Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:13

https://doi.org/10.28919/cmbn/9049

ISSN: 2052-2541

HYPERTENSION MODELLING USING NONPARAMETRIC ORDINAL LOGISTIC REGRESSION BASED ON MULTIVARIATE ADAPTIVE REGRESSION SPLINE

MAYLITA HASYIM[1,2], NUR CHAMIDAH[3,4,*], TOHA SAIFUDIN[3,4]

[1]Doctoral Study Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia

[2]Study Program of Mathematics Education, Faculty of Social and Humanities, Universitas Bhinneka PGRI, Tulungagung 66221, Indonesia

[3]Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia

[4]Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia

Abstract: Hypertension is one of the most common diseases in the world and is an important risk factor for heart disease and a significant contributor to mortality and morbidity in both developed and developing countries. Systematic reviews have been conducted to assess the prevalence of hypertension and its risk factors. To model the hypertension status, in this study we developed an ordinal logistic regression model into a nonparametric regression model and called it a Nonparametric Ordinal Logistic Regression (NOLR) model. Therefore, this study aims to model hypertension status based on several influencing factors, and identify these factors that the most influential on hypertension status using the NOLR model approach, because we assume an ordinal scale response variable with q categories to have an asymmetric distribution, namely a multinomial distribution. Next, to estimate the NOLR model of Hypertension status, we use a Multivariate Adaptive Regression Spline (MARS) estimator, because it can

*Corresponding author

E-mail address: nur-c@fst.unair.ac.id

Received November 30, 2024

accommodate interactions between risk factors expressed in basis functions, making it suitable for high-dimensional data cases. Furthermore, selection of the best model is based on minimum value of Generalized Cross-Validation (GCV). The results are that the best MARS model of hypertension status has BF = 16, MI = 3, and MO = 1 with GCV value of 0.0353874, and R^2 value of 49.13508%. Also, there are three predictor variables, namely age, body mass index and total cholesterol that significantly affect the hypertension status. In addition, the obtained estimation of nonparametric ordinal logistic regression model using the MARS estimator is valid for predicting the hypertension status with an accuracy value of 69.25%, sensitivity value of 66.47% and specificity value of 84.06%.

Keywords: hypertension; nonparametric ordinal logistic regression; MARS; age; gender; BMI; cholesterol.
2020 AMS Subject Classification: 62G05, 62G08, 62P10, 65D07.

1. INTRODUCTION

The global epidemic of hypertension is mainly uncontrolled, and hypertension remains the leading cause of non-communicable disease deaths worldwide [1]. Approximately 22% of people worldwide who are 18 years of age or older suffer from high blood pressure. Only 8.8% of people with high blood pressure are aware that they have it, even though 34% of Indonesians aged 15 and over have high blood pressure, which is higher than the global average [2,3]. Hypertension is a preventable illness associated with demographic and lifestyle factors [4,5]. Some critical factors influencing hypertension are age, gender, body mass index, and cholesterol [6,7]. In this study, we analyze the relationship between hypertension status and several influencing factors, namely age, gender, body mass index, and cholesterol. In this study we use nonparametric regression for modeling the hypertension status influenced by age, gender, body mass index, and cholesterol, because the relationship between hypertension status and the influencing factors has no specific pattern. The estimation method used is Multivariate Adaptive Regression Spline (MARS), because it can obtain good predictive results for handling data that has changed behavior at certain sub-intervals, indicating a change in data behavior patterns [8,9]. The MARS method is needed to model hypertension status to obtain more accurate results.

Several previous researches on hypertension cases were performed by many researchers. For examples, Kurniawan et al. [3] developed and validated a hypertension risk-prediction model using machine learning; Andriani and Chamidah [7] have modeled hypertension risk factors using logistic regression, and the result showed that the logit link function is better than the gompit link function for modeling hypertension risk factors; Adiwati and Chamidah [10] analyzed hypertension risk factors using penalized spline, and found that the nonparametric

approach is better than the outcome, so it can be used to model the risk of hypertension;Rohkuswara and Syarif [11] analyzed the relationship between obesity and the incidence of stage-1 hypertension; Qu et al. [12] analyzed the relationship between BMI (body mass index) and hypertension using restricted cubic spline functions, and the result showed that there was a non-linear relationship between the continuous change of the BMI and hypertension; and Amalia et al. [13] used a penalized spline estimator approach to model the risk of hypertension based on parameters related to fat, sugar, and salt consumption, and the result showed that for modeling the risk of hypertension based on consumption of salt, sugar, and fat, the use of the penalized spline estimator of nonparametric logistic regression is better than that of logit link function of parametric logistic regression.

Those previous researches mentioned above analyzed risk factors using various methods, including spline. Splines are polynomials with different segments combined at vertices to adapt effectively to the local characteristics of the function or data [14–28]. However, the spline method used in the previous researches could be improved when it is applied to high-dimensional data with multi-predictors. So, the MARS can overcome this weakness. The MARS method is focused on overcoming the problems of high dimensions, has many variables, as well as a large sample size, so the MARS method can be used on high-dimensional data and produces accurate response predictions [8,29]. Another advantage of MARS is that it can accommodate interactions between predictor variables expressed in basis functions, making it suitable for high-dimensional data cases. The MARS is also ideal for calculating data classification accuracy cases that require response variables to be categorical [30,31]. This means that MARS method is very suitable for the data structure in this study, which is the response variable has a categorical scale.

Classification in the MARS is based on a logistic regression approach [8]. Logistic regression analyses the relationship between categorical response variables and categorical and continuous predictor variables [32,33]. By converting the logit to a logit link function form, the logistic regression equation is applied to a probability function. It is derived from the shape of the estimated probability function of a successful event or specific event occurring [32]. This link logit function is the MARS model, or is referred to as the MARS logit model. Researchers such as Wibowo and Mehrani [34], Permatasari et al. [35], and Meilisa [36] have studied MARS with binary category responses in modeling the classification cases. However, in this study the response variable has an ordinal scale so the binary MARS method cannot be used. This study aims to model the hypertension status based on several influencing factors and identify the

factors that most influence the hypertension status using nonparametric ordinal logistic regression (NOLR) model approach based on MARS estimator, and selection of the best model is performed based on minimum value of Generalized Cross-Validation (GCV). In order to lower the prevalence of hypertension in Indonesia, the findings of this study could be utilized to forecast the likelihood that a person with the condition will exhibit the risk factors.

2. MATERIALS AND METHODS

This section provides information about the dataset we used for the analysis and some literature reviews for modeling the data.

2.1. Dataset

The dataset of hypertension status Age, Gender, BMI, and Cholesterol are presented in Table 1.

Table 1. Dataset of Hypertension Status, Age, Gender, BMI, and Cholesterol.

Patient Number	Hypertension Status (Y)	Age (X_1)	Gender (X_2)	Body Mass Index (X_3)	Total Cholesterol (X_4)
1	1	44	1	35.56	170
2	1	55	1	30.78	170
3	1	53	2	34.95	190
⋮	⋮	⋮	⋮	⋮	⋮
488	2	60	2	40,35	183

Table 1 provides information regarding the dataset used in this study. This dataset consists of the hypertension status as a response variable, and age, gender, body mass index (BMI), and total cholesterol as predictor variables that are thought to influence the hypertension. The data were recorded from 488 patients. The hypertension status as response variable (Y) has an ordinal scale with category 1 as pre-hypertension, category 2 as stage-1 hypertension, and category 3 as stage-2 hypertension [37]. The age variable (X_1) has an interval scale. The gender variable (X_2) has a nominal scale with category 1 as male and category 2 as female. While the body mass index variable (X_3) and total cholesterol (X_4) have scale of ratio.

2.2. Ordinal Logistic Regression

When the response variable is polychotomous and has an ordinal scale, ordinal logistic regression is a regression used to examine the relationship between predictor and response variables [38]. The model that can be used for ordinal logistic regression is the cumulative

logit model, where the ordinal nature of the Y response is outlined in the cumulative probability. Therefore, comparing the cumulative probability that is less than or equal to the response variable category on predictor variables written in vector form **X** yields the cumulative logit model, and it can be written as $P(Y \le r|X_i)$ [33]. The cumulative probability, $P(Y \le r|X_i)$, is given by the following equation [33]:

$$P\left(Y \le r \mid X_i\right) = \pi\left(x_i\right) = \left(\frac{\exp\left(\beta_{0r} + \sum_{m=1}^{M} \beta_m x_{im}\right)}{1 + \exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)} \right) \tag{1}$$

where $x_i = \left(x_{i1}, x_{i2}, ..., x_{im}\right)$ is predictor variable of the *i*-th observation $\left(i = 1, 2, ..., n\right)$ for each *p* predictor variable, while $r = 1, 2, ..., R$ is the response variable category. The logit transformation is used to parse the ordinal logistic regression parameters to estimate them. The logit of $P\left(Y \le r \mid X_i\right)$ can be presented as follows [33]:

$$g_r\left(x\right) = \text{logit}\, P\left(Y \le r \mid X_i\right) = \ln\left(\frac{P\left(Y \le r \mid X_i\right)}{1 - P\left(Y \le r \mid X_i\right)} \right)$$

$$= \ln\left(\frac{\dfrac{\exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}}{1 - \dfrac{\exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}} \right) = \beta_{0r} + \sum_{k=1}^{p} \beta_k x_{ik} \tag{1}$$

where the value β_k for each $k = 1, 2, ..., m$ in each ordinal logistic regression model is the same. For example, if there are three response categories, i.e., $r = 1, 2, 3$ then the cumulative probability of *r* category response is given as follows [33]:

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

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

Based on Equations (3) and (4), the cumulative probability of each category of response variables is obtained, namely,

$$P\left(Y_r = 1 \mid x_i\right) = \pi_1\left(x\right) = \left(\frac{\exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)} \right) \tag{5}$$

$$\begin{aligned} P\left(Y_r = 2 \mid x_i\right) &= \pi_2\left(x\right) = P\left(Y \le 2 \mid x_i\right) - P\left(Y_r = 1 \mid x_i\right) \\ &= \frac{\exp\left(\beta_{02} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\beta_{02} + \sum_{k=1}^{p} \beta_k x_{ik}\right)} - \frac{\exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)}{1 + \exp\left(\beta_{01} + \sum_{k=1}^{p} \beta_k x_{ik}\right)} \end{aligned} \tag{6}$$

If Equation (2) has three response categories, i.e., $r = 1, 2, 3$, then the cumulative logit model for each response category can be expressed as follows:

$$\begin{aligned} \hat{g}_1\left(x\right) &= \ln\left(\frac{P\left(Y \le 1 \mid x\right)}{1 - P\left(Y \le 1 \mid x\right)} \right) = \ln\left(\frac{P\left(Y \le 1 \mid x\right)}{P\left(Y > 1 \mid x\right)} \right) \\ &= \ln\left(\frac{P\left(Y = 1 \mid x\right)}{P\left(Y = 2 \mid x\right) + P\left(Y = 3 \mid x\right)} \right) = \theta_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \\ \hat{g}_2\left(x\right) &= \ln\left(\frac{P\left(Y \le 2 \mid x\right)}{1 - P\left(Y \le 2 \mid x\right)} \right) = \ln\left(\frac{P\left(Y \le 2 \mid x\right)}{P\left(Y > 2 \mid x\right)} \right) \\ &= \ln\left(\frac{P\left(Y = 1 \mid x\right) + P\left(Y = 2 \mid x\right)}{P\left(Y = 3 \mid x\right)} \right) = \theta_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \end{aligned} \tag{7}$$

In this study we use the link logit function of ordinal logistic regression as the basis for classification in MARS. The following discussion focuses on MARS.

2.3. Multivariate Adaptive Regression Spline (MARS)

The Multivariate Adaptive Regression Spline (MARS) technique was developed in 1991 by Jerome H. and automates the process of creating accurate predictive models for continuous and binary response variables. One of the versatile techniques for modeling the

high-dimensional regression data is MARS. The number of basis functions is to be parameters of the MARS model which is an extension of the basis spline functions [4]. The following terminology must be taken into account in MARS [8]:

a. Knots, which are the points on a regression line that comprise a region of a regression function;
b. Basis Function (BF), which is a group of functions used to characterize the relationship between the predictor and the response variable; and
c. Interaction, is the interaction between variables and the maximum number of interactions (MI) 1, 2, and 3.

Furthermore, the MARS model is formulated as follows [8]:

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

where β_0 is main of basis function; β_m is coefficient of basis function; M is maximum of basis function (non-constant basis function); K_m is degree of interaction; s_{km} equals to 1 if the data is to the right of the knot point, and equals to -1 if the data is to the left of the knot point; $x_{v(k,m)}$ is predictor variable; and t_{km} is knot point of predictor variable $x_{v(k,m)}$.

The best model is selected based on the minimum value of the following GCV (Generalized Cross-Validation):

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

where y_i is response variable; $\hat{f}_M(x_i)$ is the estimated value of response variable on M basis functions; n is the number of observations; K_m is degree of interaction; $C(\hat{M}) = C(M) + dM$; d is a value when each basis function reaches the optimization that is $d = 2$ (for additive model) and $d = 3$ (for interaction model).

If more than one model has minimum value of GCV, the next step is to use the criteria for maximum value of the R^2 and the highest classification accuracy value [14]. Formula for the R^2 is presented as follows [14]:

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

where n is the number of observations; $f(x_i)$ is response value at *i*-th observation; $\hat{f}(x_i)$ is estimated response value at *i*-th observation; and $\bar{f}(x_i)$ is average value of response at *i*-th observation.

Furthermore, by considering equations (7) and (8), the MARS logit model with three ordinal categories of response variables can be expressed as follows [33]:

(11)
$$\begin{aligned}
\hat{g}_1(x) &= \ln\frac{P(Y\le 1)}{\left(1-P(Y\le 1)\right)} = \ln\frac{P(Y\le 1)}{P(Y>1)} \\
&= \ln\frac{P(Y=1)}{P(Y=2)+P(Y=3)} = \theta_1 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right] \\
\hat{g}_2(x) &= \ln\frac{P(Y\le 2)}{\left(1-P(Y\le 2)\right)} = \ln\frac{P(Y\le 2)}{P(Y>2)} \\
&= \ln\frac{P(Y=1)+P(Y=2)}{P(Y=3)} = \theta_2 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]
\end{aligned}$$

If there are three response categories, namely $r = 1,2,3$, then the cumulative probability of the r_i category response can be expressed as follows [32,38]:

(12)
$$\begin{cases}
P(Y\le 1\,|\,x_i) = \dfrac{\exp\left(\theta_1 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)}{1+\exp\left(\theta_1 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)} \\
P(Y\le 2\,|\,x_i) = \dfrac{\exp\left(\theta_2 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)}{1+\exp\left(\theta_2 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)}
\end{cases}$$

Equations in (12) yield the cumulative probability for each category of response variables as follows:

(2) $$P(Y_r = 1\,|\,x_i) = \pi_1(x) = \frac{\exp\left(\theta_1 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)}{1+\exp\left(\theta_1 + \left[\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)} - t_{km}\right)\right]\right]\right)}$$

$$P\left(Y_r=2\mid x_i\right)=\pi_2\left(x\right)=P\left(Y\le 2\mid x_i\right)-P\left(Y_r=1\mid x_i\right)$$

$$=\frac{\exp\left(\theta_2+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right]\right)}{1+\exp\left(\theta_2+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right]\right)} \quad (3)$$

$$-\frac{\exp\left(\theta_1+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right]\right)}{1+\exp\left(\theta_1+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right]\right)}$$

$$P\left(Y_r=3\mid x_i\right)=\pi_3\left(x\right)=1-\pi_1\left(x\right)-\pi_2\left(x\right) \quad (4)$$

Finally, for $\pi_1\left(x\right)=P\left(Y=1\right);\pi_2\left(x\right)=P\left(Y=2\right);\pi_3\left(x\right)=P\left(Y=3\right)$, then we obtain:

$$\ln\frac{\pi_1}{\pi_2+\pi_3}=\theta_1+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right] \quad (5)$$

$$\ln\frac{\pi_1+\pi_2}{\pi_3}=\theta_2+\left[\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_M}\left[s_{km}.\left(x_{v(k,m)}-t_{km}\right)\right]\right]$$

2.4. Evaluation of Classification Procedures

Validation is carried out using testing data information to determine whether the strength of the logistic regression model can predict data outside the model. Then a comparison is made between the model prediction results and their actual grouping, so a classification table (confusion matrix) can be made. The elements of the confusion matrix are utilized to find three model strengths: accuracy, sensitivity, and specificity. If the three measurements are close to 1, then it can be stated that the classification has been done well [39]. The confusion matrix is presented in Table 2 [40].

Table 2. Confusion matrix 3*x*3

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

Based on the confusion matrix in Table 2, three values of model strength can be obtained: accuracy, sensitivity, and specificity. Accuracy is a metric used to measure the extent to which a classification system that has been built can classify accurately, both for data that has positive and negative labels. The higher the accuracy value obtained, the better the classification system makes predictions [41]. The maximum value of accuracy is 100%. The formula for calculating accuracy in the case of classification with several classes (multi-class) can be found in Equation (10) [40].

$$Accuracy(\%) = \frac{n_{11} + n_{22} + n_{33}}{n_1 + n_2 + n_3 + n_4} \times 100\% \tag{6}$$

Sensitivity describes how well a classification system can identify and classify positive data into the positive class. The higher the sensitivity value obtained, the better the classification system recognizes objects with favorable conditions [41]. Specificity is the degree of reliability of the model to detect negative labeled data correctly. Specificity measures the proportion of true negatives that are correctly identified, where true negative is the number of negative data belonging to a specific category that is correctly classified for that category. The sensitivity and specificity formulas used for each category are as follows [40,42]:

For Category 1, we have Sensitivity and Specificity as follows:

$$\begin{cases} Sensitivity_{y_1} = \dfrac{n_{11}}{n_{11} + n_{12} + n_{13}} = \dfrac{n_{11}}{n_1} \\ Specificity_{y_1} = \dfrac{n_{22} + n_{23} + n_{32} + n_{33}}{n_{22} + n_{23} + n_{32} + n_{33} + n_{21} + n_{31}} \end{cases} \tag{18}$$

For Category 2, we have Sensitivity and Specificity as follows:

$$\begin{cases} Sensitivity_{y_2} = \dfrac{n_{22}}{n_{21} + n_{22} + n_{23}} = \dfrac{n_{22}}{n_2} \\ Specificity_{y_2} = \dfrac{n_{11} + n_{13} + n_{31} + n_{33}}{n_{11} + n_{13} + n_{31} + n_{33} + n_{12} + n_{32}} \end{cases} \tag{19}$$

For Category 3, we have Sensitivity and Specificity as follows:

$$\begin{cases} Sensitivity_{y_3} = \dfrac{n_{33}}{n_{31} + n_{32} + n_{33}} = \dfrac{n_{33}}{n_3} \\ Specificity_{y_2} = \dfrac{n_{11} + n_{12} + n_{21} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22} + n_{13} + n_{23}} \end{cases} \tag{20}$$

3. RESULTS AND DISCUSSIONS

3.1. Characteristics of Data

In the following, we present the boxplots for age against with the hypertension status (Figure 1), for the body mass index against with the hypertension status (Figure 2), and the total cholesterol against with the hypertension status (Figure 3).

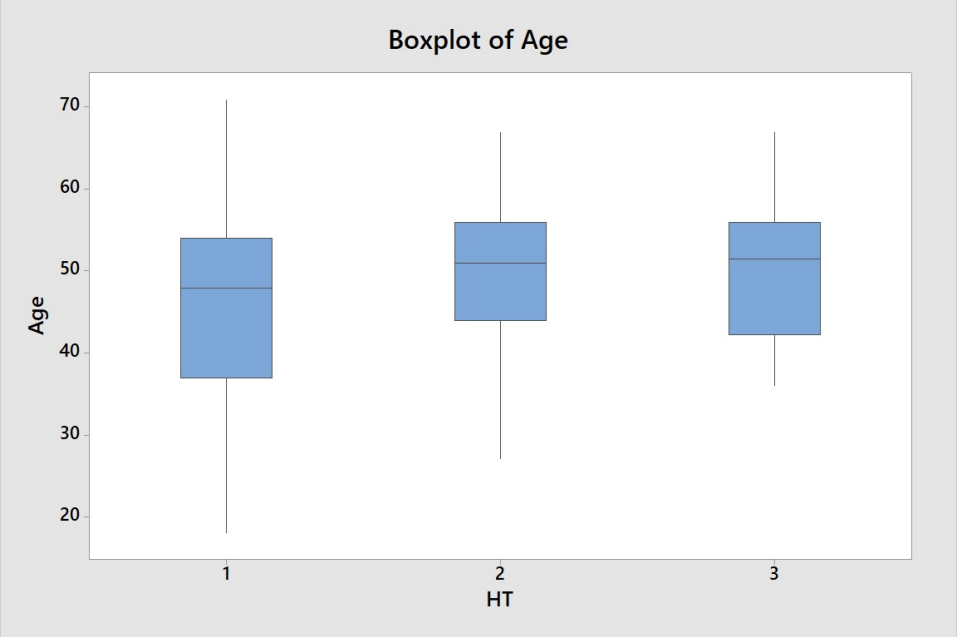

Figure 1. Boxplots for the age against the hypertension status.

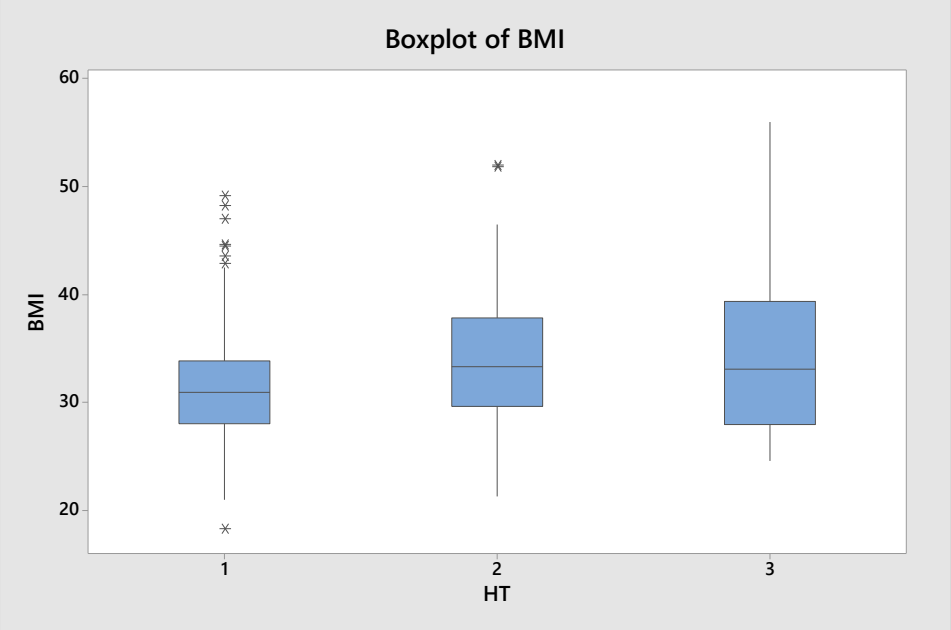

Figure 2. Boxplots the body mass index against the hypertension status.

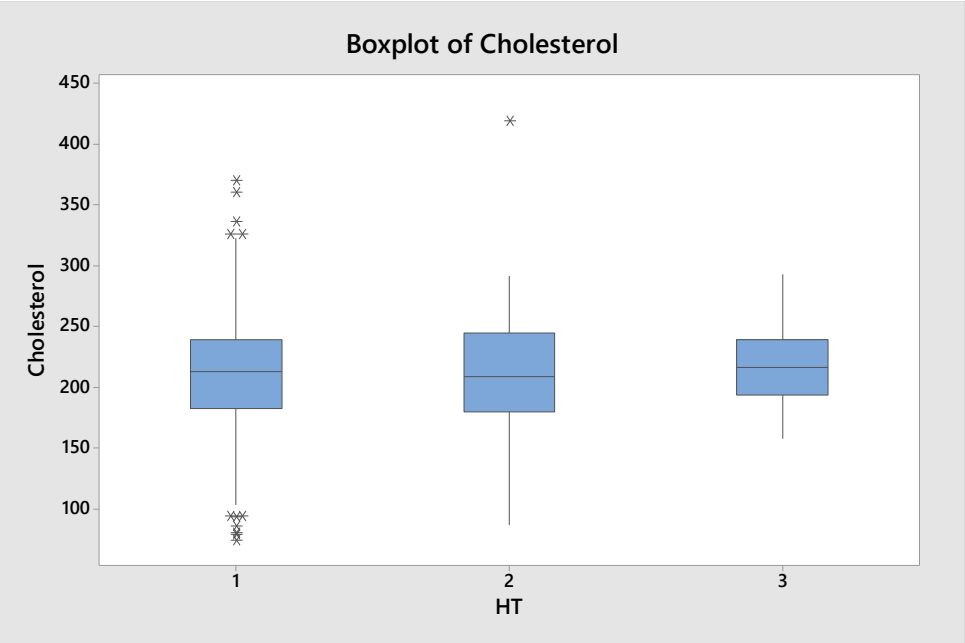

Figure 3. Boxplots for the total cholesterol against the hypertension status.

From Figure 1 to Figure 3, it can be seen that the median line of three boxplots lies outside of the box of a comparison box plot, so there is likely to be a difference between the three categories of hypertension based on age, body mass index and total cholesterol. Figure 1 shows that the interquartile range in category 1 is longer than the other categories, so it can be concluded that the age data in Category 1 (pre-hypertension) is more spread out than the age data in Category 2 (stage-1 hypertension) and 3 (stage-2 hypertension). Figure 2 shows that the interquartile range in Category 3 is longer than the other categories, so it can be concluded that the body mass index data in Category 3 (stage-2 hypertension) is more spread out than the body mass index data in Category 1 (pre-hypertension) and Category 2 (stage-1 hypertension). Figure 2 also shows several observations in Category 1 and Category 2 located outside the whiskers of the box plot. Figure 3 shows that the interquartile range in Category 2 is longer than the other categories, so it can be concluded that the total cholesterol data in Category 2 (stage-1 hypertension) is more spread out than the total cholesterol data in Category 1 (pre-hypertension) and category 3 (stage-2 hypertension). Figure 3 also shows several observations in Category 1 that are located outside the whiskers of the box plot. Furthermore, a general description of the predictor variables is known through descriptive statistics shown in Table 3, Table 4, and Table 5, respectively.

Table 3. Descriptive statistics of age, body mass index, and total cholesterol based on hypertension

Variable	Hypertension Status	Total Count	Mean	Variance	Minimum	Maximum	Range
Age	Pre-hypertension	355	46.29	121.44	18	71	53
	Stage-1 Hypertension	101	49.41	78.85	27	67	40
	Stage-2 Hypertension	32	50.09	70.9	36	67	31
BMI	Pre-hypertension	355	31.44	21.53	18.29	49.12	30.83
	Stage-1 Hypertension	101	33.89	32.04	21.37	51.95	30.58
	Stage-2 Hypertension	32	34.84	66.26	24.61	56	31.39
Total Cholesterol	Pre-hypertension	355	209.79	2147.4	74	370	296
	Stage-1 Hypertension	101	211.14	2242.97	87	419	332
	Stage-2 Hypertension	32	218.59	999.82	158	293	135

Table 3 provides the following information: (a) Patients with pre-hypertension status have an average age of 46.29 and a diversity of 121.44, with a range of 53, ranging from a minimum of 18 to a high of 71. Patients with stage 1 hypertension status have an average age of 49.41 and a range of 78.85, with a minimum of 27 and a maximum of 67 or 40. Additionally, patients with stage-2 hypertension status had an average age of 50.09 and a diversity of 70.9, with a range of 31 from a minimum of 36 to a maximum of 67. Thus, it can be concluded that stage-2 hypertension status has the highest average of 50.09 years old. (b) Patients with stage-1 hypertension status have an average body mass index of 31.4,4 and the variation is 21.53 with a minimum value of 18.29 and a maximum value of 49.12, so a range value of 30.83. The average body mass index for stage-1 hypertension status is 33.89, and the variation is 32.04, with a minimum value of 21.37 and a maximum value of 51.95, so a range value of 30.58. Meanwhile, the average body mass index for stage-2 hypertension status was 34.84 and the diversity is 66.26, with a minimum value of 24.61 and a maximum value of 56, so a range value of 31.39. Thus, it can be concluded that stage-2 hypertension status has the highest average of 34.84.

The average total cholesterol of patients with pre-hypertension status is 209.79, and the variation is 2147.4, with a minimum value of 74 and a maximum value of 370, so there is a

range value of 296. The average total cholesterol of patients with stage-1 hypertension status is 211.14, and the variation is 2242.97, with a minimum value of 87 and a maximum value of 419, so a range value of 332. Furthermore, patients with stage-2 hypertension status have the highest average of 218.59. The gender variable has a nominal scale with the category male and category female, so descriptive statistics of the gender variable shown by Table 4.

Table 4. Descriptive Statistics of Gender Variable

Gender	Hypertension Status			Total
	Pre-hypertension	Stage 1 Hypertension	Stage 2 Hypertension	
Male	163	54	23	**240**
Female	192	47	9	**248**
Total	**355**	**101**	**32**	**488**

Based on Table 4, it was noted that from the total sample, the number of male samples was 240 (49%) and the number of female samples was 249 (51%). The number of female samples was greater than that of male samples although the difference was small. Furthermore, the percentage of patients included in pre-hypertension status came from 67.91% male and 77.10% female. The percentage of patients included in stage-1 hypertension status came from 22.5% male and 18.88% female, while the percentage of patients included in stage-2 hypertension status came from 9.58% male and 4.01% female.

3.2. Estimation Result of Nonparametric Ordinal Logistic Regression based on MARS Model on Real Data

The next step is randomly dividing the 488 data into 2 groups, namely training data and testing data. The comparison of training and testing data is selected based on the minimum GCV value so that the training data taken is 90% or 439 observations, in comparison, the testing data is 10% or 49 observations. The lowest GCV value is used to determine the optimal model. The basis function (BF), maximum interaction (MI), and minimum observation (MO) are the three factors that must be taken into account when creating the MARS model. Each region defines a function known as the basis function (BF). Two to four times the number of predictor variables is the often employed basis function (BF). In this study, the number of predictor variables that are suspected of influencing hypertension status is four variables, hence, the number of basis functions (BF) that will be combined in forming variables is 8, 12, and 16.

The amount of interactions that can take place in the model is known as the maximum interaction (MI). In this investigation, maximum interactions (MI) of 1, 2, and 3 were employed. According to Friedman [8], if the maximum interaction (MI) is more than 3, the GCV value will increase and the resulting model will be more complex. If the maximum interaction (MI) used is 1, the model has no interaction between variables. If the maximum interaction (MI) used is 2, there can be interaction between variables in the model of up to 2. Likewise, if the maximum interaction (MI) used is 3, it means that there can be interaction between variables in the model of up to 3 variables. Minimum observation (MO) is the number of observations between the minimum knots. The minimum observation (MO) used in this study are 0, 1, 2 and 3, because above that value the GVC increases. Based on the R-code output, we obtained combination between the number of basis functions (BF), maximum interaction (MI), and minimum observation (MO), the results obtained are shown in Table 5.

Table 5. Parameter estimation results for parametric and nonparametric regression model

Model Number	BF	MI	MO	GCV	Rsq
[1,]	8	1	0	0.0365587	0.4491944
[2,]	8	1	1	0.0358759	0.4594804
[3,]	8	1	2	0.0359755	0.4579803
[4,]	8	1	3	0.0361921	0.4547178
[5,]	8	2	0	0.0369330	0.4517588
[6,]	8	2	1	0.0371872	0.4479853
[7,]	8	2	2	0.0367667	0.4484003
[8,]	8	2	3	0.0368437	0.4530848
[9,]	8	3	0	0.0369330	0.4517588
[10,]	8	3	1	0.0361882	0.4628143
[11,]	8	3	2	0.0367667	0.4484003
[12,]	8	3	3	0.0370153	0.4505366
[13,]	12	1	0	0.0363052	0.4668021
[14,]	12	1	1	0.0356877	0.4758713
[15,]	12	1	2	0.0358075	0.4741117
[16,]	12	1	3	0.0359411	0.4766610
[17,]	12	2	0	0.0367865	0.4712409
[18,]	12	2	1	0.0364828	0.4699159
[19,]	12	2	2	0.0367667	0.4484003
[20,]	12	2	3	0.0366219	0.4678935
[21,]	12	3	0	0.0367865	0.4712409

Model Number	BF	MI	MO	GCV	Rsq
[22,]	12	3	1	0.0364835	0.4825117
[23,]	12	3	2	0.0367667	0.4484003
[24,]	12	3	3	0.0368247	0.4764041
[25,]	16	1	0	0.0362152	0.4668021
[26,]	16	1	1	0.0356977	0.4758713
[27,]	16	1	2	0.0357275	0.4741117
[28,]	16	1	3	0.0359722	0.4766610
[29,]	16	2	0	0.0376865	0.4712409
[30,]	16	2	1	0.0373928	0.4699159
[31,]	16	2	2	0.0371567	0.4484003
[32,]	16	2	3	0.0369419	0.4678935
[33,]	16	3	0	0.0365865	0.4712409
[34,]*	16	3	1	0.0353874	0.4913508
[35,]	16	3	2	0.0367667	0.4484003
[36,]	16	3	3	0.0367647	0.4829298

Based on Table 5, from all the models obtained from the combination of BF, MI, and MO, based on the minimum GCV value, the best MARS model was selected and considered the most appropriate of the existing models. The combination of BF, MI, and MO that shows the best MARS model in model number 34 has BF = 16, MI = 3, and MO = 1 with GCV values of 0.0353874 and R^2 value of 49.13508%. Then, it was three predictor variables that were significant and affected hypertension status. It was X_1 (age), X_3 (body mass index), and X_4 (total cholesterol) using training data 90% and testing data 10%. The best MARS model of hypertension status that has been obtained as follows:

$$\hat{f}(x) = 3.095 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 - 0.00118BF_{10} + 0.008014BF_{12} \quad (7)$$

where

$BF_1 = h(50 - X_1)$;
$BF_2 = h(X_1 - 50)$;
$BF_3 = h(51.95 - X_3)_2$;
$BF_4 = h(360 - X_4)$;
$BF_5 = h(X_1 - 42)BF_6$;
$BF_6 = h(X_4 - 226)$;
$BF_7 = h(X_1 - 55)BF_6$;
$BF_8 = h(245 - X_4)BF_3$;
$BF_9 = h(224 - X_4)BF_1$;
$BF_{10} = h(24.22 - X_3)BF_1 * BF_4$;
$BF_{11} = h(X_3 - 37.24)$
$BF_{12} = h(X_1 - 53)BF_{11} * BF_6$

In interpreting the MARS model above, we will use the interpretation of three basis functions as an example. Interpretation for three basis functions of the best MARS model. from Equation (17) as follows,

$$BF_2 = h(X_1 - 50) \tag{8}$$

Coefficient BF_2 will be statistically significant if the patient is over 50 years old. Each BF_2 increase of one unit with the patient's age is more than 50, increases the risk of hypertension $\exp(0.02687) = 1.02723$, while other basis functions included in the model are considered constant. Age is a factor that influences the occurrence of hypertension [43,44]. Increased arterial stiffness and endothelial dysfunction are linked to aging in hypertension, particularly systolic hypertension in the elderly. Age is linked to hypertension because of normal bodily changes that alter the vascular system, including the heart, blood vessels, and hormones. These changes raise blood pressure, which in turn causes hypertension [45,46].

$$\begin{aligned} BF_5 &= h(X_1 - 42)\, BF_6; \\ BF_6 &= h(X_4 - 226) \end{aligned} \tag{9}$$

Coefficient BF_5 will be statistically significant if patient age is over 42 years old with total cholesterol more than 226 mg/dL. Each BF_5 increase of one unit with the patient's age is more than 42 years old and total cholesterol more than 226 mg/dL, increases the risk of hypertension is $\exp(0.003354) = 1.00336$ at the same time other basis functions included in the model are considered constant. Cholesterol is a complex problem in the human body. Even young toddlers might be impacted by unhealthy cholesterol levels. However, those between the ages of 40 and 59 are frequently diagnosed with excessive cholesterol [47]. Narrowing and stiffening of the walls of blood vessels due to the accumulation of cholesterol in the vessels can cause blood pressure to increase[45]. Numerous supporting research support the idea that elevated blood cholesterol levels are common in people with hypertension [48].

$$\begin{aligned} BF_{12} &= h(X_1 - 53)\, BF_{11} * BF_6 \\ BF_{11} &= h(X_3 - 37.24) \\ BF_6 &= h(X_4 - 226) \end{aligned} \tag{10}$$

Coefficient BF_{12} will be statistically significant if the patient is over 53 with a body mass

index of more than 37.24 and total cholesterol of more than 226 mg/dL. Each BF_{12} increase of one unit with the patient's age is more than 53 years old with body mass index more than 37.24 and total cholesterol more than 226 mg/dL, an increase in the risk of hypertension is $\exp(0.008014) = 1.00805$ while other basis functions included in the model are considered constant. Other studies also concluded that high body mass index (BMI) was significantly associated with hypertension [12]. The relationship between excess adiposity and increased blood pressure is well-established [49].

Based on Equation (11) dan (21), the logit link function is written as follows:

$$\hat{g}_1(x) = -0.46311 + \begin{bmatrix} 3.095 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{bmatrix} \tag{11}$$

$$\hat{g}_2(x) = 0.027271 + \begin{bmatrix} 3.095 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{bmatrix} \tag{12}$$

Based on estimated models in Equations (25) and (26), we obtain the estimated probability models as follows:

Probability of Pre-hypertension Status is given by:

$$\pi_1(x) = \frac{\exp\left(-0.46311 + \begin{bmatrix} 3.095 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{bmatrix} \right)}{1 + \exp\left(-0.46311 + \begin{bmatrix} 3.095 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{bmatrix} \right)}$$

(13)

$$\pi_1(x) = \frac{\exp\begin{pmatrix} 2.63189 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{pmatrix}}{1 + \exp\begin{pmatrix} 2.63189 - 0.02687BF_1 + 0.00108BF_2 - 0.03634BF_3 - 0.03584BF_4 \\ + 0.003354BF_5 + 0.001245BF_7 - 0.002288BF_8 - 0.01006BF_9 \\ -0.00118BF_{10} + 0.008014BF_{12} \end{pmatrix}}$$

Probability of Stage-1 hypertension Status is given by:

$$\pi_2(x)=\frac{\exp\left(0.02727+\begin{bmatrix}3.095-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{bmatrix}\right)}{1+\exp\left(0.027271+\begin{bmatrix}3.095-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{bmatrix}\right)}$$

(14)

$$\pi_2(x)=\frac{\exp\begin{pmatrix}3.12227-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{pmatrix}}{1+\exp\begin{pmatrix}3.12227-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{pmatrix}}$$

Probability of Stage-2 hypertension Status is given by:

$$\pi_3(x)=\frac{1}{1+\exp\left(0.027271+\begin{bmatrix}3.095-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{bmatrix}\right)}$$

(15)

$$\pi_3(x)=\frac{1}{1+\exp\begin{pmatrix}3.12227-0.02687BF_1+0.00108BF_2-0.03634BF_3-0.03584BF_4\\+0.003354BF_5+0.001245BF_7-0.002288BF_8-0.01006BF_9\\-0.00118BF_{10}+0.008014BF_{12}\end{pmatrix}}$$

Regarding model interpretation, we only offer one example of discussion in this study utilizing one of the test results, which is a patient who is 57 years old, has a body mass index of 38.95, and total cholesterol of 238 mg/dL. In the meantime, other testing data undergoes the same computation procedure. Additionally, we derive the estimated probability models' values as follows:

$$\pi_1(x)=\frac{\exp\begin{pmatrix}2.63189-0+0.00756-0.47242-4.37248+0.60372+\\0.02988-0.208208-0-0+0.155792\end{pmatrix}}{1+\exp\begin{pmatrix}2.63189-0+0.00756-0.47242-4.37248+0.60372+\\0.02988-0.208208-0-0+0.155792\end{pmatrix}}$$

(16)

$$=\frac{\exp(-1.62427)}{1+\exp(-1.62427)}=\frac{0.197056}{1+0.197056}=0.164617$$

$$\pi_1(x) = \frac{\exp\begin{pmatrix}3.12227-0+0.00756-0.47242-4.37248+0.60372+\\0.02988-0.208208-0-0+0.155792\end{pmatrix}}{1+\exp\begin{pmatrix}3.12227-0+0.00756-0.47242-4.37248+0.60372+\\0.02988-0.208208-0-0+0.155792\end{pmatrix}} \quad (17)$$

$$= \frac{\exp(-1.13389)}{1+\exp(-1.13389)} = \frac{0.32178}{1+0.32178} = 0.243445$$

$$\pi_3(x) = 1-\pi_1(x)-\pi_2(x) \quad (18)$$

$$= 1-0.164617-0.243445 = 0.591938$$

We can explain that in comparison to other levels of hypertension, a patient aged 57 with a body mass index of 38.95 and total cholesterol of 238 mg/dL has the highest chance of risk, i.e., stage-2 hypertension, based on the values of the estimated probability models shown in Equations (29) to (31). The patient was assigned a score of three and was diagnosed with stage-2 hypertension based on the training data's highest likelihood value. An identical mathematical procedure can be used to estimate the results of various training data.

3.3. Evaluation of Classification Procedures

The next step is to calculate the classification value between the actual observation value and the predicted value obtained from the model that has been formed. Classification accuracy describes how well a system or method classifies data. A confusion matrix, composed of actual observations and predicted results, can be used to measure classification accuracy in ordinal logistic regression.

Table 6. Confusion matrix for training data

Observation	**Prediction**			**Total**
	Pre-hypertension	Stage 1 Hypertension	Stage 2 Hypertension	
Pre-hypertension	148	52	3	202
Stage-1 Hypertension	25	111	12	148
Stage-2 Hypertension	4	39	45	88

The results of the confusion matrix value show that 148 pre-hypertension patients were correctly classified, while 111 stage-1 hypertension patients were correctly classified and 45 stage-2 hypertension patients were correctly classified. This can be interpreted that the model formed can classify pre-hypertension patients correctly by 33.71%. In comparison, stage-1 hypertension patients can be correctly classified by 25.28% and stage 2 hypertension patients

can be correctly classified by 10.25% of the total observations for each hypertension status.

If viewed as a whole, the prediction results that the model correctly classifies, then the accuracy value obtained is 69.25%. The accuracy value is not good enough to assess the classification performance in this study, so other measurements are needed, namely the sensitivity value, which is the percentage of correct positive prediction results from the total prediction results that are correctly classified and specificity is the percentage of classification performance to show the correct negative prediction results from the total negative prediction results in each category as follows:

Table 7. Sensitivity and specificity

Status	Sensitivity	Sensitivity
Pre-hypertension	73.27%	87.71%
Stage-1 Hypertension	75.0%	68.73%
Stage-2 Hypertension	51.14%	95.73%

Table 7 shows that the sensitivity or percentage of correct prediction results in pre-hypertension patients from the total prediction results that were correctly classified was 73.27%. In comparison the percentage for stage-1 hypertension was 75.0% and the percentage of correct prediction results in stage-2 hypertension patients, based on the total prediction results that were correctly classified, was 51.14%. The specificity value for patients classified into pre-hypertension was 87.71%. At the same time, stage-1 hypertension had a specificity value of 68.73%, then and the specificity value for patients classified into stage-2 hypertension was 95.73%. Based on the calculation of the classification accuracy from the accuracy, sensitivity, and specificity values, it was found that the model formed was not good enough to classify hypertension patients.

Next, the classification accuracy for the entire testing data is presented in Table 8.

Table 8. Confusion matrix for testing data

Observation	Prediction			Total
	Pre-hypertension	Stage 1 Hypertension	Stage 2 Hypertension	
Pre-hypertension	13	5	1	19
Stage-1 Hypertension	3	15	3	21
Stage-2 Hypertension	1	4	9	14

Based on Table 8, we can determine the accuracy values for testing data as follows,

$$Accuracy(\%) = \frac{13+15+9}{19+21+14} x100\% = 68.52\% \quad (19)$$

For testing data, the model estimation accuracy value is 68.52%. This demonstrates the validity of the ordinal logistic nonparametric regression model that was estimated using the acquired MARS estimator to ascertain the risk of hypertension in test data. The estimated model can explain 68.52% of the incidence of hypertension in the test data. This will suggest that the MARS estimator-estimated ordinal logistic nonparametric regression model is likewise appropriate for forecasting the risk of hypertension.

4. CONCLUSIONS

In this study, two conclusions were obtained. The first conclusion is that the best MARS model of hypertension status has BF = 16, MI = 3 and MO = 1 with GCV values of 0.0353874, and R^2 value of 49.13508%. Then, it was three predictor variables that significant and affected hypertension status. It was age, body mass index and total cholesterol. Analysis results using the nonparametric ordinal logistic model based on the MARS estimator show that all patients who were predicted to suffer from stage-2 hypertension by the nonparametric ordinal logistic regression model based on the MARS estimator were also sufferers of stage-2 hypertension based on the doctor's diagnosis with an accuracy value of 69.25%, a sensitivity value of 73.27% for pre-hypertension, 75.0% for stage-1 hypertension and 51.14% for stage-2 hypertension. A specificity value of 87.71% for pre-hypertension, 68.73% for stage-1 hypertension and 95.73% for stage-2 hypertension. Predictions using the nonparametric ordinal logistic model based on the MARS estimator provide relatively the same results as a doctor's diagnosis. These findings suggest that the MARS estimator-based nonparametric ordinal logistic model is a reliable method for estimating the risk of hypertension. Thus, the findings of this study can serve as the foundation for early warning systems in the future to inform people who are elderly and have high body mass index and high cholesterol about their risk of developing hypertension.

ACKNOWLEDGMENTS

The authors thank The Center for Education Financial Services and The Indonesia Endowment Funds for Education for funding this research. Also, the authors thank Dr. Drs. Budi Lestari, PG.Dip.Sc.,M.Si., for providing pre-review, constructive comments and suggestions for improving the quality of this paper.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] M. Burnier, B.M. Egan, Adherence in Hypertension: A Review of Prevalence, Risk Factors, Impact, and Management, Circ. Res. 124 (2019), 1124–1140. https://doi.org/10.1161/CIRCRESAHA.118.313220.

[2] A. Kurnianto, D.K. Sunjaya, F.R. Rinawan, D. Hilmanto, Prevalence of Hypertension and Its Associated Factors among Indonesian Adolescents, Int. J. Hypertens. 2020 (2020), 4262034. https://doi.org/10.1155/2020/4262034.

[3] R. Kurniawan, B. Utomo, K.N. Siregar, K. Ramli, B. Besral, R.J. Suhatril, O.A. Pratiwi, Hypertension Prediction Using Machine Learning Algorithm among Indonesian Adults, IAES Int. J. Artif. Intell. 12 (2023), 776-784. https://doi.org/10.11591/ijai.v12.i2.pp776-784.

[4] K. Peltzer, S. Pengpid, The Prevalence and Social Determinants of Hypertension among Adults in Indonesia: A Cross-Sectional Population-Based National Survey, Int. J. Hypertens. 2018 (2018), 5610725. https://doi.org/10.1155/2018/5610725.

[5] D. Sun, J. Liu, L. Xiao, et al. Recent Development of Risk-Prediction Models for Incident Hypertension: An Updated Systematic Review, PLOS ONE 12 (2017), e0187240. https://doi.org/10.1371/journal.pone.0187240.

[6] P. Santhanam, R.S. Ahima, Machine Learning and Blood Pressure, J. Clin. Hypertens. 21 (2019), 1735–1737. https://doi.org/10.1111/jch.13700.

[7] P. Andriani, N. Chamidah, Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia, J. Phys.: Conf. Ser. 1306 (2019), 012027. https://doi.org/10.1088/1742-6596/1306/1/012027.

[8] J.H. Friedman, C.B. Roosen, An Introduction to Multivariate Adaptive Regression Splines, Stat. Methods Med. Res. 4 (1995), 197–217. https://doi.org/10.1177/096228029500400303.

[9] M. Hasyim, D.S. Rahayu, N.E. Muliawati, et al. Bootstrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) to Analyse the Lecturer Research Performance in Private University, J. Phys.: Conf. Ser. 1114 (2018), 012117. https://doi.org/10.1088/1742-6596/1114/1/012117.

[10] T. Adiwati, N. Chamidah, Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia, IOP Conf. Ser.: Mater. Sci. Eng. 546 (2019), 052003. https://doi.org/10.1088/1757-899X/546/5/052003.

[11] T.D. Rohkuswara, S. Syarif, Hubungan Obesitas Dengan Kejadian Hipertensi Derajat 1 Di Pos Pembinaan Terpadu Penyakit Tidak Menular (Posbindu PTM) Kantor Kesehatan Pelabuhan Bandung Tahun 2016, J. Epidemiol. Kesehat. Indones. 1 (2017), 3. https://doi.org/10.7454/epidkes.v1i2.1805.

[12] Y. Qu, H. Niu, L. Li, et al. Analysis of Dose-Response Relationship between BMI and Hypertension in Northeastern China Using Restricted Cubic Spline Functions, Sci. Rep. 9 (2019), 18208. https://doi.org/10.1038/s41598-019-54827-2.

[13] Z.N. Amalia, D.R. Hastuti, F. Istiqomah, N. Chamidah, Hypertension Risk Modeling Using Penalized Spline Estimator Approach Based on Consumption of Salt, Sugar, and Fat Factors, AIP Conf. Proc. 2264 (2020), 030005. https://doi.org/10.1063/5.0023456.

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

[15] D. Ruppert, Selecting the Number of Knots for Penalized Splines, J. Comput. Graph. Stat. 11 (2002), 735–757. https://doi.org/10.1198/106186002853.

[16] B. Lestari, Fatmawati, I.N. Budiantara, Spline Estimator and Its Asymptotic Propertiesin Multiresponse Nonparametric Regression Model, Songklanakarin J. Sci. Technol. 42 (2020), 533548. https://doi.org/10.14456/SJST-PSU.2020.68.

[17] W. Ramadan, N. Chamidah, B. Zaman, L. Muniroh, B. Lestari, Standard Growth Chart of Weight for Height to Determine Wasting Nutritional Status in East Java Based on Semiparametric Least Square Spline Estimator, IOP Conf. Ser.: Mater. Sci. Eng. 546 (2019), 052063. https://doi.org/10.1088/1757-899X/546/5/052063.

[18] N. Chamidah, B. Lestari, T. Saifudin, Modeling of Blood Pressures Based on Stress Score Using Least Square Spline Estimator in Bi-Response Non-Parametric Regression, Int. J. Innov., Creat. Change 5 (2019), 1200–1216.

[19] Fatmawati, I.N. Budiantara, B. Lestari, Comparison of Smoothing and Truncated Spline Estimators in Estimating Blood Pressure Models, Int. J. Innov., Creat. Change 5 (2019), 1177–1199.

[20] N. Chamidah, B. Lestari, A. Massaid, et al. Estimating Mean Arterial Pressure Affected by Stress Scores Using Spline Nonparametric Regression Model Approach, Commun. Math. Biol. Neurosci. 2020 (2020), 72. https://doi.org/10.28919/cmbn/4963.

[21] N. Chamidah, B. Lestari, A. Y. Wulandari, et al. Z-Score standard growth chart design of toddler weight using least square spline semiparametric, AIP Conf. Proc. 2329 (2021), 060031. https://doi.org/ 10.1063/5.0042285.

[22] N. Chamidah, B. Lestari, I.N. Budiantara, et al. Consistency and Asymptotic Normality of Estimator for Parameters in Multiresponse Multipredictor Semiparametric Regression Model, Symmetry 14 (2022), 336. https://doi.org/10.3390/sym14020336.

[23] B. Lestari, N. Chamidah, D. Aydin, E. Yilmaz, Reproducing Kernel Hilbert Space Approach to Multiresponse Smoothing Spline Regression Function, Symmetry 14 (2022), 2227. https://doi.org/10.3390/sym14112227.

[24] N.Chamidah, B. Zaman, L. Muniroh, et al. Multiresponse Semiparametric Regression Model Approach to

Standard Growth Charts Design for Assessing Nutritional Status of East Java Toddlers, Commun. Math. Biol. Neurosci. 2023 (2023), 30. https://doi.org/10.28919/cmbn/7814.

[25] B. Lestari, N. Chamidah, I. Nyoman Budiantara, D. Aydin, Determining Confidence Interval and Asymptotic Distribution for Parameters of Multiresponse Semiparametric Regression Model Using Smoothing Spline Estimator, J. King Saud Univ. - Sci. 35 (2023), 102664. https://doi.org/10.1016/j.jksus.2023.102664.

[26] D. Aydın, E. Yılmaz, N. Chamidah, et al. Right-Censored Nonparametric Regression with Measurement Error, Metrika (2024). https://doi.org/10.1007/s00184-024-00953-5.

[27] D. Aydın, E. Yılmaz, N. Chamidah, B. Lestari, Right-Censored Partially Linear Regression Model with Error in Variables: Application with Carotid Endarterectomy Dataset, Int. J. Biostat. 20 (2024), 245–278. https://doi.org/10.1515/ijb-2022-0044.

[28] N. Chamidah, B. Lestari, I.N. Budiantara, D. Aydin, Estimation of Multiresponse Multipredictor Nonparametric Regression Model Using Mixed Estimator, Symmetry 16 (2024), 386. https://doi.org/10.3390/sym16040386.

[29] M. Hasyim, 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. 974 (2018), 012065. https://doi.org/10.1088/1742-6596/974/1/012065.

[30] J.H. Friedman, Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines, Technical Reports, No. 108, 1991. https://purl.stanford.edu/ty209wk6792.

[31] R.M. Adnan, Z. Liang, S. Heddam, et al. Least Square Support Vector Machine and Multivariate Adaptive Regression Splines for Streamflow Prediction in Mountainous Basin Using Hydro-Meteorological Data as Inputs, J. Hydrol. 586 (2020), 124371. https://doi.org/10.1016/j.jhydrol.2019.124371.

[32] C. R. Bilder, J. M. Tebbs, An introduction to categorical data analysis, J. Amer. Stat. Assoc. 103 (2008), 19 pages. https://doi.org/10.1198/jasa.2008.s251.

[33] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, 2000.

[34] A. Wibowo, Pemodelan MARS dan Regresi Logistik Rumah Tangga Miskin Kalimantan Tengah Tahun 2016, J. Mat. Pendidik. Mat. 3 (2018), 1. https://doi.org/10.26594/jmpm.v3i1.1023.

[35] E.O. Permatasari, F. Nasuha, C.L. Prawirosastro, Modeling the Level of Open Unemployment in Central Java with Multivariate Adaptive Regression Splines (MARS) Approach, J. ASRO 12 (2021), 66. https://doi.org/10.37875/asro.v12i01.382.

[36] M. Meilisa, B.W. Otok, J.D.T. Purnomo, Factor Affecting the Severity of the Dengue Fever Using Multivariate Adaptive Regression Spline (Mars) Method, AIP Conf. Proc. 2556 (2023), 050003. https://doi.org/10.1063/5.0131508.

[37] L. Green, JNC 7 Express: New Thinking in Hypertension Treatment, Am. Fam. Physician 68 (2003), 228-230.

[38] I. Zain, Z. Zakariyah, Analisis Regresi Logistik Ordinal pada Prestasi Belajar Lulusan Mahasiswa di ITS Berbasis SKEM, J. Sains Seni ITS 4 (2015), 121–126.

[39] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 6th ed, Pearson Prentice Hall, Upper Saddle River, 2007.

[40] J. Ha. M. Kambe, J. Pe, Data Mining: Concepts and Techniques, Elsevier, 2012. https://doi.org/10.1016/C2009-0-61819-5.

[41] A. Miladitiya, Sensitivitas dan Spesifisitas Lingkar Pinggang Dalam Mengidentifikasi Kelebihan Berat Badan dan Obesitas Pada Wanita Dewasa, Interest: J. Ilmu Kesehatan 7 (2018), 22-28. https://doi.org/10.37341/interest.v7i1.64..

[42] W. Härdle, L. Simar, Applied Multivariate Statistical Analysis, Springer, Berlin, Heidelberg, 2003. https://doi.org/10.1007/978-3-662-05802-2.

[43] E. Salsabila, S.L. Utami, S. Sahadewa, Faktor Risiko Usia dan Jenis Kelamin dengan Kejadian Hipertensi di Klinik Paradise Surabaya Oktober 2023, Calvaria Med. J. 2 (2024), 64–68.

[44] N. Chamidah, B. Lestari, H. Susilo, et al. Spline Estimator in Nonparametric Ordinal Logistic Regression Model for Predicting Heart Attack Risk, Symmetry 16 (2024), 1440. https://doi.org/10.3390/sym16111440.

[45] A. Adila, S.E. Mustika, Hubungan Usia dan Jenis Kelamin Terhadap Kejadian Kanker Kolorektal, J. Kedokt. (Sains Teknol. Medik) 6 (2023), 53–59. https://doi.org/10.30743/stm.v6i1.349.

[46] M. Guèze, L. Napitupulu, Trailing Forest Uses Among the Punan Tubu of North Kalimantan, Indonesia, in: V. Reyes-García, A. Pyhälä (Eds.), Hunter-Gatherers in a Changing World, Springer, Cham, 2017: pp. 41–58. https://doi.org/10.1007/978-3-319-42271-8_3.

[47] A.M. Nawi, Z. Mohammad, K. Jetly, et al. The Prevalence and Risk Factors of Hypertension among the Urban Population in Southeast Asian Countries: A Systematic Review and Meta-Analysis, Int. J. Hypertens. 2021 (2021), 6657003. https://doi.org/10.1155/2021/6657003.

[48] H. Maryati, Hubungan Kadar Kolesterol Dengan Tekanan Darah Penderita Hipertensi di Dusun Sidomulyo Desa Rejoagung Kecamatan Plos Kabupaten Jombang, J. Keperawat. 8 (2017), 128–137.

[49] O.A. Shariq, T.J. McKenzie, Obesity-Related Hypertension: A Review of Pathophysiology, Management, and the Role of Metabolic Surgery, Gland Surg. 9 (2020), 80–93. https://doi.org/10.21037/gs.2019.12.03.