Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 1, 2026
DOI: 10.30829/zero.v10i1.28733
Page: 365-378

Nonlinear Ordinal Logistic Regression and Multivariate Adaptive Regression Splines (NORL-MARS) for Prediction of Diabetes Mellitus Risk

[1] **Any Tsalasatul Fitriyah**
Universitas Islam Negeri Mataram, Mataram, 83116, Indonesia

[2] **Maylita Hasyim**
Universitas Bhinneka PGRI, Tulungangung, 66221, Indonesia

[3] **Nur Chamidah**
Airlangga University, Surabaya, 60115, Indonesia

[4] **Toha Saifudin**
Airlangga University, Surabaya, 60115, Indonesia

[5] **Vita Fibriyani**
Universitas Merdeka Pasuruan, Pasuruan, 67129, Indonesia

Article Info

Article history:

Accepted 15 May, 2026

Keywords:

Diabetes Mellitus;
MARS;
Nonparametric Regression;
Ordinal Logistic Regression

ABSTRACT

Diabetes Mellitus (DM) is a high-risk metabolic disease with increasing prevalence in Indonesia, requiring an effective classification model based on significant risk factors. This study uses Nonparametric Ordinal Logistic Regression based on the Multivariate Adaptive Regression Spline estimator (NOLR-MARS). Unlike conventional parametric ordinal regression, this model does not assume a fixed functional pattern but rather determines the form of the relationship based on data patterns through basis functions, making it more flexible in handling complex predictor variable interactions. Using 664 records from the Non-Alcoholic Fatty Liver Disease (NAFLD) cohort, we explore the relationship between metabolic factors, included age, sex, Body Mass Index (BMI), LDL cholesterol, and hypertension—and DM risk. This NOLR-MARS integration addresses the nonlinear relationship while maintaining the ordinal nature of DM stages, a combination often overlooked in traditional models. Based on Generalized Cross Validation (GCV) selection, the best model achieved 74.92% accuracy for in-sample data and 80.30% for out-sample data. Furthermore, a sensitivity of 70% and a specificity of 92.86% were obtained for stage 2 DM. Factors such as age, BMI, LDL cholesterol, and hypertension significantly influenced DM status. The results showed that the NORL-MARS model had good predictive performance. The novelty of this study lies in the integration of the MARS estimator into an ordinal logistic regression framework for more granular DM risk assessment. Although this model shows potential as a screening tool in high-risk metabolic cohorts, further clinical application requires external validation to ensure broader generalizability.

Corresponding Author:

Any Tsalasatul Fitriyah,
Universitas Islam Negeri Mataram
Email: any.tsalasatul@uinmataram.ac.id

1. INTRODUCTION

Indonesia is currently transitioning from infectious to non-communicable diseases (NCDs), which account for 70% of global deaths, primarily in low- and middle-income countries [1]. Diabetes mellitus (DM) is the most prevalent and deadly NCD, caused by the body's inability to use insulin efficiently for blood sugar regulation [2], [3]. The prevalence of DM is rising globally and in Indonesia, which ranked seventh worldwide with 10.7 million patients in 2019 and a projected 16.6 million cases by 2045 [4], [5]. Since DM is preventable through lifestyle changes, studying its risk factors is essential [6].

Approximately 90% of global cases are Type II DM [7], influenced by non-modifiable factors like age and gender, as well as modifiable factors including abnormal lipid metabolism, high cholesterol, obesity, and hypertension [3], [8]. Several statistical approaches have been used to model DM risk factors. Various statistical approaches have been employed to model the risk factors associated with DM. Several studies, such as those by [7], [9], [4], [10], [11] analyzed factors associated with type-2 DM, which showed that age, BMI, family history, physical activity, and diet had a significant association with the risk of type-2 DM. Other studies by [12], [13], [8] concluded that blood pressure, hypertension, and cholesterol were significant causal factors influencing the risk of DM. Methodologically [14] employed bi-response logistic and probit regression, while others like [15], [16]. utilized standard logistic regression. Beyond parametric methods, research using nonparametric approaches has also been conducted, such as [17] using local linear estimation and [18] using Chi-Square tests. In terms of classification, [19] applied Smooth Support Vector Machine (SSVM), while [20] implemented Neural Networks and Multiple Regression.

These studies mostly use parametric regression methods, such as logistic regression. This approach often struggles to capture complex, non-linear relationships and interactions between predictors, despite the foundational ordinal regression theories established by McCullagh [21]. To address these limitations, nonparametric regression is proposed as an alternative. Nonparametric regression is an approach that does not assume a specific relationship pattern between the response variable and the predictor [22], [23]. In a nonparametric regression approach, the shape of the estimated relationship pattern model is determined based on patterns in the data [24], [25]. The ability of nonparametric regression to find the shape of the regression curve pattern is supported by the presence of parameters in each type of nonparametric regression method/approach, which makes regression curve pattern estimation more flexible [26], [27]. This flexibility is essential for metabolic data where the risk factor interactions are often high-dimensional and non-linear.

One of the estimation methods in the nonparametric approach is the Multivariate Adaptive Regression Spline (MARS), This estimation method is able to produce good prediction results when handling data that exhibits behavioral changes within certain sub-intervals, indicating changes in the data's behavioral pattern [28], [29], [30] and is able to handle high-dimensional and complex data interactions [28], [31], [32] In addition, MARS also accommodates interactions between predictor variables expressed in basis functions and classifies categorical responses [28], [33], [34]. Several studies related to MARS estimation have been conducted by [35] to develop a flowing bottom-hole pressure (FBHP) estimation model. In addition [36], [37], [38] also conducted research with MARS. The models produced by MARS in these studies have been well verified and validated, showing high accuracy and stability in various applications.

Although several previous studies have used regression methods to analyze DM risk factors, and others have applied the MARS method in different contexts, no study has combined these two methods to predict DM risk factors using MARS. This study is the first to combine ordinal logistic regression with MARS basis functions for diabetes risk prediction. The data structure in this study has a response variable with an ordinal scale, this is very suitable for analysis using the MARS method. Preserving the ordinal structure is clinically vital, as it allows for a more granular assessment: a patient in the Stage-2 DM category requires significantly more intensive clinical intervention than one in the pre-DM or Stage-1 stages, a distinction that binary classification models often overlook. MARS uses logistic regression techniques for classification [29], [34]. In this study, we proposes a Nonparametric Ordinal Logistic regression model based on the MARS estimator (NORL-MARS) for ordinal DM status with three categories. Early prediction using this robust model is essential to help reduce healthcare costs and increase the effectiveness of prevention efforts. This approach aligns with the increasing global burden of non-communicable diseases, demanding more effective management strategies to reduce social and economic impacts.

2. RESEARCH METHOD

2.1 Dataset and Research Variables

The data used in this study are secondary data taken from the website: https://www.kaggle.com/datasets/utkarshx27/non-alcohol-fatty-liver-disease This dataset was selected because NAFLD and DM share a common pathophysiological mechanism rooted in insulin resistance and metabolic syndrome [39]. Clinical evidence suggests that NAFLD is a strong predictor for the development of DM [40], [41], making this cohort highly relevant for modeling metabolic risk factors. However, it is explicitly acknowledged as a methodological constraint that while NAFLD is a validated predictor of DM, findings from this cohort may not generalize to populations without hepatic metabolic dysfunction. External validation in dedicated diabetes screening cohorts is essential before clinical deployment.

The research dataset consists of DM status (Y) as an ordinal response variable, categorized into three levels based on HbA1c thresholds: pre-DM (<5.8%), stage-1 DM (5.8% – 6.4%), and stage-2 DM (≥6.5%). These stages were explicitly mapped based on the clinical diagnosis and metabolic profiles established in the original NAFLD study. The initial data extraction yielded 1,876 subjects who met the inclusion criteria for metabolic monitoring. Then, we provide data on age (X_1) as an interval variable (years); gender (X_2) as a nominal variable with 1 for male and 2 for female; BMI (X_3) in kg/m^2 and LDL cholesterol (X_4) in mg/dL as ratio variable; and hypertension status of users (X_5) as an ordinal variable with 1 as pre-hypertension, 2 for stage-1 hypertension and 3 for stage-2 hypertension as variables considered to affect DM status. For the MARS algorithm, categorical predictors such as gender and hypertension status were processed using dummy coding to allow the model to capture non-linear interactions between specific categories.

Preprocessing involved standardized procedures to ensure data integrity. From the 1,876 extracted records, 63.8% (n=1,198) were removed due to missing values in critical laboratory or anthropometric fields, resulting in 678 complete cases. Subsequently, outliers were identified and removed using the Interquartile Range (IQR) method with a standard 1.5xIQR multiplier threshold. This threshold was selected to strike a balance between preserving natural clinical variation and eliminating biologically implausible values, such as extreme physiological readings likely resulting from recording errors which could otherwise disproportionately distort the spline-fitting process in the MARS algorithm. This step removed an additional 2% (n=14) of records, resulting in a final analytical sample of 664 patients. Continuous predictors (Age, BMI, LDL) were then standardized using Z-score transformation to ensure statistical stability.

The NOLR-MARS model estimation was analyzed using R Studio (version 2024.12.0+467 "Kousa Dogwood") with the 'earth' package. The dataset was split into 598 patients (90%) as in-sample data and 66 patients (10%) as out-sample data. Given the specific nature of the cohort (n=664), this single holdout split was deemed sufficient to maintain a representative number of subjects in the test set for all three DM stages, while bootstrap-based uncertainty estimates were calculated to ensure the robustness of the accuracy metrics. Although the dataset shows a higher prevalence of stage-2 DM, no class-weighted loss functions or resampling were applied to preserve the natural clinical distribution. To mitigate potential bias toward the majority class, model performance was evaluated using stage-specific sensitivity and specificity (one-vs-rest), which provides a more granular assessment of the model's ability to distinguish between each progression stage regardless of class frequency.

Table 1. Dataset of DM Cases

Patient Number	DM Status (Y)	Age (X_1)	Gender (X_2)	BMI (X_3)	LDL Cholesterol (X_4)	Hypertension Status (X_5)
1	1	47	2	25.91	144	2
2	1	33	1	31.02	102	1
3	1	35	2	34.95	81	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
664	3	55	2	42.04	170	3

2.2 Ordinal Logistic Regression

Ordinal Logistic Regression (OLR) is a technique that looks at the association between predictor and response variables when the response variable is polychotomous and has an ordinal scale [42], [43]. The cumulative logit model is suitable for OLR, as it accounts for the ordinal nature of the response by representing it through cumulative probabilities. Consequently, it compares the cumulative probability of being less than or equal to a given response variable category based on predictor variables expressed in vector form **X** yields the cumulative logit model, or can be written $P = (Y \leq r|X_i)$ which is written in equation (1) [44].

$$P = (Y \leq r|X_i) = \pi(x_i) = \left(\frac{exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}\right) \quad (1)$$

where P is the cumulative probability that the DM risk (Y) falls in category r or lower, given the predictor X_i, $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ represents the vector of predictor variables for observation $(i = 1,2,\cdots,n)$, while $r = 1,2,\cdots,R$ is the response variable category. The logit transformation is applied to process OLR parameters to estimate them $P = (Y \leq r|X_i)$ described in equation (2) [44]:

$$g_r(x) = logitP(Y \leq r|X_i) = \ln \frac{P(Y \leq r|X_i)}{1 - P(Y \leq r|X_i)}$$
$$g_r(x) = \beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik} \quad (2)$$

where the value β_k for each $k = 1,2,\cdots,p$ in each OLR model is the same. The model in Equation (2) follows the proportional odds assumption, also known as the parallel lines assumption. This assumption implies that the relationship between each predictor (x_{ik}) and the cumulative odds is the same across all response categories *(r)*. Consequently, the slope coefficients (β_k) remain constant, and only the intercept values (β_{0r})vary for each threshold, representing the point where the response shifts from one category to the next. For example, if there are three response categories, namely $r = 1,2,3$ then the cumulative probability of r category response is described as equation (3) and (4) [44],

$$P = (Y \leq 1|x_i) = \left(\frac{exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}\right) \tag{3}$$

$$P = (Y \leq 2|x_i) = \left(\frac{exp\left(\beta_{02} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + exp\left(\beta_{02} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}\right) \tag{4}$$

The cumulative probability of each category of response variables is derived using Equation (3-4), specifically explained in equations (5-6),

$$P = (Y_r = 1|x_i) = \pi_1(x) \tag{5}$$

$$P = (Y_r = 2|x_i) = \pi_2(x) = P(Y \leq 2|x_i) - P(Y_r = 1|x_i) \tag{6}$$

By applying Equation (2), the cumulative logit model for the three response categories can be formulated, with the detailed expression shown in Equation (7).

$$\begin{aligned} \widehat{g_1}(x) &= \ln\left(\frac{P(Y \leq 1|x)}{1 - P(Y \leq 1|x)}\right) = \ln\left(\frac{P(Y \leq 1|x)}{(Y > 1|x)}\right) \\ &= \ln\left(\frac{P(Y = 1|x)}{P(Y = 2|x) + P(Y = 3|x)}\right) = \beta_{01} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \\ \widehat{g_2}(x) &= \ln\left(\frac{P(Y \leq 2|x)}{1 - P(Y \leq 2|x)}\right) = \ln\left(\frac{P(Y \leq 2|x)}{(Y > 2|x)}\right) \\ &= \ln\left(\frac{P(Y = 1|x) + P(Y = 2|x)}{P(Y = 3|x)}\right) = \beta_{02} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \end{aligned} \tag{7}$$

As shown in Equation (7), the model maintains the same slope coefficients $\beta_k$for both logits, which satisfies the proportional odds assumption. The only difference between the equations lies in the threshold intercepts β_{01} and β_{02}. In this study, the linear predictor $\sum_{k=1}^{p} \beta_k x_{ik}$ in the OLR model is replaced by the MARS basis functions. Therefore, the integration of NOLR-MARS allows the model to capture non-linearities and interactions between metabolic risk factors through the basis functions within the logit link framework.

2.3 Multivariate Adaptive Regression Splines (MARS)

The problem of nonparametric regression often encountered when using multi-predictor data is the high-dimensional case (curse of dimensionality) [22], [23]. One method that can overcome high-dimensional cases is Multivariate Adaptive Regression Splines (MARS), an adaptive approach that combines truncated spline regression and Recursive Partitioning Regression (RPR) [28], [29] . While RPR models are not continuous at the knot, MARS overcomes this by creating a continuous knot model that distinguishes between linear and composite functions [24], [33]. This study utilizes the quantile method to determine knot points based on data density, effectively placing knots even in sparse regions to minimize overfitting and reduce the risk of unrepresentative knots compared to a greedy approach [45], [46].Furthermore, the forward step in the MARS method produces more representative knot points, thereby accelerating the search for the optimal basis function without reducing model accuracy [47], with the model function written in Equation (8).

$$\hat{f}(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_M} \left[s_{km}.\left(x_{v(k,m)} - t_m\right)\right] \tag{8}$$

where β_0 is the intercept and β_m is the coefficient of the m-th basis function. The value M indicates the maximum number of basis functions used in the model, while K_m denotes the degree of interaction for the m-th term. The term s_{km} is a directional indicator which takes the value of either 1 or -1, depending on whether the data is to the right or left of the knot point. Furthermore, $x_{v(k,m)}$ represents the predictor variable, and t_{km} is the corresponding knot point of $x_{v(k,m)}$.

The smallest GCV (Generalized Cross Validation) is produced by the best model and is represented in equation (9):

$$GCV(M) = \frac{ASR}{\left[1 - \frac{C(\widehat{M})}{n}\right]^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - \hat{f}_M(x_i)\right]^2}{\left[1 - \frac{C(\widehat{M})}{n}\right]^2} \tag{9}$$

where, y_i denotes the response variable, and $\hat{f}_M(x_i)$ is the estimated value of the response variable using M basis functions. The term $C(\widehat{M}) = C(M) + dM$ refers to a complexity measure that adjusts based on the number of basic functions, where d is a constant representing the value added when each basis function reaches optimization.

If more than one model has a minimum GCV, the next step is to use the criteria for the maximum R_{square} value and the highest classification accuracy value. R_{square} formula is presented in equation (10)

$$R_{square} = 1 - \frac{\sum_{i=1}^{n}\left(f(x_i) - \hat{f}(x_i)\right)^2}{\sum_{i=1}^{n}\left(f(x_i) - \bar{f}(x_i)\right)^2} \tag{10}$$

where, n is number observations, with $f(x_i)$ is response observation value to i, $\hat{f}(x_i)$ is predicted response value to i, and $\bar{f}(x_i)$is average response value to *i*. Since traditional R^2 is only applicable to continuous linear regression, $Pseudo - R^2$ is employed here to evaluate the goodness-of-fit for the ordinal categorical response. This metric, combined with GCV, ensures that the selected NOLR-MARS model achieves an optimal balance between complexity and predictive accuracy.

Additionally, Equations (7) and (8) are integrated into the NOLR-MARS model. For simplicity, the term $\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_m\right)\right]$ is denoted as the basis function $B_m(x)$. By replacing the linear predictor in the cumulative logit model with these MARS basis functions, the equations for the three ordinal categories can be formulated in equation (11) [44].

$$\begin{aligned} \widehat{g_1}(x) &= \ln\left(\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}\right) = \beta_{01} + \sum_{m=1}^{M} \beta_m B_m(x) \\ \widehat{g_2}(x) &= \ln\left(\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}\right) = \beta_{02} + \sum_{m=1}^{M} \beta_m B_m(x) \end{aligned} \tag{11}$$

where β_{01} and $\beta_{02}$are the threshold intercepts that satisfy the proportional odds assumption. If there are three response categories, namely $r = 1,2,3$, then the cumulative probability of the r_i category response is described in equation (12),

$$\begin{aligned} P(Y \leq 1|x_i) &= \frac{exp(\beta_{01} + \sum_{m=1}^{M}\beta_m B_m(x))}{1 + exp\,[\beta_{01} + \sum_{m=1}^{M}\beta_m B_m(x)]} \\ P(Y \leq 2|x_i) &= \frac{exp(\beta_{02} + \sum_{m=1}^{M}\beta_m B_m(x))}{1 + exp\,[\beta_{02} + \sum_{m=1}^{M}\beta_m B_m(x)]} \end{aligned} \tag{12}$$

If there are three response categories, the cumulative probability for each category is first computed using Equation (12). The probability for each specific category is then derived from these cumulative probabilities and expressed individually in Equations (13–15). Finally, the cumulative logit formulation for these three categories is presented in Equation (16).

$$P(Y_r = 1|x_i) = \pi_1(x) = \frac{exp(\beta_{01} + \sum_{m=1}^{M}\beta_m B_m(x))}{1 + exp\,[\beta_{01} + \sum_{m=1}^{M}\beta_m B_m(x)]} \tag{13}$$

$$P(Y_r = 2|x_i) = \pi_2(x) == \frac{exp(\beta_{02} + \sum_{m=1}^{M}\beta_m B_m(x))}{1 + exp\,[\beta_{02} + \sum_{m=1}^{M}\beta_m B_m(x)]} - \pi_1(x) \tag{14}$$

$$P(Y_r = 3|x_i) = \pi_3(x) = 1 - \pi_1(x) - \pi_2(x) \tag{15}$$

For $\pi_1(x) = P(Y = 1)$; $\pi_2(x) = P(Y = 2)$; $\pi_3(x) = P(Y = 3)$, then it can be written,

$$\begin{aligned} \ln\frac{\pi_1}{\pi_2 + \pi_3} &= \beta_{01} + \sum_{m=1}^{M} \beta_m B_m(x) \\ \ln\frac{\pi_1 + \pi_2}{\pi_3} &= \beta_{02} + \sum_{m=1}^{M} \beta_m B_m(x) \end{aligned} \tag{16}$$

2.4 Metrics Evaluation

From the final model formed as in Equation (11), then a confusion matrix is created by applying each model to predict the test data. From the confusion matrix, the classification performance obtained from each model is then measured in the form of accuracy, sensitivity, and specificity. If the three measurements are close to 100%, then it can be stated that the classification has been done well [48]. The confusion matrix is presented in Table 2.

Table 2. Confusion Matrix 3x3

Observation	Prediction			Total
	y_1	y_2	y_3	
y_1	n_{11}	n_{12}	n_{13}	n_1
y_2	n_{21}	n_{22}	n_{23}	n_2
y_3	n_{31}	n_{32}	n_{33}	n_3

The confusion matrix in Table 2 can yield three measures of model strength: accuracy, sensitivity, and specificity. Validation is carried out using testing data information to determine whether the strength of the NOLR-MARS model can predict data outside the model. The maximum value of accuracy is 100%. The accuracy calculation algorithm when classifying data with multiple classes (multi-class) is presented in equation (17).

$$Accuracy(\%) = \frac{n_{11} + n_{22} + n_{33}}{N} \times 100\% \tag{17}$$

The ability of a classification system to recognize and categorize positive data into the positive class is known as sensitivity. Specificity is the degree of reliability of the model to detect negative labeled data correctly [49]. Since the response variable consists of three ordinal categories, sensitivity and specificity are calculated for each category individually using a one-vs-rest approach, ensuring a detailed evaluation of the model's performance across all DM statuses. The sensitivity and specificity formulas used for each category are presented in equations (18-20) [25],

Category 1
$$\begin{aligned} Sensitivity_{y_1} &= \frac{n_{11}}{n_{11} + n_{12} + n_{13}} = \frac{n_{11}}{n_1} \times 100\% \\ Specifity_{y_1} &= \frac{n_{22} + n_{23} + n_{32} + n_{33}}{n_{22} + n_{23} + n_{32} + n_{33} + n_{21} + n_{31}} \times 100\% \end{aligned} \tag{18}$$

Category 2
$$\begin{aligned} Sensitivity_{y_2} &= \frac{n_{22}}{n_{21} + n_{22} + n_{23}} = \frac{n_{22}}{n_2} \times 100\% \\ Specifity_{y_2} &= \frac{n_{11} + n_{13} + n_{31} + n_{33}}{n_{11} + n_{13} + n_{31} + n_{33} + n_{12} + n_{32}} \times 100\% \end{aligned} \tag{19}$$

Category 3
$$\begin{aligned} Sensitivity_{y_2} &= \frac{n_{33}}{n_{31} + n_{32} + n_{33}} = \frac{n_{33}}{n_3} \times 100\% \\ Specifity_{y_2} &= \frac{n_{11} + n_{12} + n_{21} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22} + n_{13} + n_{23}} \times 100\% \end{aligned} \tag{20}$$

3. RESULT AND DISCUSSION

3.1 Characteristic of Data

The descriptive statistic table between the variables age, BMI, and LDL cholesterol viewed from the DM status variable can be seen in Table 3.

Table 3. Descriptive Statistics of Age, BMI, and LDL Cholesterol Based on DM Status

Variable	Status	Total	Minimum	Maximum	Mean	Variance
Age	Pre-DM	267	18	61	42.007	93.113
	Stage-1 DM	165	27	65	47.927	108.861
	Stage-2 DM	232	31	71	51.418	75.751
BMI	Pre-DM	267	18.12	40.27	30.621	16.944
	Stage-1 DM	165	20.29	41.23	31.848	21.814
	Stage-2 DM	232	23.92	56	34.002	29.027
LDL Cholesterol	Pre-DM	267	31.2	186	121.38	892.89
	Stage-1 DM	165	58	202	142.16	1368.84
	Stage-2 DM	232	79	291	153.76	1286.51

Table 3 presents the descriptive statistics of age, BMI, and LDL cholesterol stratified by DM status. The data reveals a clear positive monotonic trend, where the mean values of age, BMI, and LDL cholesterol escalate in tandem with the severity of Diabetes Mellitus. Notably, Stage-2 DM patients exhibit the highest mean values for both BMI and LDL cholesterol, suggesting a significant increase in these metabolic parameters as the disease progresses. Furthermore, due to the variance in LDL levels which peaks at Stage-1 DM (1368.84), it can be stated that the data is heterogeneous. This heterogeneity necessitates a non-linear modeling approach to effectively capture the complex and non-linear risk patterns that standard parametric models typically assume away. Additionally, the distribution of nominal gender variables and ordinal hypertension status is further detailed in the contingency table presented in Table 4, providing a comprehensive overview of the patients' clinical characteristics.

Table 4. Contingency Table of Gender and Hypertension Status Variable

Variable		DM Status			Total
		Pre-DM	Stage-1 DM	Stage-2 DM	
Gender	Male	98 (31.8%)	80 (25.9%)	131 (42.4%)	309
	Female	169 (47.6%)	85 (23.9%)	101 (28.5%)	355
Hypertension Status	Pre-Hypertension	74 (58.3%)	31 (24.4%	22 (17.3%)	127
	Stage-1 Hypertension	179 (46.5%)	87 (22.6%)	119 (30.9%)	385
	Stage-2 Hypertension	14 (9.2%)	47 (31%)	91 (59.8%)	152

Table 4 illustrates the distribution of gender and hypertension status across the DM categories. The analysis reveals a distinct associative pattern where the severity of hypertension aligns with the progression of DM stages. Specifically, while pre-hypertension is more prevalent in the pre-DM group, a significant shift is observed as patients progress; stage-2 DM patients are more likely to exhibit stage-2 hypertension, suggesting a co-dependent escalation of vascular and metabolic risks. Gender distribution also shows a higher proportion of females in the pre-DM category, though the overall sample remains relatively balanced. These findings indicate that hypertension status, more significantly than gender, serves as a progressive indicator of diabetes severity within the studied population.

3.2 Estimation Process

Parameter estimation is carried out for the NOLR model, which is applied to the DM risk case. In obtaining the best MARS model, the lowest GCV value is determined. The basic function (BF) is a function that is defined in each region. The basic function (BF), which is frequently used, is two to four times the number of predictor variables. Five predictor variables are thought to have an impact on DM status in this study. As a result, 10, 15, and 20 BF will be merged to produce variables. The maximum interaction (MI) is the number of interactions that can occur in the model. MI of 1, 2, and 3 were used in this study. This investigation employed MI of 2 and 3 [29], [31]. If the MI used is 2, there can be interaction between variables in the model of up to 2. Likewise, if the MI used is 3, there can be interaction between variables in the model of up to 3. Minimum observation (MO) is the number of observations between the minimum knots. The MO used in this study are 0, 1, 2, and 3 because above that value, the GCV increases. To obtain the NOLR-MARS model, in-sample data from 598 patients (90%) were used, while out-sample data from 66 patients (10%) were used to evaluate the model that had been formed. The combined results of BF, MI, and MO in this research have been shown in Table 5.

Table 5. Combination of Pruning MARS in The Case of DM Status Model

Model Number	BF	MI	MO	$\mathbf{R^2}$	GCV
[1,]	10	2	1	0.5079772	0.0523525
[2,]	10	2	2	0.5077737	0.0523679
[3,]	10	2	3	0.5004233	0.0524664
[4,]	10	3	1	0.5099973	0.0521997
[5,]	10	3	2	0.5085140	0.0523119
[6,]	10	3	3	0.5003989	0.0524682
[7,]	15	2	1	0.5079772	0.0523525
[8,]	15	2	2	0.5077737	0.0523679
[9,]	15	2	3	0.5004233	0.0524664
[10,]	15	3	1	0.5171671	0.0521098
[11,]	15	3	2	0.5156834	0.0522231
[12,]	15	3	3	0.5003989	0.0524682
[13,]	20	2	1	0.5079772	0.0523525

Model Number	BF	MI	MO	R^2	GCV
[14,]	20	2	2	0.5077737	0.0523679
[15,]	20	2	3	0.5004233	0.0524664
[16,]*	20	3	1	0.5171671	0.0521008
[17,]	20	3	2	0.5156834	0.0522231
[18,]	20	3	3	0.5003989	0.0524682

According to Table 5, the best MARS model was chosen with GCV values of 0.0521008 and an R^2 value of 51.71671%. The best MARS model is represented by model number 16, which originally had BF = 20, MI = 3, and MO = 1. However, after performing basis function pruning to optimize model performance and reduce complexity, the final model included 12 active basis functions. This pruning was done based on the Generalized Cross-Validation (GCV) score, which showed that the model with 12 active basis functions had the lowest GCV, thus ensuring the best balance between predictive accuracy and model complexity. The best MARS model obtained, after pruning, is written in equation (21):

$$\begin{aligned}\hat{f}(x) = 11.716 + 0.4542BF_1 - 0.01699BF_2 - 2.065BF_3 + 1.706BF_4 + 1.046BF_5 \\ +0.03262BF_6 + 0.06025BF_7 - 0.1526BF_8 - 0.03189BF_{10} + 0.05801BF_{12}\end{aligned} \quad (18)$$

with,

$BF_1 = h(X_1 - 52)$; $BF_5 = h(X_5 = 3)$; $BF_9 = h(X_5 = 1)$;
$BF_2 = h(54 - X_1)$; $BF_6 = h(X_3 - 30.12)BF_1$; $BF_{10} = h(51 - X_1)BF_{11}BF_9$;
$BF_3 = h(56 - X_1)$; $BF_7 = h(X_1 - 51)BF_5$; $BF_{11} = h(51.8 - X_3)$;
$BF_4 = h(X_1 - 55)$; $BF_8 = h(94 - X_4)BF_9$; $BF_{12} = h(X_4 - 102)BF_7$

In interpreting the optimal NOLR-MARS model, we analyze the basis functions through the lens of cumulative logits. For instance, Equation (22) defines BF_4 as a piecewise linear spline denoted by the hinge function *h*:

$$BF_4 = h(X_1 - 55) \quad (19)$$

This basis function identifies a mathematical nodal point at age 55, acting as a non-linear threshold. In the context of the cumulative logit link function, the coefficient for BF_4 (1.706) signifies that for patients older than 55, each additional year of age increases the log-odds of transitioning to a more severe DM stage by 1.706. This corresponds to an odds ratio of $exp(1.706) = 5.51$, indicating a sharp increase in the risk profile—a nuance that standard linear models, which assume a constant risk across all ages, fail to capture. Biologically, this threshold at 55 aligns with the physiological decline in insulin efficiency [50], [51]. While traditional models suggest a gradual risk increase, the NOLR-MARS model mathematically demonstrates that the risk increases abruptly after the age of 55, providing more precise clinical insights. Higher-order interactions are further captured through the following product basis functions in equation (23)

$$\begin{aligned}BF_8 &= h(94 - X_4)BF_9 \\ BF_9 &= h(X_5 = 1)\end{aligned} \quad (20)$$

The basis function BF_8 is active if the patient has LDL cholesterol less than 94 mg/dL and a pre-hypertension history. Each one-unit increase in BF_8 increases the odds of transitioning to a higher DM stage by a factor of $exp(0.1526) = 1.1649$, assuming that other basic functions (BF) included in the model remain constant. BF_8 represents the interaction between two predictor variables: LDL cholesterol and hypertension status, with a mathematical knot identified at 94 mg/dL. Cholesterol management is a complex issue in the human body, as elevated LDL cholesterol contributes to the development of metabolic conditions such as DM, while hypertension further complicates the risk of progression to more severe stages of DM. The average value of cholesterol and LDL cholesterol levels in the diabetes group is higher than in the non-diabetes group [52], [53]. Thus, stage-2 diabetes is associated with low cholesterol absorption and increased cholesterol synthesis [54].

$$\begin{aligned}BF_{12} &= h(X_4 - 102)BF_7 \\ BF_7 &= h(X_1 - 51)BF_5 \\ BF_5 &= h(X_5 = 3)\end{aligned} \quad (21)$$

Similarly, the BF_{12} coefficient is active if the patient has LDL cholesterol levels greater than 102 mg/dL, an age greater than 51 years, and a history of stage-2 hypertension. Each one-unit increase in BF_{12} increases the odds of transitioning to a higher DM stage by a factor of $exp(0.05801) = 1.0597$, assuming that other basic functions (BF) in the model remain constant. BF_{12} represents the interaction of three predictor variables: LDL cholesterol, age, and hypertension status. A study with DM patients in Benghazi showed that DM patients had a history of hypertension, which was associated with advanced age and being overweight [55]. In addition, hypertension is prevalent in stage-2 DM, suggesting that insulin resistance may play an essential role in the pathogenesis of hypertension [6].

In the best MARS model, there are 4 (four) predictor variables that influence the risk of DM, namely age, BMI, LDL cholesterol and hypertension status. Then, Equation (11) and Equation (21) give the following representation of equation (25-26) of the logit link function,

$$\hat{g}_1(x) = 0.88748 + 0.4542BF_1 - 0.01699BF_2 - 2.065BF_3 + 1.706BF_4 + 1.046BF_5 + 0.03262BF_6 + 0.06025BF_7 - 0.1526BF_8 - 0.03189BF_{10} + 0.05801BF_{12} \quad (22)$$

$$\hat{g}_2(x) = 1.66579 + 0.4542BF_1 - 0.01699BF_2 - 2.065BF_3 + 1.706BF_4 + 1.046BF_5 + 0.03262BF_6 + 0.06025BF_7 - 0.1526BF_8 - 0.03189BF_{10} + 0.05801BF_{12} \quad (23)$$

We derive the estimated probability models as follows using the estimated models in Equation (25) and Equation (26). The estimated probability models for each DM status are presented in Equations (27-29).

$$\pi_1(x) = \frac{exp(\hat{g}_1(x))}{1 + exp(\hat{g}_1(x))} \quad (24)$$

Probability of Stage-1 DM Status

$$\pi_2(x) = \frac{exp(\hat{g}_2(x))}{1 + exp(\hat{g}_2(x))} \quad (25)$$

Probability of Stage-2 DM Status

$$\pi_3(x) = \frac{1}{1 + exp(\hat{g}_2(x))} \quad (26)$$

In this study, we only provide one example of model interpretation discussion using one of the test results, a 53-year-old patient with a body mass index of 45.04, LDL cholesterol of 212 mg/dL and stage-2 hypertension status. Meanwhile, the same computing process is applied to different testing data. Furthermore, we obtain the following values for the estimated probability models,

$$\pi_1(x) = \frac{exp(-1.6181)}{1 + exp(-1.6181)} = 0.16544 \quad (27)$$

$$\pi_2(x) = \frac{exp(-0.8398)}{1 + exp(-0.8398)} = 0.301686 \quad (28)$$

$$\pi_3(x) = 1 - \pi_1(x) - \pi_2(x) = 0.532874 \quad (29)$$

Based on Equations (30-32), a 53 years old patient with a BMI of 45.04, LDL of 212 mg/dL, and stage-2 hypertension shows the highest probability for stage-2 DM (53.29%). Consequently, the patient was diagnosed with stage-2 DM diagnosis, based on the maximum likelihood in the in-sample data. The NOLR-MARS framework is mathematically justified by its ability to capture complex non-linear thresholds and high-order interactions such as those identified in the Basis Functions which standard parametric models often fail to detect. Model robustness is further validated in Table 5, where Model 16 emerged as optimal, achieving the lowest GCV (0.0521008) and a high Nagelkerke Pseudo-R^2 of 51.71671%. This Pseudo-R^2 value is specifically utilized as a measure of goodness-of-fit for ordinal outcomes, indicating that the selected predictors explain a substantial portion of the model's deviance. These metrics, alongside strong out-sample accuracy, confirm the model's superior performance in capturing intricate DM risk patterns compared to traditional linear methods.

3.3 Evaluation of Classification

Classification accuracy measurement plays an important role in evaluating the performance of the NOLR-MARS model produced. A model with high classification accuracy lies in its ability to measure how well the NOLR-MARS model can classify the observed data. This provides a measure of goodness of fit and helps assess the predictive power of the model.

The classification value between the actual observation value and the anticipated value derived from the developed model is then determined. The ability of a system or process to classify data accurately is known as classification accuracy. NOLR-MARS classification accuracy can be assessed using a confusion matrix made up of actual observations and anticipated outcomes. Table 6 describes the confusion matrix.

Table 6. Confusion Matrix and Accuracy

Data	Observation	Prediction			Total	Accuracy
		Pre-DM	Stage-1 DM	Stage-2 DM		
In-sample	Pre-DM	268	62	2	332	74.92%
	Stage-1 DM	25	98	5	128	
	Stage-2 DM	36	20	82	138	
Out-sample	Pre-DM	27	3	2	32	80.30%
	Stage-1 DM	3	19	2	24	
	Stage-2 DM	1	2	7	10	

The performance of the NOLR-MARS model is evaluated through two data partitions: in-sample data for model construction and out-sample data for evaluating the developed model, using a confusion matrix to determine its classification accuracy, providing a measure of goodness of fit and predictive power [56]. As shown in Table 6, the model demonstrates high diagnostic consistency, where the in-sample results reflect a strong goodness-of-fit in capturing complex metabolic patterns during the calibration phase. More importantly, the sustained accuracy in the out-sample phase serves as a critical measure of the model's generalizability, confirming that the identified basis functions are robust and not prone to overfitting when applied to unseen clinical data.

This predictive performance offers a significant advantage over established benchmarks, such as standard Logistic Regression [16] and bi-response probit models [14], which typically struggle with the non-linear thresholds and high-order interactions inherent in diabetes progression. By replacing the traditional linear predictor with MARS-based basis functions, the NOLR-MARS framework maintains its superior predictive power while preserving the essential ordinal structure of the disease stages. To ensure a balanced detection of chronic disease stages, further evaluation of sensitivity and specificity is presented in the following tables 7.

Table 7. Sensitivity and Specificity

Data	Status	Sensitivity	Confidence Interval (95%)	Specificity	Confidence Interval (95%)
In-sample	Pre-DM	59.82%	(54.55% - 65.09%)	77.06%	(72.54% - 81.58%)
	Stage-1 DM	76.56%	(69.21% - 83.91%)	82.55%	(75.97% - 89.13%)
	Stage-2 DM	59.42%	(51.23% - 67.61%)	98.49%	(96.46% - 100%)
Out-sample	Pre-DM	50.94%	(33.61% - 68.27%)	88.23%	(77.07% - 99.39%)
	Stage-1 DM	79.17%	(62.91% - 95.43%)	88.1%	(75.83% - 100.00%)
	Stage-2 DM	70%	(41.64% - 98.36%)	92.86%	(79.39% - 100.00%)

The balance between sensitivity and specificity in Table 7, alongside their reported confidence intervals, further validates the model's reliability for clinical screening. The narrow confidence intervals in the in-sample data reflect the model's high precision, while the relatively wider intervals for the out-sample data, particularly in Stage-2 DM are primarily attributed to the smaller sample size, yet still maintain a consistent performance trend. The inclusion of these uncertainty ranges provides a transparent assessment of the model's stability across different DM stages. This performance is clinically essential to minimize false negatives, which in diabetes screening can lead to delays in diagnosis and unchecked hyperglycemia. Unlike local linear nonparametric approaches that often show weaker interaction capture in high-dimensional metabolic data [17], the NOLR-MARS model effectively manages this complexity through its 12 pruned basis functions. The robustness of the MARS algorithm in capturing non-linear patterns has been well-documented in various high-complexity modeling studies, where it consistently demonstrates superior predictive performance compared to traditional regression models. This proven mathematical flexibility further justifies its application in clinical modeling, particularly for capturing the intricate metabolic thresholds found in DM progression [57], [58].

Clinically, the proposed framework addresses the limitations of standard screening by providing a robust, non-linear approach to capture complex risk patterns. It offers a significant improvement over black-box ensemble structures through its mathematical interpretability [59], allowing clinicians to trace the logic behind each risk prediction. However, it is important to emphasize that since this study utilizes a NAFLD-origin cohort, the primary threat to external validity remains its generalizability. Specifically, this introduces a potential selection bias, as patients with NAFLD may exhibit more advanced metabolic dysfunction compared to the general population. Furthermore, the cohort's temporal window (1997–2014) represents a limitation, as it may not fully capture contemporary lifestyle shifts or the most recent diagnostic evolutions in diabetes management. Therefore, future research is highly recommended to validate this framework using modern, primary, general population Diabetes Mellitus datasets to confirm these non-linear metabolic patterns across broader and more current clinical settings.

4. CONCLUSION

This study integrates the Multivariate Adaptive Regression Splines (MARS) into an Ordinal Logistic Regression framework (NOLR-MARS) to model complex diabetes mellitus (DM) risk stages. By replacing the linear predictor with 12 pruned basis functions (GCV=0.0521) and training on 598 in-sample subjects, the model effectively captures nonlinear metabolic interactions and high-order thresholds, specifically involving age, BMI, LDL cholesterol, and hypertension. Methodologically, this approach achieved a robust 80.3% predictive accuracy when tested on 66 out-sample subjects, demonstrating its superiority in handling metabolic data complexity compared to traditional parametric methods. Based on these findings, we recommend that the Indonesian Ministry of Health pilot NOLR-MARS-based risk stratification in community health centers (Puskesmas) serving populations with high metabolic syndrome prevalence to prioritize early intervention resources. A significant limitation of this research is the reliance on a NAFLD-specific cohort, which serves as a primary constraint on the model's generalizability. Therefore, while these findings provide a rigorous mathematical basis for clinical screening, future work must prioritize external validation using primary, general population datasets to confirm the model's broader clinical applicability. In the spirit of open science, the research code and a simplified dataset are available upon request to facilitate replication and further development.

5. REFERENCES

[1] N. J. Sathi, Md. A. Islam, Md. S. Ahmed, and S. M. S. Islam, “Prevalence, trends and associated factors of hypertension and diabetes mellitus in Bangladesh: Evidence from BHDS 2011 and 2017–18,” *PLoS One*, vol. 17, no. 5, p. e0267243, May 2022, doi: 10.1371/journal.pone.0267243.

[2] P. K. Roy, M. H. R. Khan, T. Akter, and M. S. Rahman, “Exploring socio-demographic-and geographical-variations in prevalence of diabetes and hypertension in Bangladesh: Bayesian spatial analysis of national health survey data,” *Spat. Spatiotemporal Epidemiol.*, vol. 29, pp. 71–83, Jun. 2019, doi: 10.1016/j.sste.2019.03.003.

[3] L. S. B. Dahman, A. M. Daakeek, H. S. Alghazali, A. M. Kaity, and M. S. Obbed, “Diagnosis and Classification of Diabetes Mellitus,” *Diabetes Care*, vol. 37, no. Supplement_1, pp. S81–S90, Jan. 2014, doi: 10.2337/dc14-S081.

[4] N. Susanti, A. H. Nazli, D. Wahyuni, and W. Y. yasmin, “Faktor – Faktor yang Mempengaruhi Diabetes Mellitus di Puskesmas Tuntungan,” *PREPOTIF : Jurnal Kesehatan Masyarakat*, vol. 8, no. 2, pp. 4293–4299, Aug. 2024.

[5] S. Sunarti *et al.*, “Peningkatan Pengetahuan Masyarakat tentang Mitos dan Fakta Diabetes Melitus serta Diet Seimbang Penderita Diabetes Melitus,” *Jurnal Pengabdian Masyarakat: Darma Bakti Teuku Umar*, vol. 5, no. 2, p. 64, Jan. 2024, doi: 10.35308/baktiku.v5i2.7230.

[6] G. Jia and J. R. Sowers, “Hypertension in Diabetes: An Update of Basic Mechanisms and Clinical Disease,” *Hypertension*, vol. 78, no. 5, pp. 1197–1205, Nov. 2021, doi: 10.1161/HYPERTENSIONAHA.121.17981.

[7] C. C. Rani and N. S. Mulyani, “Faktor-faktor yang berhubungan dengan kejadian diabetes mellitus tipe-II pada pasien rawat jalan,” *Jurnal SAGO Gizi dan Kesehatan*, vol. 2, no. 2, p. 122, Sep. 2021, doi: 10.30867/gikes.v2i2.258.

[8] G. Li *et al.*, “Incidence and Risk Factors of Gestational Diabetes Mellitus: A Prospective Cohort Study in Qingdao, China,” *Front. Endocrinol. (Lausanne).*, vol. 11, Sep. 2020, doi: 10.3389/fendo.2020.00636.

[9] K. P. Irjayanti, S. Zaenal, and Suhartatik, “Faktor-Faktor yang Mempengaruhi Terjadinya Peningkatan Diabetes Melitus Tipe 2,” *Jurnal Ilmiah Mahasiswa & Penelitian Keperawatan*, vol. 1, no. 6, pp. 805–813, 2022.

[10] F. R. Mahmud, Sudirman, and N. Afni, “Faktor-Faktor Yang Berhubungan Dengan Penyakit Diabetes Melitus di Ruang Poli Interna Rsud Mokopido Kabupaten Tolitoli,” *Jurnal Kolaboratif Sains*, vol. 1, no. 1, pp. 168–175, Nov. 2018.

[11] S. Alam, Md. K. Hasan, S. Neaz, N. Hussain, Md. F. Hossain, and T. Rahman, “Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management,” *Diabetology*, vol. 2, no. 2, pp. 36–50, Apr. 2021, doi: 10.3390/diabetology2020004.

[12] N. K. Hasibuan, S. Dur, and I. Husein, “Faktor Penyebab Penyakit Diabetes Melitus dengan Metode Regresi Logistik,” *G-Tech: Jurnal Teknologi Terapan*, vol. 6, no. 2, pp. 257–264, Sep. 2022, doi: 10.33379/gtech.v6i2.1696.

[13] B. Lestari, Fatmawati, and I. N. Budiantara, “Smoothing Spline Estimator in Multiresponse Nonparametric Regression for Predicting Blood Pressure and Heart Rate,” *International Journal of Academic and Applied Research (IJAAR)*, vol. 3, no. 9, pp. 1–8, Sep. 2019.

[14] Suliyanto and M. Rifada, “Modeling of Risk for Diabetes Mellitus and Hypertension Using Bi-response Probit Regression,” in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, Singapore: Springer Singapore, 2019, pp. 383–389. doi: 10.1007/978-981-13-7279-7_47.

[15] H. Al-Rimmawi, “Prediction of Type 2 Diabetes using logistic regression techniques,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 15, no. 1, Jan. 2024, doi: 10.61841/turcomat.v15i1.13875.

[16] R. D. Joshi and C. K. Dhakal, “Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[17] W. A. Anam, A. Massaid, N. A. Amesya, and N. Chamidah, “Modeling of diabetes mellitus risk based on consumption of salt, sugar, and fat factors using local linear estimator,” in *AIP Conference Proceedings 2264*, AIP Publishing, Sep. 2020, p. 030009. doi: 10.1063/5.0023498.

[18] W. O. R. Alifu, R. Andriani, and W. Ode, “Faktor- Faktor Yang Berhubungan dengan Kejadian Diabetes Melitus di Wilayah Kerja Puskesmas Sampolawa Kabupaten Buton Selatan,” *Kampurui Jurnal Kesehatan Masyarakat (The Journal of Public Health)*, vol. 2, no. 2, pp. 6–12, Dec. 2020, doi: 10.55340/kjkm.v2i2.228.

[19] R. A. Nugroho, Tarno, and A. Prahutama, "Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode Smooth Support Vector Machine (Ssvm)," *Jurnal Gaussian*, vol. 6, no. 3, pp. 439–448, 2018.

[20] K. Gholipour, M. Asghari-Jafarabadi, S. Iezadi, A. Jannati, and S. Keshavarz, "Modelling the prevalence of diabetes mellitus risk factors based on artificial neural network and multiple regression," *Eastern Mediterranean Health Journal*, vol. 24, no. 08, pp. 770–777, Aug. 2018, doi: 10.26719/emhj.18.012.

[21] P. McCullagh, "Regression Models for Ordinal Data," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 42, no. 2, pp. 109–127, Jan. 1980, doi: 10.1111/j.2517-6161.1980.tb01109.x.

[22] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*. CRC Press, 1999. doi: 10.1201/9781482273144.

[23] W. Härdle, *Applied Nonparametric Regression*. Cambridge University Press, 1990. doi: 10.1017/CCOL0521382483.

[24] M. Hasyim and D. D. Prastyo, "Modelling lecturer performance index of private university in Tulungagung by using survival analysis with multivariate adaptive regression spline," *J. Phys. Conf. Ser.*, vol. 974, p. 012065, Mar. 2018, doi: 10.1088/1742-6596/974/1/012065.

[25] W. K. Härdle, L. Simar, and M. R. Fengler, *Applied Multivariate Statistical Analysis*, 6th ed. Springer, 2024.

[26] N. Chamidah, B. Lestari, H. Susilo, M. Y. Alsagaff, I. N. Budiantara, and D. Aydin, "Spline Estimator in Nonparametric Ordinal Logistic Regression Model for Predicting Heart Attack Risk," *Symmetry (Basel).*, vol. 16, no. 11, p. 1440, Oct. 2024, doi: 10.3390/sym16111440.

[27] M. Hasyim *et al.*, "Bootstrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) to Analyse the Lecturer Research Performance in Private University," *J. Phys. Conf. Ser.*, vol. 1114, p. 012117, Nov. 2018, doi: 10.1088/1742-6596/1114/1/012117.

[28] J. H. Friedman and C. B. Roosen, "An introduction to multivariate adaptive regression splines," *Stat. Methods Med. Res.*, vol. 4, no. 3, pp. 197–217, Sep. 1995, doi: 10.1177/096228029500400303.

[29] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, Mar. 1991, doi: 10.1214/aos/1176347963.

[30] J. H. Friedman, "Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines. Technical Report No. 108," Stanford, 1991.

[31] J.-H. Kim, "MARS Modeling for Ordinal Categorical Response Data: A Case Study," *Commun. Stat. Appl. Methods*, vol. 7, no. 3, pp. 711–720, Dec. 2000.

[32] A. Wibowo, "Pemodelan Mars dan Regresi Logistik Rumah Tangga Miskin Kalimantan Tengah Tahun 2016," *JMPM: Jurnal Matematika dan Pendidikan Matematika*, vol. 3, no. 1, p. 1, Mar. 2018, doi: 10.26594/jmpm.v3i1.1023.

[33] 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, pp. 1417–1427, 2015, doi: 10.12988/ams.2015.5127.

[34] S. D. P. Yasmirullah, B. W. Otok, J. D. T. Purnomo, and D. D. Prastyo, "Modification of Multivariate Adaptive Regression Spline (MARS)," *J. Phys. Conf. Ser.*, vol. 1863, no. 1, p. 012078, Mar. 2021, doi: 10.1088/1742-6596/1863/1/012078.

[35] O. E. Agwu, S. Alatefi, A. Alkouh, and R. R. Suppiah, "Modelling the Flowing Bottom Hole Pressure of Oil and Gas Wells Using Multivariate Adaptive Regression Splines," *J. Pet. Explor. Prod. Technol.*, vol. 15, no. 2, p. 22, Feb. 2025, doi: 10.1007/s13202-025-01933-9.

[36] S. Keawsawasvong, K. Kounlavong, N. T. Duong, V. Q. Lai, V. N. Khatri, and A. Eskandarinejad, "Seismic Stability Assessment of Rock Slopes Using Multivariate Adaptive Regression Splines," *Transportation Infrastructure Geotechnology*, vol. 11, no. 4, pp. 2296–2318, Aug. 2024, doi: 10.1007/s40515-024-00374-x.

[37] M. B. Jumber, M. T. Damtie, and D. Tegegne, "Integration of Multivariate Adaptive Regression Splines and Weighted Arithmetic Water Quality Index Methods for Drinking Water Quality Analysis," *Water Conservation Science and Engineering*, vol. 9, no. 1, p. 6, Jun. 2024, doi: 10.1007/s41101-024-00239-x.

[38] A. G. P. Varshini and K. A. Kumari, "Software Effort Estimation Using Stacked Ensemble Technique and Hybrid Principal Component Regression and Multivariate Adaptive Regression Splines," *Wirel. Pers. Commun.*, vol. 134, no. 4, pp. 2259–2278, Feb. 2024, doi: 10.1007/s11277-024-11010-9.

[39] A. M. Allen, H. K. Van Houten, L. R. Sangaralingham, J. A. Talwalkar, and R. G. McCoy, "Healthcare Cost and Utilization in Nonalcoholic Fatty Liver Disease: Real-World Data From a Large U.S. Claims Database," *Hepatology*, vol. 68, no. 6, pp. 2230–2238, Dec. 2018, doi: 10.1002/hep.30094.

[40] C. D. Byrne and G. Targher, "NAFLD: A Multisystem Disease," *J. Hepatol.*, vol. 62, no. 1, pp. S47–S64, Apr. 2015, doi: 10.1016/j.jhep.2014.12.012.

[41] H. Tilg, A. R. Moschen, and M. Roden, "NAFLD and Diabetes Mellitus," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 14, no. 1, pp. 32–42, Jan. 2017, doi: 10.1038/nrgastro.2016.147.

[42] Zakariyah and I. Zain, "Analisis Regresi Logistik Ordinal pada Prestasi," *Jurnal Sains dan Seni ITS*, vol. 4, no. 1, pp. 121–126, 2015.

[43] M. Rifada, N. Chamidah, and R. A. Ningrum, "Estimation of Nonparametric Ordinal Logistic Regression Model using Generalized Additive Models (GAM) Method Based on Local Scoring Algorithm," in *AIP Conference Proceedings*, 2022. doi: 10.1063/5.0111771.

[44] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.

[45] D. Ariesta, N. Gusriani, and K. Parmikanti, "Estimasi Parameter Model Regresi Nonparametrik B-Spline pada Angka Kematian Maternal," *Jurnal Matematika UNAND*, vol. 10, no. 3, pp. 342–354, Jul. 2021, doi: 10.25077/jmu.10.3.342-354.2021.

[46] X. Ju, V. C. P. Chen, J. M. Rosenberger, and F. Liu, "Fast knot optimization for multivariate adaptive regression splines using hill climbing methods," *Expert Syst. Appl.*, vol. 171, p. 114565, Jun. 2021, doi: 10.1016/j.eswa.2021.114565.

[47] E. K. Koc and C. Iyigun, "Restructuring forward step of MARS algorithm using a new knot selection procedure based on a mapping approach," *Journal of Global Optimization*, vol. 60, no. 1, pp. 79–102, Sep. 2014, doi: 10.1007/s10898-013-0107-5.

[48] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Pearson Prentice Hall, 2007.

[49] A. Miladitiya, "Sensitivitas dan Spesifisitas Lingkar Pinggang Dalam Mengidentifikasi Kelebihan Berat Badan Dan Obesitas Pada Wanita Dewasa," *Interest : Jurnal Ilmu Kesehatan*, vol. 7, no. 1, May 2018, doi: 10.37341/interest.v7i1.64.

[50] H. Kim, "Evidence on the Optimal Capital Structure for Korea Composite Stock Price Index (KOSPI)-listed Firms: Employing Inter- and Intra-industry Analyses," *Korea Open Acces Journal (KOAJ)*, vol. 13, no. 2, pp. 173–191, Apr. 2017, doi: 10.16980/jitc.13.2.201704.173.

[51] K. Komariah and S. Rahayu, "Hubungan Usia, Jenis Kelamin dan Indeks Massa Tubuh dengan Kadar Gula Darah Puasa pada Pasien Diabetes Melitus Tipe 2 di Klinik Pramtama Rawat Jalan Proklamasi, Depok, Jawa Barat," *Jurnal Kesehatan Kusuma Husada*, pp. 41–50, Jan. 2020, doi: 10.34035/jk.v11i1.412.

[52] Y.-X. Cao *et al.*, "The Longitudinal Association of Remnant Cholesterol with Cardiovascular Outcomes in Patients with Diabetes and Pre-Diabetes," *Cardiovasc. Diabetol.*, vol. 19, no. 1, p. 104, Dec. 2020, doi: 10.1186/s12933-020-01076-7.

[53] O. Castañer *et al.*, "Remnant Cholesterol, Not LDL Cholesterol, Is Associated With Incident Cardiovascular Disease," *J. Am. Coll. Cardiol.*, vol. 76, no. 23, pp. 2712–2724, Dec. 2020, doi: 10.1016/j.jacc.2020.10.008.

[54] R. Anggraini, "Korelasi Kadar Kolesterol Dengan Kejadian Diabetes Mellitus Tipe 2 Pada Laki-Laki," *Medical and Health Science Journal*, vol. 2, no. 2, Aug. 2018, doi: 10.33086/mhsj.v2i2.588.

[55] Y. Akalu and Y. Belsti, "Hypertension and Its Associated Factors Among Type 2 Diabetes Mellitus Patients at Debre Tabor General Hospital, Northwest Ethiopia," *Diabetes Metab. Syndr. Obes.*, vol. Volume 13, pp. 1621–1631, May 2020, doi: 10.2147/DMSO.S254537.

[56] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, May 2000, doi: 10.1093/bioinformatics/16.5.412.

[57] M. B. Jumber, M. T. Damtie, and D. Tegegne, "Integration of Multivariate Adaptive Regression Splines and Weighted Arithmetic Water Quality Index Methods for Drinking Water Quality Analysis," *Water Conservation Science and Engineering*, vol. 9, no. 1, p. 6, Jun. 2024, doi: 10.1007/s41101-024-00239-x.

[58] R. M. Adnan, Z. Liang, S. Heddam, M. Zounemat-Kermani, O. Kisi, and B. Li, "Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs," *J. Hydrol. (Amst).*, vol. 586, p. 124371, Jul. 2020, doi: 10.1016/j.jhydrol.2019.124371.

[59] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.