+ \quad (1)$$

Where $\hat{\gamma}_0$ is the estimated parameter a constant basis function? $\hat{\gamma}_m$ are the estimated parameter of the m nonconstant basis functions? M is the number of nonconstant basis functions. D_m is maximum interaction on the basis function m. S_{dm} signs basis functions, value ±1.

$x_{v(d,m)i}$ is variable x to v, where v is an index of one of the variable x related to the d and the interaction m basis function in the MARS function. r_{dm} is the value of knots in the interaction and the m basis functions.

The best model for MARS is obtained by choosing optimal basis function by minimizing the value of GCV through stepwise procedure (forward and backward), GCV can be obtained by the following equation:

$$GCV(M) = \frac{MSE}{\left[1 - \frac{C(\tilde{M})}{n}\right]^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_M(\mathbf{x}_i)\right]^2}{\left[1 - \frac{C(\tilde{M})}{n}\right]^2} \quad (2)$$

with $C(\tilde{M}) = C(M) + d.M$, $C(M) = trace\left[\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\right] + 1$, and

$\hat{f}_M(\mathbf{x}_i) = \hat{\gamma}_0 + \sum_{m=1}^{M}\hat{\gamma}_m\prod_{d=1}^{D_m}\left[s_{dm}(x_{v(d,m)i} - r_{dm})\right]_+$ where n is the sample size. M is the number of basic functions in MARS.

$C(\tilde{M})$ is the complex function. d is the degree of interaction, y_i is the value of the i response variable. $\hat{f}_M(x_i)$ is the estimated value of the response variable on the basis function M.

The classification of the MARS model is based on regression analysis. If the classification on the response variable consists of two values, it is said to be a binary response regression. The probability model can be used with the following equation.

$$P(Y = 1 \mid X = \mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{f(x)}}{1 + e^{f(x)}} \quad (3)$$

and

$$P(Y = 0 \mid X = \mathbf{x}) = 1 - \pi(\mathbf{x}) = \frac{1}{1 + e^{f(x)}} \quad (4)$$

with $f(x) = logit\ \{\pi(x)\}$.

To find out the exact proportion of samples classified using TAR (Total Accuracy Rate).

(TABLE 1)

From Table 1, we get the formula to calculate the incorrect sample proportion is classified (APER) and the right sample proportion is classified (TAR).

$$APER(\%) = \frac{n_{10} + n_{01}}{n} \times 100\% \quad (5)$$

and

$$TAR(\%) = 1 - APER \quad (6)$$

Where n is the sample size or number of observations, n_{00} is the number of observations from y_0 the right classified as y_0, n_{11} is the number of observations from y_1 the right classified as y_1, n_{01} is the number of observations of y_0 which are incorrectly classified as y_1, n_{10} is the number of observations of y_1 which are incorrectly classified as y_0.

Bootstrap Aggregating

The bootstrap aggregating (bagging) was first introduced in 1996 by Breiman. Bagging can be used to reduce the variance of estimators in classification and regression. This technique can also improve stability, accuracy, and predictive power. The minimum number of replications for classification purposes is 50 times, and optimum when the highest accuracy value has been obtained. Bagging is used to correct unstable estimators or classifications, especially high dimensional problems [12].

The bagging algorithm is as follows:

1. Bootstrap sample constructs $£_i^* = (y_i^*, \boldsymbol{xi}^*)$, $i = 1,2,...,n$ according to the empirical distribution of pairs $£_i = (y_i, x_i)$, $i = 1,2,...,n$.
2. Calculate the bootstrap $\hat{\theta}_n^*(\boldsymbol{x}_i)$ estimator with the plug-in principle, namely $\hat{\theta}_n^*(x) = h_n(£_1^*,...,£_n^*)$.
3. The aggregate bootstrap estimator is $\hat{\theta}_{nB}^*(x) = E^*[(\boldsymbol{x}_i)]$.

Bagging MARS

The bagging algorithm in MARS modelling is as follows.

1. Take a bootstrap sample set of £ consisting of $\{(y_i, x_i), i = 1, 2,..., n\}$, and do bootstrap replication on the data, so that it gets $£_i^* = (y_i^*, x_i^*)$, $i = 1,2,...,n$ or called (£(B)).
2. Do MARS modelling on (£(B)).
3. Predict the response variable from the MARS model that has been produced.
4. Repeat step 1 through step 3 until the bootstrap replication.

5. Make predictions on the response variable based on the selection of predictions that often appear on each observation from bootstrap replication (majority vote).
6. Calculate the accuracy of the prediction classification of the bagging MARS model.

Illustration of bagging MARS can be seen in Figure 1.

Research Methodology

This research used secondary data from the Health Department of Pasuruan Regency. The research variables can be seen in Table 2.
(TABLE 2)

1. The analysis step used in this research are
2. Do descriptive statistics to determine the characteristics of DM type II patients.
3. Form the best MARS model for the initial dataset by combining the number BF = 14, 21, 28. MI = 1, 2, 3. MO = 0, 1, 2, 3, 5, 10.
4. Obtain the best MARS model for the initial data set based on the smallest GCV.
5. Form a data set from the best MARS model that will be used as data to analyze using bagging MARS.
6. Do bootstrap sampling with returns for 50, 100, 150 and 200.
7. Do MARS modelling on each sampling B bootstrap replication.
8. Determine the prediction of the response variable from the bagging MARS model based on maximum voting.
9. Get the value of the accuracy of the MARS classification and bagging MARS at each B bootstrap replication.
10. Comparing the classification accuracy between MARS and bagging MARS.

Results and Discussion

Characteristics of Type 2 Diabetes Mellitus Patients

Characteristics of type 2 Diabetes Mellitus (DM) patients and the risk factors that influence it can be seen from descriptive statistics. The following are the characteristics of type 2 DM patients and risk factors that influence them.
(TABLE 3)
Table 3 shows that the age range of type 2 DM patients is from the age of 29 to 77 years old, with the mean age of patients being 52.9 years old and variance 225.98. Besides, the duration of suffering from type 2 DM is to 7 years, with a mean duration of type 2 DM for 4.06 years and a variance of 4.017. Based on Figure 2, the majority of type 2 DM indicated no disease complications. This can be seen from the percentage of patients who do not have complications is 56.25% compared with the percentage of patients who have complications that equal to 43.75%. The majority of disease complications occur in patients who do not attend diabetes gymnastics.
(Fig. 2)
Characteristics of Patients Based on Disease Complication

Bagging MARS Analysis

The formation of the MARS model is done by trial and error, combining the maximum number of Basis Function (BF), Maximum Interaction (MI) and Minimum Observation (MO). The maximum number of basic functions is 2 to 4 times the number of predictor variables [10]. In this research, the predictor variables used were seven variables, so that the maximum number of basic functions used were 14, 21 and 28. The maximum interaction used were 1, 2 and 3. Table The minimum observations used are 0, 1, 2, 3, 5 and 10. The next step is to form the MARS model by combining Basis Function (BF = 14, 21, 28), Maximum Interactions (MI = 1, 2, 3), and Minimum Observations (MO = 0, 1, 2, 3, 5, 10). The following are the results of a combination of BF, MI, and MO for the selection of the best MARS models.
(TABLE 4)
From Table 4, The best MARS models are obtained in combination of BF = 28, MI = 3 and MO = 1 with GCV value = 0.063. The best model MARS is:

$$\begin{aligned}\hat{f}(x) = {} & 0.177 + 0.396BF_1 - 0.089BF_2 + 0.619BF_3 + 0.015BF_4 + 0.010BF_5 \\ & - 0.109BF_6 - 0.147BF_7 + 0.310BF_8 - 0.179BF_9 - 0.121BF_{10} \\ & - 0.033BF_{11} - 0.026BF_{12} + 0.113BF_{13} + 0.217BF - 0.478BF_{15} \\ & + 0.028BF_{16} + 0.028BF_{17} + 0.087BF_{18} + 0.066BF_{19}\end{aligned} \quad (7)$$

Where,

$$BF_1 = x_5 \qquad BF_6 = h(4 - x_4) \qquad BF_{11} = h(51 - x_1)(x_5) \qquad BF_{16} = h(51 - x_1)(x_5)(x_6)$$
$$BF_2 = x_6 \qquad BF_7 = h(x_4 - 4) \qquad BF_{12} = h(x_1 - 51)(x_5) \qquad BF_{17} = h(x_1 - 51)(x_5)(x_6)$$
$$BF_3 = x_7 \qquad BF_8 = (x_5)(x_6) \qquad BF_{13} = h(4 - x_4)(x_7) \qquad BF_{18} = h(5 - x_4)(x_5)(x_6)$$
$$BF_4 = h(58 - x_1) \qquad BF_9 = (x_5)(x_7) \qquad BF_{14} = h(x_4 - 4)(x_7) \qquad BF_{19} = h(x_4 - 5)(x_5)(x_6)$$
$$BF_5 = h(x_1 - 58) \qquad BF_{10} = (x_6)(x_7) \qquad BF_{15} = (x_5)(x_6)(x_7)$$

Interpretation of several basis functions of the equation (7).

1. 0.619 BF3 = 0.619 x_7
   Type 2 DM patients with attending to diabetes gymnastics tend to not get disease complications as exp (0.619) = 1.857 times compared to patients with not attend to diabetes gymnastics.
2. 0.310 BF8 = 0.310 (x_5) (x_6)
   Type 2 DM patients who are non-obesity and do not hypertension tend to not get disease complications as exp (0.310) = 1.364 times compared to patients who have obesity and hypertension.

3. 0.087 BF18 = 0.087 $h(5\text{-}x_4)$ (x_5) (x_6)
   Type 2 DM patients with less than the 5-year duration of suffering, non-obesity, and not hypertension tend to not get disease complications as exp (0.087) = 1.091 times compared with patients more than 5 years duration of suffering, have obesity, and hypertension.

The probability value of diabetics experiencing complications and not experiencing complications can be calculated using equation (3), equation (4) and the best MARS model, whit the following result.

$$\hat{f}(x) = 0.177 + 0.396(1) - 0.089(1) + 0.619(1) + 0.015(1) + 0.010(1) - 0.109(1) - 0.147(1) + 0.310(1) - 0.179(1) - 0.121(1) - 0.033(1) - 0.026(1) + 0.113(1) + 0.217(1) - 0.478(1) + 0.028(1) + 0.028(1) + 0.087(1) + 0.066(1) = 0,885$$

So that,

$$P(Y = 1 \mid X = \mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{0.885}}{1 + e^{0.885}} = 0.708 \quad (8)$$

and

$$P(Y = 0 \mid X = \mathbf{x}) = 1 - \pi(\mathbf{x}) = 1 - 0.708 = 0.292 \quad (9)$$

From equations (8) and (9) it is found that the probability of not occurring disease complications in patients is 0.708 and the occurrence of complications is 0.202.

The importance of predictor variables in the classification of the MARS model can be seen in Table 5.

(TABLE 5)

Table 5 shows that there are five most important variables in the model formation and influencing the disease complications of type 2 diabetes mellitus. The most important variable is diabetes gymnastics (X7) with the importance of 100 percent.

Prediction values between MARS and bagging MARS can be seen in Figure 3. Figure 3 shows that the predictive value for MARS and bagging MARS is different for each bootstrap replication. However, the predicted value produced is not much different from the MARS prediction value.

(Fig. 3)

A Comparison of the accuracy classification MARS method and bagging MARS can be seen in Table 6. Based on Table 6, the classification accuracy of the MARS method and bagging MARS produces the same value that is equal to 90.63 %. This shows that the bagging method cannot always improve the classification accuracy. Bagging does not always work well but has the potential to reduce the forecasting Mean squared error.

(TABLE 6)

Conclusions

The classification accuracy between the MARS model and bagging MARS results in the same classification accuracy that is equal to 90.63%. The best model of MARS in the health complication data of type 2 diabetes mellitus is a combination of BF = 28, MI = 3, and MO =1 with GCV values of 0.063. The probability of not occurring disease complications in patients is 0.708 and the occurrence of complications is 0.202. Based on the importance of predictor variable, the most important variable is the diabetes gymnastics (X7), using one interaction that has the most dominant influence type 2 DM patients with attending to diabetes gymnastics tend to not get disease complications as 1.857 times compared to patients with not attend to diabetes gymnastics. Using two interactions that have the most dominant influence in type 2 DM patients who are non-obesity and not hypertension tend to not get disease complications as 1.364 times compared to patient's obesity and have hypertension. Using three interactions that have the most dominant influence in type 2 DM patients with less than the 5-year duration of suffering, non-obesity, and not hypertension have a tendency to not get disease complications as 1.091 times compared with patients more than 5 years duration of suffering, have obesity, and hypertension.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper

Funding/Support

No external funding was received for this research.

REFERENCES

1. WHO, Global report on diabetes? 2016. Accessed 20 April 2019. http://www.who.int/diabetes/global-report
2. Depkes, Menkes: Mari Kita Cegah Diabetes dengan Cerdik, 2016. Accessed 20 March 2019. http://www.depkes.go.id/article/print/16040700002/menkes-mari-kita-cegah-diabetes-dengan-cerdik.html (2)
3. Kemenkes RI, Hasil Utama RISKESDAS 2018 (Kementerian Kesehatan RI, 2018). (3)
4. Kemenkes RI, Riset Kesehatan Dasar 2013 (Kementerian Kesehatan RI, 2013). (4)
5. S. Hasanah, B. W. Otok & Purhadi, "Propensity Score Matching Using Support Vector Machine in Case of Diabetes Mellitus (DM)," *In The 2nd International Conference on Biomedical Engineering* (IBIOMED), pp. 132-137, 2018.
6. B. W. Otok, A. Amalia, Purhadi and S. Andari, "Propensity Score Matching of the Gymnastics for Diabetes Mellitus Using Logistic Regression," *In International Conference and Workshop on Mathematical Analysis and its Application* (ICWOMAA), AIP Conference Proceedings, vol. 1913, no. 6, pp. 1-8, 2017.
7. L. Brieman, "Bagging Predictors," *Machine Learning*, vol. 26, no. 2, pp. 123-140,1996.

8. R, Pane, B. W. Otok, I. Zain, and I. N. Budiantara, "Bootstrap Inference Longitudinal Semiparametric Regression Model," *In Proceedings of The 7th SEAMS UGM International Conference on Mathematics and its Application*, AIP Conference Proceedings, vol. 1707, pp. 1-9, 2016.
9. J. H. Friedman, "Multivariate Adaptive regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1-67, 1991.
10. A. P. Ampulembang, B. W. Otok, A. T. Rumiati, and Budiasih, "Bi-Responses Nonparametric Regression Model Using MARS and Its Properties," *Applied Mathematical Sciences*, vol. 9, no. 29, pp. 1417-1427, 2015.
11. B. W. Otok, "Bootstrap Approach in the Multivariate Adaptive Regression Splines (MARS) Model," Dr. Dissertation, Gadjah Mada University, 2008.
12. P. Buhlmann and B. Yu, "Analyzing Bagging," *The Annals of Statistics*, vol. 30, no. 4, pp. 927-961, 2002.

TABLE 1. Binary Response Classification

Observation	Observation Predictions	
	y_0	y_1
y_0	n_{00}	n_{01}
y_1	n_{10}	n_{11}

TABLE 2. Research Variables

Variable	**Scala Data**	**Category**
Disease Complication (Y)	Nominal	0 = Complication of diabetes
		1 = Non-complication of diabetes
Age (X_1)	Ratio	-
Gender (X_2)	Nominal	0 = Women
		1 = Man
Genetic History (X_3)	Nominal	0 = Have a genetic history
		1 = No genetic history
Duration of Suffering (X_4)	Ratio	-
Obesity (X_5)	Nominal	0 = Obesity (if Body Mass Index (BMI) $\geq$ 27.5)
		1 = Non-obesity (if Body Mass Index (BMI) < 27.5)
Hypertension (X_6)		0 = Hypertension (if systolic blood pressure > 130 mmHg)
		1 = Non hypertension (if systolic blood pressure $\leq$ 130 mmHg)
Diabetes Gymnastics (X_7)	Nominal	0 = Not attend to diabetes gymnastics
		1 = Attend to diabetes gymnastics

TABLE 3 Characteristics of Continuous Type Variables

Variable	**Mean**	**Variance**	**Minimum**	**Maximum**
Age (X_1)	52.890	225.980	29	77
Duration of Suffering (X_4)	4.060	4.017	1	7

TABLE 4 Combining Basis Function (BF), Maximum Interaction (MI), and Minimum Observation (MO)

No	BF	MI	MO	GCV	R^2
1	14	1	0	0.110	0.555
2	14	1	1	0.110	0.555
3	14	1	2	0.110	0.555
4	14	1	3	0.110	0.555
5	14	1	5	0.110	0.555
6	14	1	10	0.110	0.555
7	14	2	0	0.093	0.621
8	14	2	1	0.093	0.621
9	14	2	2	0.093	0.621
10	14	2	3	0.093	0.621
11	14	2	5	0.093	0.621
12	14	2	10	0.093	0.621
13	14	3	0	0.093	0.621
14	14	3	1	0.093	0.621
15	14	3	2	0.093	0.621
16	14	3	3	0.093	0.621
17	14	3	5	0.093	0.621
18	14	3	10	0.093	0.621
19	21	1	0	0.096	0.609
20	21	1	1	0.096	0.609
21	21	1	2	0.096	0.609
22	21	1	3	0.096	0.609
23	21	1	5	0.096	0.609
24	21	1	10	0.103	0.580
25	21	2	0	0.079	0.679
26	21	2	1	0.079	0.678
27	21	2	2	0.079	0.678
28	21	2	3	0.079	0.678
29	21	2	5	0.079	0.680
30	21	2	10	0.079	0.679
31	21	3	0	0.079	0.680
32	21	3	1	0.076	0.690
33	21	3	2	0.077	0.689
34	21	3	3	0.076	0.690
35	21	3	5	0.079	0.681
36	21	3	10	0.079	0.680
37	28	1	0	0.092	0.628
38	28	1	1	0.092	0.625
39	28	1	2	0.092	0.625
40	28	1	3	0.092	0.625
41	28	1	5	0.092	0.625
42	28	1	10	0.099	0.597
43	28	2	0	0.069	0.718
44	28	2	1	0.067	0.726
45	28	2	2	0.067	0.726
46	28	2	3	0.067	0.726
47	28	2	5	0.065	0.738
48	28	2	10	0.069	0.718
49	28	3	0	0.071	0.710
50	28	3	1	0.063*	0.743
51	28	3	2	0.064	0.740
52	28	3	3	0.063	0.742
53	28	3	5	0.064	0.738
54	28	3	10	0.072	0.709

*) Smallest GCV Value

TABLE 5 Importance Rate of Predictor Variables

Variable	Importance Rate
Diabetes gymnastic (X_7)	100.00
Obesity (X_5)	64.30
Hypertension (X_6)	64.30
Age (X_1)	58.90
Long suffering (X_4)	30.40

TABLE 6 Comparison of The Accuracy Classification MARS and Bagging MARS

MARS	Bagging MARS	
	Replication	Classification Accuracy
90.63%	50	90.63%
	100	90.63%
	150	90.63%
	200	90.63%

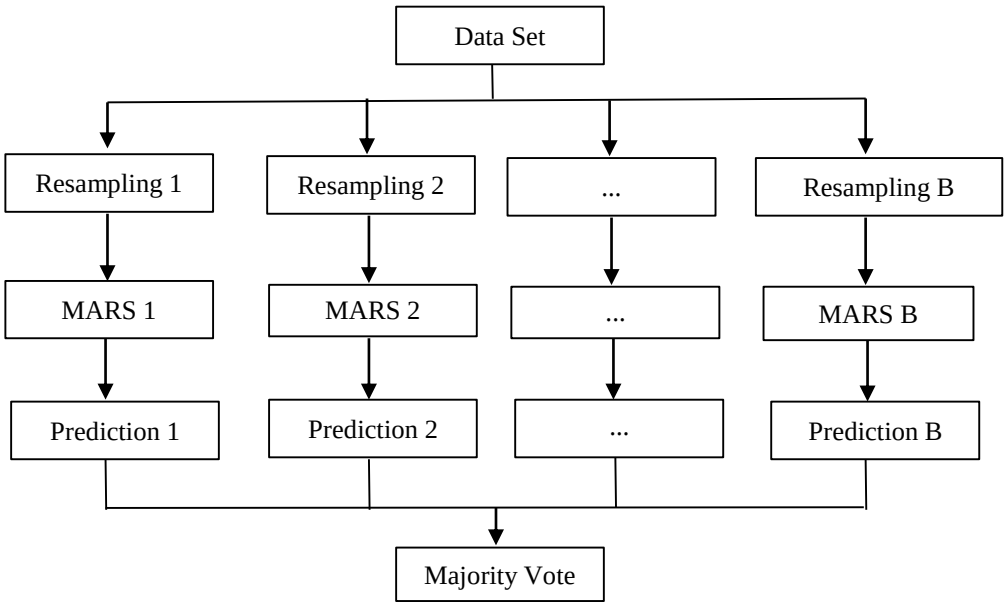

Fig. 1. Illustration of Bagging MARS

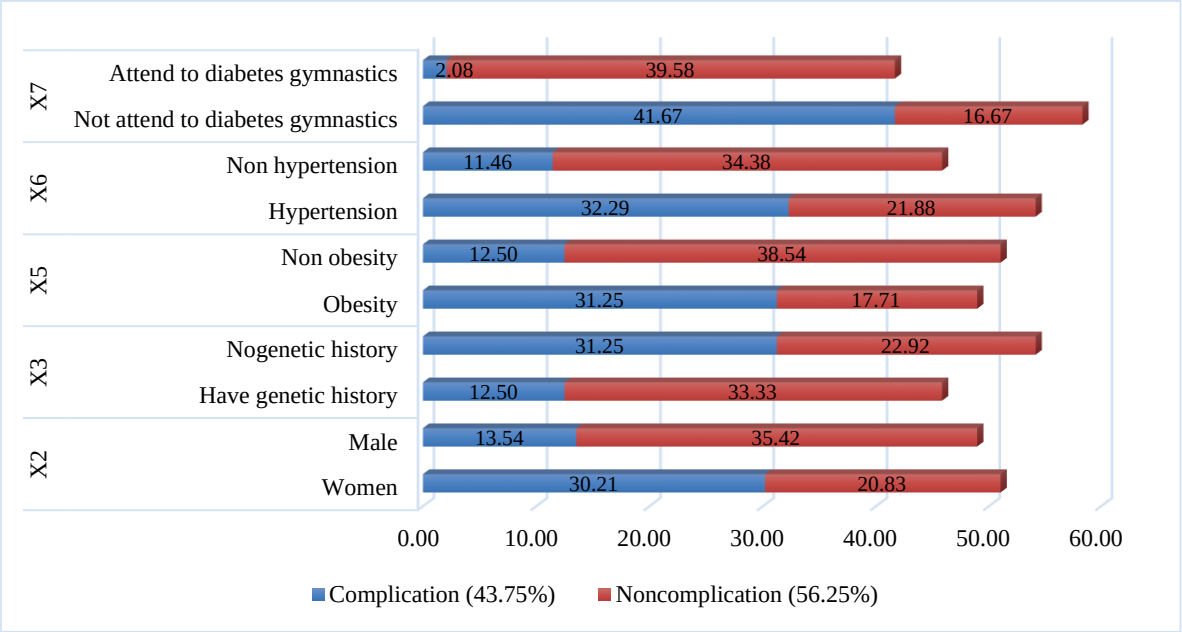

Fig. 2. Characteristics of Patients Based on Disease Complication

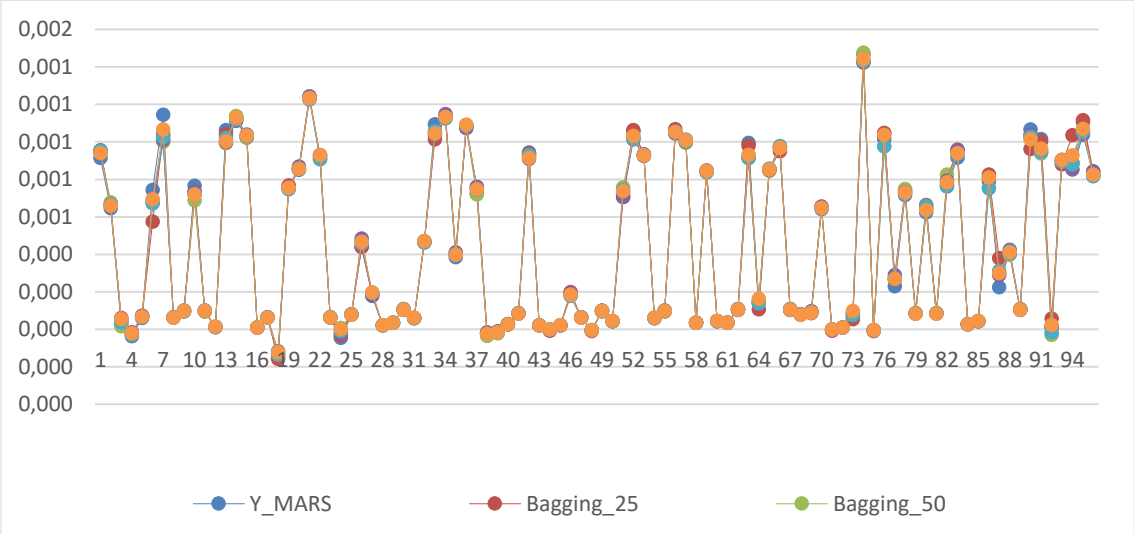

Fig. 3. Prediction Value of MARS and Bagging MARS