Cutting-edge Approach to Classifying Type II Diabetes Mellitus: Bagging, Boosting, Random Forest, and MARS Ensemble (Bag-Boost-RF-MARS)

Harison[1], Bambang Widjanarko Otok[2], Harun Al Azies[3,4]

[1]Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Riau, Pekanbaru, Indonesia
[2]Departement of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[4]Research Center for Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

DOI: https://doi.org/10.56293/IJASR.2023.5617

IJASR 2023
VOLUME 6
ISSUE 6 NOVEMBER – DECEMBER **ISSN: 2581-7876**

Abstract: This article aims to determine the prevalence of diabetes mellitus (DM) in East Java Province, which ranks ninth at 6.8, with Surabaya City leading in diabetes cases. To improve quality, the Social Health Insurance Administration Body (BPJS) has created a referral system, and the As-Shafa clinic in Sidoarjo serves as a reference for BPJS users and diabetes patients. 37 of the 75 female patients have normal blood glucose levels, while 38 have high levels. Thirteen males had normal readings, while 38 had excessive blood glucose. With BF = 15, MI = 1, and MO = 3, the best MARS model highlights gender and BMI as dominant variables. The MARS method obtains 0.6733 accuracy on training data and 0.6000 accuracy on testing data. Bagging MARS and random forest MARS both achieve 40% accuracy, whereas boosting MARS achieves 44.1%. The MARS approach beats other methods in identifying blood sugar levels in type 2 diabetes mellitus patients at the As-Shafa Clinic in Sidoarjo.

Keywords: MARS Ensemble, Diabetes Mellitus, Bagging, Boosting, Random Forest

1. Introduction

Hyperglycemia is a medical condition in the form of an increase in blood glucose levels beyond normal limits, namely high fasting blood glucose levels if >125 mg/dL, and is one of the typical signs of diabetes mellitus (DM) (Genuth et al., 2021). An increase in blood glucose levels that exceeds normal/hyperglycemia is one of the typical signs of diabetes mellitus (DM) (Aulia et al., 2021). There are two main categories of DM, namely, type I DM and type II DM. Type I or insulin-dependent diabetes is characterized by a lack of insulin production and type II diabetes or non-insulin independent due to the use of insulin that is less effective by the body. Insulin is a hormone that regulates the balance of blood sugar levels (Mitchell & Begg, 2021). Based on the Result of the National Basic Health Research (Riskesdas) in Indonesia, there are 10 million people with DM and 17.9 million people at risk of suffering from DM. The high prevalence of diabetes in Indonesia led the BPJS to establish a referral program, which is health services for participants with chronic diseases, including patients with diabetes mellitus. Patients who participate in BPJS and who have been diagnosed with diabetes mellitus will be referred to a first-level health facility (Faskes). In East Java Province, Indonesia, one of the clinics that have become a referral for BPJS participants who also suffer from DM is As-Shafa Clinic located in Sidoarjo Regency, East Java Province, Indonesia.

There are two risk factors for DM, namely non-modifiable and modifiable factors (Arifin et al., 2022). Non-modifiable risk factors are race and ethnicity, age, gender, family history of diabetes mellitus, history of childbirth with babies weighing over 4000 grams, and birth history with a low birth weight (less than 2500 grams) (Khazaei et al., 2021). Meanwhile, modifiable risk factors are closely related to unhealthy lifestyle behaviors, namely being overweight which causes abdominal/central obesity, lack of physical activity, hypertension, poor diet (Smith et al., 2022). Several studies on diabetes have been carried out, including Otok in 2020 which analyzes the factors for the

occurrence of complications in patients with type 2 diabetes. The result is that patients who exercise are less likely to have complications (B. W. Otok et al., 2020). Furthermore, Akolo and Otok in 2017 also researched peripheral diabetic neuropathy cases with the results of the analysis that the variables that directly affect the type of DM were obesity, age, gender (B. Otok et al., 2017). The research on type II diabetes mellitus indicates that many factors influence type II diabetes mellitus. Therefore, to explain the pattern of the relationship between response variables and predictor variables, a regression curve can be used. The regression curve approach that is often used is the parametric regression approach, which takes the form of a linear curve. However, not all relationship models between variables can be approximated with a parametric approach, because there is no information about the shape of the relationship between the response variable and the predictor variable. If the parametric model assumptions are not met, the regression curve can be estimated using a nonparametric regression model approach. Several nonparametric regression models are widely used, one of which is the multivariate adaptive regression spline (MARS) (Prihastuti Yasmirullah et al., 2021). MARS is a nonparametric regression approach developed by Friedman in 1990 using the spline function to estimate the model (Shahbaz et al., 2020). The review of the use of the MARS method for this modeling is based on the unclear model of the relationship between the response variables of the predictor variables (Mehdizadeh, 2020).

In the MARS method, there is a continuous response MARS and a categorical response MARS. In MARS, the categorical response uses Boostrap in MARS (Hasyim et al., 2018), while for continuous response MARS, MARS modeling on entrance exam results on GPA (Guerrero-Roldán et al., 2021). The level of precision of a classification method can be increased to provide better classification results and reduce the level of classification error, so the resampling method is performed in the preparation of the model to reduce the level of classification error. Bagging (bootstrap aggregation) and Boosting are relatively new ensemble methods but have become popular (Tuysuzoglu & Birant, 2020). One of the newer ensemble methods is the Random Forest which has been developed from the bagging process. Random forest was first introduced by Breiman in 2001(Arfiani & Rustam, 2019; Parmar et al., 2019). In his research, he showed the advantages of random forest, among other things, it can produce fewer errors, perform well in classification, can handle efficiently very large amounts of training data and is an efficient method for estimating missing data (Parmar et al., 2019). Previous random forest research conducted by (Sulaiman et al., 2011) investigated web caching by comparing classification accuracy using the CART, MARS, Random Forest, and Tree Net methods. Research on the application of the random forest method in the analysis of conductors (Yao et al., 2020). The ensemble method researches the classification of poverty in the Jombang regency and it was found that the random forest provides the best classification accuracy (B. W. Otok & Seftiana, 2012). The number of factors that influence the risk of type II DM, a classification analysis was performed to determine the classification determination of risk factors for DM with a case study at As-Shafa Clinic, Sidoarjo, East Java, Indonesia. One of the classification methods is the MARS method, but the results of the MARS classification are sometimes unsatisfactory, therefore resampling is carried out using the MARS Boosting and MARS bagging methods and the random forest which is one of the latest methods of 'entire Bagging process to obtain the best method of classification of factors risk of DM.

2. Materials and Methods

2.1 Data sources and research variables

The data used in this study are secondary in the form of medical records of patients with type II diabetes mellitus at As-Shafa Clinic, Sidoarjo in March 2017 with a total of 126 patients. The variables used in this study are presented in Table 1.

Table 1. Research Variables of Studies

Variable	Category
Age (X1)	-
Gender (X2)	X2(1): Male; X2 (2): Female
Body mass index (BMI) (X3)	-
Blood pressure (X4)	X4(1): There is Hypertension; X4 (2): No Hypertension
Sports Activities (X5)	X5(1): Active; X5 (2): Inactive/Less
Age (X1)	-

The response variable (Y) in this study is the blood glucose level of patients with type II diabetes, the experts state that the classification of blood glucose levels is based on fasting plasma glucose (FPG) [17][18]. The determination of the level of blood glucose levels can be seen in Table 2. If the blood sugar level is <100-125 mg/dl, it is classified as normal blood sugar which is classified as "0" and if the blood sugar value is at least 126 mg/dl, it is classified as high blood sugar and is classified with the code "1".

Table 2. Blood Sugar Classification Diabetes Diagnosis

Fasting Plasma Glucose (FPG) (mg/dl)	Qualification
Age (X1)	-
<100 mg/dl	Normal fasting glucose
100–125 mg/dl	IFG (impaired fasting glucose)
≥126 mg/dl	Provisional diagnosis of diabetes (the diagnosis must be confirmed)

2.2 Multivariate Adaptive Regression Splines (MARS) Concept

The MARS model is used to solve the problem of large-dimensional data (SABANCI & CENGİZ, 2022; Sukhinov et al., 2021). In addition, the MARS model also produces an accurate classification of response variables and produces a continuous node model based on the smallest generalized cross-validation (GCV) value (Yasmirullah et al., 2021). GCV is a method for obtaining optimal nodes. MARS was developed by Friedman (1990) for a nonparametric regression model approach between the response variables of several predictor variables in a piecewise regression (Yasmirullah et al., 2021). In general, the MARS model according to Friedman can be written in the following equation (Through et al., 2020; Yasmirullah et al., 2021)(1):

$$y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(x) + \varepsilon_i \quad (1)$$

with, α_0 is the constant coefficient of the basis function B_0, and α_m is the coefficient of the m-th basis function, as well as $B_m(x) = \prod_{k=1}^{Km}\left[S_{km}\left(x_{v(k,m)} - t_{km}\right)\right]$, so that when written in matrix form it is as follows:

$$\mathbf{y}=\mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (2)$$

$x_{v(k,m)}$ is the independent variable, t_{km} is the knot value of the independent variable $x_{v(k,m)}$, where M is the number of base functions, K_m is the number of interactions on the m-th basis function. S_{km} is a value that is worth 1 if the data is to the right of the knot point or -1 if the data is to the left of the knot point, v is the number of predictor variables, and k is the number of interactions. The contribution measure used in the MARS method uses the GCV criteria. GCV is used because it has optimal asymptotic properties (Bülbül & Purutçuoglu, 2021). The GCV method in general is like equation (3)

$$\text{GCV(M)} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_M(x_i)\right]^2 \Big/ \left[1 - \frac{C(M)}{n}\right]^2 \quad (3)$$

with,

M = the number of base functions

C(M) = number of model parameters = Trace $\left(\mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\right)+1$

B = matrix base function

n = amount of data

y_i = value of the response variable on the i-th observation

$\hat{f}(x_i)$ = the estimated value of the response variable on the i-th observation

Elements to take into account in the formation of the model MARS is as follows (Abed et al., 2023).

Knots are the end of one regression line (region) and the beginning of another regression line (region). At each knot point, it is expected that there will be continuity of the basis function between one region and another, with a minimum distance between knots or minimum observations between knots symbolized by "MO" of 0.1, 2, and 3.
The base function is a set of functions used to explain the relationship between the response variable and the prediction variable. The base function, which is symbolized by "BF" consists of one or more variables and is a parametric function defined in each region. In general, the selected basis function is a polynomial form with a continuous derivative at each node.

Interaction is a cross-product between interrelated or correlated variables. The maximum number of interactions symbolized by "MI" allowed is 1, 2, and 3. If it is greater than 3, the model will be difficult to interpret.

2.3 Classification Accuracy

Total Accuracy Rate (TAR) is used to calculate the classification accuracy of the grouping results. The TAR value can represent the proportion of the sample that is properly classified (Azies & Anuraga, 2021; Sai et al., 2020; Scheiber et al., 2023; Susilaningrum & Al Azies, 2017). Determination of the accuracy of the binary response MARS classification with calculations in table 3 of the following classifications (Nikita & Nikitas, 2020).

Table 3. Binary Response MARS Classification

Observation Results	Estimated Observation	
	y_0	y_1
y_0	n_{00}	n_{01}
y_1	n_{10}	n_{11}

where,

y_0 : classified as normal blood sugar
y_1 : classified as high blood sugar
n : number of observations
n_{00} : the number of observations from y_0 that are properly classified as y_0
n_{11} : the number of observations from y_1 that are correctly classified as y_1
n_{01} : the number of observations from y_0 that are incorrectly classified as y_1
n_{10} : the number of observations from y_1 that are incorrectly classified as y_0
The Total Accuracy Rate (TAR) value is obtained by the following calculation

$$\text{TAR}(\%) = \frac{n_{00} + n_{11}}{\boldsymbol{n}} \text{X } 100\% \quad (4)$$

2.4 Step analysis

A descriptive analysis was used in this study to investigate factors among the respondents. The information was then separated into two categories: training data and testing data. The MARS approach was used to construct a model for type II diabetes patient data at the As-Shafa Clinic in Sidoarjo. The MARS model was created by taking into account changes in:

The maximum number of base functions (BF) is 10, 15, or 20.
The maximum number of interactions (MI) with options one, two, and three.
The minimum distance between knots/minimum observation (MO) with values ranging from 0 to 10, with selections of 0, 1, 2, 3, 5, and 10.
The least generalized cross-validation (GCV) value was used to identify the optimal MARS model (Wu et al., n.d.).
The best MARS model was then used to classify the training and testing data, and the classification determination

was determined (Agr Sci-Tarim Bili & Çanga, 2022). Furthermore, the classification accuracy was evaluated using the bagging, boosting, and random forest MARS methods, which were based on the findings of the best model identified before. The bagging, boosting, and random forest MARS methods with the highest percentage accuracy value were chosen as having the best classification accuracy.

3. Results and Discussion

Characteristics of Respondents

To determine the characteristics of the data, a descriptive analysis was carried out to find out the general description of the data used in a study. Table 4 is a tabulation between blood glucose levels (Y) and gender (X2). It is known that from 75 female patients, 37 patients have normal blood glucose levels and 38 others have high blood glucose levels, while for male gender, it is known that of 51 patients, 13 had normal blood glucose levels and 38 had high blood glucose levels.

Table 4. Tabulation of blood glucose levels (Y) and gender (X2)

Type II DM patients	Gender	
	Female	Male
Normal blood glucose level	37	13
High blood glucose levels	38	38
Total	75	51

Modeling Blood Sugar Levels Using MARS

The results of MARS modeling for type II diabetes mellitus data at the As-Shafa Clinic, Sidoarjo in this study are shown in Table 5.

Table 5. MARS Modeling Results.

Model	BF	MI	MO	GCV	Model	BF	MI	MO	GCV
1	10	1	0	0.239	28	15	2	3	0.243
2	10	1	1	0.239	29	15	2	5	0.243
3	10	1	2	0.239	30	15	2	10	0.243
4	10	1	3	0.239	31	15	3	0	0.243
5	10	1	5	0.239	32	15	3	1	0.243
6	10	1	10	0.239	33	15	3	2	0.243
7	10	2	0	0.243	34	15	3	3	0.243
8	10	2	1	0.243	35	15	3	5	0.243
9	10	2	2	0.243	36	15	3	10	0.243
10	10	2	3	0.243	37	20	1	0	0.239
11	10	2	5	0.243	38	20	1	1	0.239
12	10	2	10	0.243	39	20	1	2	0.239
13	10	3	0	0.243	40	20	1	3	0.239
14	10	3	1	0.243	41	20	1	5	0.239
15	10	3	2	0.243	42	20	1	10	0.239
16	10	3	3	0.243	43	20	2	0	0.243
17	10	3	5	0.243	44	20	2	1	0.243
18	10	3	10	0.243	45	20	2	2	0.243
19	15	1	0	0.239	46	20	2	3	0.243
20	15	1	1	0.239	47	20	2	5	0.243
21	15	1	2	0.239	48	20	2	10	0.243

22*	15	1	3	0.239	49	20	3	0	0.243
23	15	1	5	0.239	50	20	3	1	0.243
24	15	1	10	0.239	51	20	3	2	0.243
25	15	2	0	0.243	52	20	3	3	0.243
26	15	2	1	0.243	53	20	3	5	0.243
27	15	2	2	0.243	54	20	3	10	0.243

*)

From all possible models based on the combination of BF, MI and MO values, the best MARS model is obtained with the criteria of having the smallest GCV value, namely the 22nd model with the combination of BF = 15, MI = 1 and MO = 3 which produces a GCV value = 0.239. Based on the 22nd model, the MARS model obtained is as follows.

$$Y = 0.4056 + 0.2787X_{21} + 0.0019\max(0, X_3 - 100) \quad (5)$$

Furthermore, from the MARS model in equation (5), it can be seen that there are two predictor variables that are included in the model and to see the extent to which these variables can be seen in Table 6.

Table 6. The level of importance of the predictor variable.

Variable	Level of importance
X2	100.0 %
X3	49.5 %
X1	0.0 %
X4	0.0 %
X5	0.0 %

In Table 6 above, it can be seen that the sex variable is the most important variable in the MARS model with an importance level of 100%, then followed by the Body mass index (BMI) variable with a large contribution to the model of 49.5%. Meanwhile, three variables have an importance level of 0.000%, which means that these variables are not included in the model because they are already represented by the variables included in the MARS model.

Accuracy of Classification of Blood Sugar Level Status Using the MARS Method

From the results of the best MARS model obtained, a classification will be carried out to find out how well the model is based on training and testing data.

Table 7. Classification of type II DM based on training data the MARS method

Actual data	Prediction data		Total
	Normal blood glucose level (0)	High blood glucose levels (1)	
Normal blood glucose level (0)	27	13	40
High blood glucose levels (1)	20	41	61
Total	47	54	101

It can be seen in Table 7 that of the 101 training data, 27 were correctly classified into the low blood pressure category and 13 were incorrectly classified from the low blood pressure category into the high blood pressure category, while 41 data were correctly classified into the high blood pressure category and 20 were incorrectly classified. classified from the category of high blood pressure into the category of low blood pressure.

Table 8. Classification of type II DM based on testing data the MARS method.

Actual data	Prediction data	Total

	Normal blood glucose level (0)	High blood glucose levels (1)	
Normal blood glucose level (0)	0	10	10
High blood glucose levels (1)	0	15	15
Total	0	25	25

It can be seen in Table 8 that of the 25 testing data, 10 were misclassified from the low blood pressure category into the high blood pressure category, while 15 data were correctly classified into the high blood pressure category. Based on the information from Tables 7 and 8, we can determine the classification in Table 9.

Table 9. The results of the accuracy of the MARS method classification

Data Source	Classification Accuracy
Training Data	67.33 %
Testing Data	60.00 %

Table 9 is the accuracy of the classification of blood sugar levels in DM-II patients based on the MARS model using the accuracy value classification can be seen in Table 9. The total classification accuracy is 67.33% for training data while for testing data is 60%.

Comparison of Classification Accuracy Using Bagging Mars, Boosting MARS and Random Forest MARS

The performance of the classification method is measured by the accuracy of the classification. After analyzing each method, the classification accuracy is obtained in table 10 below.

Table 10. The results of the accuracy of the MARS method classification.

Data Source	MARS	Bagging MARS	Boosting MARS	Random Forest MARS
Training Data	67.33 %	30.41 %	42.2 %	56.44 %
Testing Data	60.00 %	40.00 %	44.1 %	40.00 %

In Table 10 above, it can be seen that the random forest method has the highest classification accuracy, which is 56.44% on the training Data and 40% on the testing data, between the bagging and boosting methods, but the accuracy results of the MARS method are still superior to the other three methods, so it can be concluded that for the analysis of the classification of blood sugar levels in type 2 DM patients at the As-Shafa Clinic, Sidoarjo, Indonesia, it is better to use the MARS method.

Conclusions

The best model for the blood sugar level of type 2 Diabetes Mellitus patients at As-Shafa Clinic Sidoarjo, Indonesia contains two significant variables, the variables that have the highest importance for the blood sugar level of type 2 Diabetes Mellitus patients are gender and Boddy Mass Index (BMI). The level of accuracy of the classification of blood sugar levels in Type 2 Diabetes Mellitus patients at the As-Shafa Clinic Sidoarjo, Indonesia using the MARS method produces an accuracy of 60%. The classification accuracy using the Bagging mars and random forest MARS methods is the same, namely 40%, while the boosting MARS is 44.1%. The classification of the MARS method is better than the Bagging MARS, Boosting MARS and random forest MARS methods.

References

1. Abed, M. S., Kadhim, F. J., Almusawi, J. K., Imran, H., Filipe, L., Bernardo, A., & Henedy, S. N. (2023). Utilizing Multivariate Adaptive Regression Splines (MARS) for Precise Estimation of Soil Compaction Parameters. Applied Sciences 2023, Vol. 13, Page 11634, 13(21), 11634. https://doi.org/10.3390/APP132111634
2. Agr Sci-Tarim Bili, J., & Çanga, D. (2022). Use of MARS Data Mining Algorithm Based on Training and Test Sets in Determining Carcass Weight of Cattle in Different Breeds. Journal of Agricultural Sciences, 28(2), 259–268. https://doi.org/10.15832/ANKUTBD.818397

3. Arfiani, A., & Rustam, Z. (2019). Ovarian cancer data classification using bagging and random forest. AIP Conference Proceedings, 2168(1). https://doi.org/10.1063/1.5132473/611710
4. Arifin, H., Chou, K. R., Ibrahim, K., Fitri, S. U. R., Pradipta, R. O., Rias, Y. A., Sitorus, N., Wiratama, B. S., Setiawan, A., Setyowati, S., Kuswanto, H., Mediarti, D., Rosnani, R., Sulistini, R., & Pahria, T. (2022). Analysis of Modifiable, Non-Modifiable, and Physiological Risk Factors of Non-Communicable Diseases in Indonesia: Evidence from the 2018 Indonesian Basic Health Research. Journal of Multidisciplinary Healthcare, 15, 2203–2221. https://doi.org/10.2147/JMDH.S382191
5. Aulia, D., Suprapto, S. I., & Soemarko, S. (2021). Relationship of Diet and Lifestyle with Blood Sugar Levels in the Elderly with Diabetes Mellitus at Internist Room in Dr. Moedjito Dwidjosiswoyo Hospital of Jombang. Journal for Quality in Public Health, 4(2), 303–313. https://doi.org/10.30994/JQPH.V4I2.187
6. Azies, H. Al, & Anuraga, G. (2021). Classification of Underdeveloped Areas in Indonesia Using the SVM and k-NN Algorithms. Jurnal ILMU DASAR, 22(1), 31–38. https://doi.org/10.19184/JID.V22I1.16928
7. Bülbül, G. B., & Purutçuoglu, V. (2021). Novel model selection criteria for LMARS: MARS designed for biological networks. Journal of Statistical Computation and Simulation, 91(9), 1749–1761. https://doi.org/10.1080/00949655.2020.1870689
8. Genuth, S. M., Palmer, J. P., & Nathan, D. M. (2021). Classification and Diagnosis of Diabetes. Diabetes in America, 3rd Edition, 2(4), 1–39. http://europepmc.org/books/NBK568014
9. Guerrero-Roldán, A. E., Rodríguez-González, M. E., Bañeres, D., Elasri-Ejjaberi, A., & Cortadas, P. (2021). Experiences in the use of an adaptive intelligent system to enhance online learners' performance: a case study in Economics and Business courses. International Journal of Educational Technology in Higher Education, 18(1), 1–27. https://doi.org/10.1186/S41239-021-00271-0/TABLES/17
10. Hasyim, M., Rahayu, D. S., Muliawati, N. E., Hayuhantika, D., Puspasari, R., Anggreini, D., Hastari, R. C., Hartanto, S., & Utomo, F. H. (2018). Bootstrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) to Analyse the Lecturer Research Performance in Private University. Journal of Physics: Conference Series, 1114(1), 012117. https://doi.org/10.1088/1742-6596/1114/1/012117
11. Khazaei, Z., Bagheri, M. M., Goodarz, E., Moayed, L., Abadi, N. E., Bechashk, S. M., Mohseni, S., Safizadeh, M., Behseresht, M., & Naghibzadeh-Tahami, A. (2021). Risk Factors Associated with Low Birth Weight Among Infants: A Nested Case-Control Study in Southeastern Iran. International Journal of Preventive Medicine, 12(1). https://doi.org/10.4103/IJPVM.IJPVM_300_20
12. Mehdizadeh, S. (2020). Using AR, MA, and ARMA Time Series Models to Improve the Performance of MARS and KNN Approaches in Monthly Precipitation Modeling under Limited Climatic Data. Water Resources Management, 34(1), 263–282. https://doi.org/10.1007/S11269-019-02442-1/TABLES/9
13. Mitchell, C. S., & Begg, D. P. (2021). The regulation of food intake by insulin in the central nervous system. Journal of Neuroendocrinology, 33(4), e12952. https://doi.org/10.1111/JNE.12952
14. Nikita, E., & Nikitas, P. (2020). On the use of machine learning algorithms in forensic anthropology. Legal Medicine, 47, 101771. https://doi.org/10.1016/J.LEGALMED.2020.101771
15. Otok, B., Purhadi, Andari, S., & Akolo, I. R. (2017). Propensity Score Stratification Using Logistic Regression Bootstrap in the Case of Peripheral Diabetic Neuropathy.
16. Otok, B. W., Putra, R. Y., Sutikno, & Yasmirullah, S. D. P. (2020). Bootstrap aggregating multivariate adaptive regression spline for observational studies in diabetes cases. Systematic Reviews in Pharmacy, 11(8), 406–413. https://doi.org/10.31838/SRP.2020.8.59
17. Otok, B. W., & Seftiana, D. (2012). The Classification of Poor Households in Jombang With Random Forest Classification And Regression Trees (RF-CART) Approach as the Solution In Achieving the 2015 Indonesian MDGs' Targets. International Journal of Science and Research (IJSR) ISSN, 3. www.ijsr.net
18. Parmar, A., Katariya, R., & Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. Lecture Notes on Data Engineering and Communications Technologies, 26, 758–763. https://doi.org/10.1007/978-3-030-03146-6_86/COVER
19. Prihastuti Yasmirullah, S. D., Otok, B. W., Trijoyo Purnomo, J. D., & Prastyo, D. D. (2021). Modification of Multivariate Adaptive Regression Spline (MARS). Journal of Physics: Conference Series, 1863(1), 012078. https://doi.org/10.1088/1742-6596/1863/1/012078
20. SABANCI, D., & CENGİZ, M. A. (2022). Random Ensemble MARS: Model Selection in Multivariate Adaptive Regression Splines Using Random Forest Approach. Journal of New Theory, 40, 27–45. https://doi.org/10.53570/JNT.1147323
21. Sai, F. L., Shahrill, M., Tan, A., & Zurimi, S. (2020). Analysis of Multivariate Adaptive Regression Spline (MARS) Model in Classifying factors affecting on Student the Study Period at FKIP Darussalam University

of Ambon. Journal of Physics: Conference Series, 1463(1), 012005. https://doi.org/10.1088/1742-6596/1463/1/012005

22. Scheiber, L., Jurado, A., Pujades, E., Criollo, R., & Suñé, E. V. (2023). Applied multivariate statistical analysis as a tool for assessing groundwater reactions in the Niebla-Posadas aquifer, Spain. Hydrogeology Journal, 31(2), 521–536. https://doi.org/10.1007/S10040-022-02580-8/TABLES/3
23. Shahbaz, M., Khraief, N., & Mahalik, M. K. (2020). Investigating the environmental Kuznets's curve for Sweden: evidence from multivariate adaptive regression splines (MARS). Empirical Economics, 59(4), 1883–1902. https://doi.org/10.1007/S00181-019-01698-1/TABLES/4
24. Smith, K. L., Danyluk, A. B., Munir, S. S., & Covassin, N. (2022). Shift Work and Obesity Risk—Are There Sex Differences? Current Diabetes Reports, 22(8), 341–352. https://doi.org/10.1007/S11892-022-01474-Z/FIGURES/1
25. Sukhinov, A., Belova, Y., Chistyakov, A., Beskopylny, A., & Meskhi, B. (2021). Mathematical Modeling of the Phytoplankton Populations Geographic Dynamics for Possible Scenarios of Changes in the Azov Sea Hydrological Regime. Mathematics 2021, Vol. 9, Page 3025, 9(23), 3025. https://doi.org/10.3390/MATH9233025
26. Sulaiman, S., Shamsuddin, S. M., & Abraham, A. (2011). Intelligent Web caching using Adaptive Regression Trees, Splines, Random Forests and Tree Net. Conference on Data Mining and Optimization, 108–114. https://doi.org/10.1109/DMO.2011.5976513
27. Susilaningrum, D., & Al Azies, H. (2017). PEMODELAN REGRESI LOGISTIK PADA FAKTOR YANG MEMPENGARUHI PHBS PADA RUMAH TANGGA PENDERITA TBC DI PESISIR SURABAYA. EKSAKTA: Berkala Ilmiah Bidang MIPA, 18(02), 121–128. https://doi.org/10.24036/EKSAKTA/VOL18-ISS02/65
28. Through, W., Sak Ceting, S., Eka Suranny, L., Christi Maharani -, F., Azizah, D. M., & Permatasari, E. O. (2020). Modeling of toddler stunting in the province of east nusa tenggara using multivariate adaptive regression splines (mars) method. Journal of Physics: Conference Series, 1490(1), 012013. https://doi.org/10.1088/1742-6596/1490/1/012013
29. Tuysuzoglu, G., & Birant, D. (2020). Enhanced Bagging (eBagging): A Novel Approach for Ensemble Learning. INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY, 17(4), 515–528. https://doi.org/10.34028/IAJIT/17/4/10
30. Wu, C., Goh, A., Zhang, W., Wu, C. Z., Goh, A. T. C., & Zhang, W. G. (n.d.). Study on Optimization of Mars Model for Prediction of Pile Drivability Based on Cross-Validation. https://doi.org/10.3850/978-981-11-2725-0-MS2-7-cd
31. Yao, R., Li, J., Hui, M., Bai, L., & Wu, Q. (2020). Feature Selection Based on Random Forest for Partial Discharges Characteristic Set. IEEE Access, 8, 159151–159161. https://doi.org/10.1109/ACCESS.2020.3019377
32. Yasmirullah, S. D. P., Otok, B. W., Purnlmo, J. D. T., & Prastyo, D. D. (2021). Multivariate adaptive regression spline (MARS) methods with application to multi drug-resistant tuberculosis (mdr-tb) prevalence. AIP Conference Proceedings, 2329(1). https://doi.org/10.1063/5.0042145/962566