BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2217–2230
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2217-2230

MODELING HYPERTENSION DISEASE RISK IN INDONESIA USING MULTIVARIATE ADAPTIVE REGRESSION SPLINE AND BINARY LOGISTIC REGRESSION APPROACHES

Nur Chamidah[1*], Ardana Tegar Hendrawan [2], Figo Surya Ardiyanto[3], Martha Sayyida Hammami[4], Nurul Izzah[5], Salsabila Niken Hariadi[6]

[1]*Mathematics Department, Faculty of Science and Technology, Universitas Airlangga*
[2,3,4,5,6]*Statistics Study Program, Faculty of Science and Technology, Universitas Airlangga*
Jl. Dr. Ir H.Soekarno, Muyorejo, Surabaya, 60115, Indonesia

Corresponding author's e-mail: * **nur-c@fst.unair.ac.id**

ABSTRACT

Article History:

Received: 15th January 2024
Revised: 14th May 2024
Accepted: 10th July 2024
Published: 14th October 2024

Keywords:

Hypertension;
Regression;
MARS;
Binary Logit Regression.

In the pursuit of the Sustainable Development Goals (SDGs), health-related challenges, especially hypertension, remain a significant global issue. The third goal of the SDGs aims to improve the quality of life and well-being of all individuals, but hypertension is a serious problem that can hinder these goals. Often referred to as the "silent killer" by the World Health Organization (WHO), hypertension is exacerbated by low awareness. Globally, more than 1.28 billion adults suffer from hypertension, with most cases in lower to middle-income countries, including Indonesia. Indonesia has an alarming rate of hypertension incidence, ranking fifth highest in the world. Riset Kesehatan Dasar (Riskesdas) 2023 and the Indonesia Family Life Survey (IFLS) are critical for understanding hypertension risk factors in Indonesia. The IFLS data, obtained from www.rand.org, includes observations from October 2014 to April 2015, totalling 85 observations. Despite being over 10 years old, this dataset was selected because it remains the most recent comprehensive data available from RAND, representing 83% of the Indonesian population. The IFLS is conducted every 7-8 years, with the next wave of data expected soon. Most studies on hypertension globally and in Indonesia use parametric regression methods. However, a research gap exists as no studies have used Multivariate Adaptive Regression Splines (MARS) on IFLS data to analyze hypertension risk factors. This study addresses this gap by comparing binary logit regression and MARS. The analysis shows the Apparent Error Rate (APPER) for MARS is 84.706%, while for binary logistic regression it is 80%, indicating MARS is better at classifying hypertension data in Indonesia. Using MARS offers a novel approach to understanding hypertension risk factors in Indonesia. Despite the data's age, it remains relevant as primary causes and risk factors for hypertension have not changed, making the findings valuable for current health policy and strategies.

.

How to cite this article:

N. Chamidah, A. T. Hendrawan, F. S. Ardiyanto, M. S. Hammami, N. Izzah and S. N. Hariadi., "MODELING HYPERTENSION DISEASE RISK IN INDONESIA USING MULTIVARIATE ADAPTIVE REGRESSION SPLINE AND BINARY LOGISTIC REGRESSION APPROACHES," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2217-2230, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

In the pursuit of achieving the Sustainable Development Goals (SDGs), there are still a number of challenges facing the world today. The third goal of the SDGs is to guarantee a healthy quality of life and improve the well-being of all individuals, regardless of age. One of the most significant health problems is hypertension, a condition in which systolic blood pressure exceeds 140 mmHg and diastolic pressure exceed 90 mm Hg [1]. According to the World Health Organization (WHO), about 46% of adults suffering from hypertension are unaware of their condition, so the disease is often referred to as "the silent killer of death". Causes of hypertension involve a number of factors that need to be weighed. These factors include gender, age, body mass index (BMI), physical activity, and the respondent's psychological condition.

Globally, there are 1.28 billion adults aged 30-79 suffering from hypertension, with more than two thirds of them living in low- and middle-income countries, including Indonesia [2]. Based on the results of Riset Kesehatan Dasar (Riskesdas) in 2023, the incidence of hypertension in Indonesia reached 30.8% of the total population. Among them, Indonesia ranks fifth as the country with the most severe hypertension in the world [3]. To obtain information related to hypertension in Indonesia can be done Indonesia Family Life Survey (IFLS). IFLS is a longitudinal survey carried out in Indonesia with samples representing about 83% of the Indonesian population [4]. Therefore, data from the IFLS survey is needed for an analysis of understanding of the risk factors of hypertension as a form of effective hypertensive prevention and management efforts.

The influence of factors associated with hypertensive patients based on the data obtained from the Indonesia Family Life Survey (IFLS) can be analyzed through the implementation of regression. Regression is a statistical method used to measure the relationship between two or more variables that have a correlation [5]. Two regression approaches that can be used are parametric and non-parametric regression. [6]. In the context of the analysis, the study compared the results of the modeling with the methods of parametric and non-parametric regression. The method used in the study is binary logit regression, whereas the method used as a comparison is the Multivariate Adaptive Regression Splines (MARS), introduced by Friedman in 1991. The main advantages of this method are its ability to solve problems that arise when data has a curse of dimensionality and its capacity to overcome the weaknesses of a parametric regression model in producing a continuous model at the cutting points [7].

There is a number of studies related to hypertension, such as the study by Naiborhu entitled Survival Analysis with Multivariate Adaptive Regression Splines (MARS) Approach in Hypertension Disease Cases. The results of the study state that the factors influencing the rate of healing of hypertensive patients are age variables, diabetes, kidneys, systolic, and diastolic [8]. Another study is entitled Comparison Classification of Hypertension Diseases using Binary Logistic Regression and C4.5 algorithm, which was made by Rumaenda. The results of the study show that the variables that influence the probability of hypertension include gender, systolic blood pressure, and the presence of other diseases with an APER of 27.4648% [9]. In this study, binary variables are used for dependent variables. Additional factors will be tested such as gender, age, BMI, physical activity, and a psychologist to see if there is an influence between the factors and the hypertension suffered. In statistics, the MARS and Logit Regression binary models can be used to estimate the regression model where the variable used is a binary variable.

Based on some of the phenomena that have been shown, the researchers wanted to describe the general overview of modeling hypertension cases in Indonesia, assess and compare the modeling of MARS models and logit regression binary on the factors that influence hypertension disease based on data from the Indonesia Family Life Survey. (IFLS). It is hoped that this study will reveal factors that play an important role in rising rates of hypertension so that readers can prevent it as soon as possible.

2. RESEARCH METHODS

2.1 Data Sources

In research using secondary data sourced from Indonesia Family Life Survey in 2014. The units of observation in this study were household life aspects that include five aspects. The Number of people that were observed was 85 people.

2.2 Research Variables

A research variable is a representation of an event, behavior, character, or trait that is calculated [10]. The variables used in this study consist of response variables and predictor variables. Variable response (Y) used in this study are people with disease Hypertension in 2014, meanwhile variable predictors (X) used are as many as 5 variables. Here are the variables Responses and predictors used in this study is shown in Table 1.

Table 1. Research Variable

Variable	Variable Name	Scale	Note
$\boldsymbol{Y}$	Diagnosis of Hypertension	Nominal	1 = Diagnosed with Hypertension 2 = Not Diagnosed with Hypertension
$\boldsymbol{X_1}$	Sex	Nominal	1 = Male 2 = Female
$\boldsymbol{X_2}$	Age	Nominal	1 = patients aged less than 45 years old 2 = patients aged more than or equal to 45 years old
$\boldsymbol{X_3}$	Body Mass Index (BMI)	Ratio	Body Mass Index (BMI) Score with unit of kg/m^2
$\boldsymbol{X_4}$	Psychological condition	Ratio	Psychological Condition Score
$\boldsymbol{X_5}$	Physical activity	Nominal	1 = High physical acivity 2 = Low

The data structure used in this study is shown in Table 2.

Table 2. Data Structures Used

No.	$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	$\boldsymbol{X_4}$	$\boldsymbol{X_5}$
1	Y_1	$X_{1.1}$	$X_{2.1}$	$X_{3.1}$	$X_{4.1}$	$X_{5.1}$
2	Y_2	$X_{1.2}$	$X_{2.2}$	$X_{3.2}$	$X_{4.2}$	$X_{5.2}$
3	Y_3	$X_{1.3}$	$X_{2.3}$	$X_{3.3}$	$X_{4.3}$	$X_{5.3}$
:	:	:	:	:	:	:
:	:	:	:	:	:	:
85	Y_{85}	$X_{1.85}$	$X_{2.85}$	$X_{3.85}$	$X_{4.85}$	$X_{5.85}$

2.3 MARS Model Estimation

Multivariate Adaptive Regression Spline (MARS) is nonparametric regression technique capable of incorporating both additive and interaction effects among predictor variables [11]. This method can yield accurate predictions for the regression curve shape based on unknown pattern relationships between response and predictors. The MARS model as follows:

$$\boldsymbol{y_i = a_0 + \sum_{m=1}^{M} a_m B_{mi}(x,t) + \varepsilon_i} \tag{1}$$

2.3.1 Basic Function Combination

MARS is a versatile approach for modeling high-dimensional regression data [12]. The calculation was conducted using the MARS application by combining Basis Function (BF), Maximum Interaction (MI), and Minimum Observation between knots (MO). The value of BF ranges from 10 to 20, MI includes 1 and 2, while MO consists of 0, 1, 2, and 3. The criterion for the best model is the lowest value of the Generalized Cross Validation (GCV). If there are similar minimum GCV values, then it can be observed from the maximum value of R^2 and the minimum value of Mean Square Error (MSE).

2.3.2 Basic Function Estimation

Basic function estimation can be examined when the best model is already found. the MARS model obtained to estimate the dependent variable. In this research, dependent variables consist of two values therefore the probability model can be determined. to determine which independent factors that influence dependent variables, it can be observed through the interpretation of odds ratios.

$$P(Y=1|X_1,X_2,\dots,X_p)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)}} \quad (2)$$

$P(Y=1|X_1,X_2,\dots,X_p)$ represents the probability that the response variable equals 1 (or success) given the values of predictor variables $X_1,X_2,\dots,X_p$ and β_0 is the intercept while $\beta_1,\beta_2,\dots,\beta_p$ are the coefficients associated with $X_1,X_2,\dots,X_p$ respectively. Thus, can be rewrite as formula below.

$$OR\ BF_i=\frac{\exp(BF_0+BF_i(X_i))}{1+\exp(BF_0+BF_i(Xi))} \quad (3)$$

where $i=1,6,7$

2.4 MARS Model Significance Test

Significance testing of the MARS model is conducted when the assumptions on the residuals have been met. This test is used to check the significance of parameters and evaluate the model's fit.

2.4.1 MARS Model Basic Function Simultaneous Test

Simultaneous testing aims to determine whether all basis function coefficients in the MARS model simultaneously exhibit an influence on the response variable by using F-test score.

Hypotheses :

H_0 : $a_m = 0$; $m=1,6,7$
H_1 : at least one of $a_m \neq 0, m=1,6,7$

Statistics test :

$$F_{score}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})/M}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/N-M-1}$$

With critical point where reject H_0 if $F_{score}>F_{\alpha\,(M;N-M-1)}$ or $p-value<\alpha$

2.4.2 MARS Model Basic Function Partial Test

Partial testing aims to determine whether each basis function coefficient in the MARS model, individually, shows an influence on the response variable by using T-test score.

Hypotheses :
H_0 : $a_m = 0$; $m=1,6,7$
H_1 : at least one of $a_m \neq 0, m=1,6,7$

Statistics test :

$$t_{score}=\frac{\hat{a}_m}{Se(\hat{a}_m)}$$

Where $Se(\hat{a}_m)=\sqrt{var(\hat{a}_m)}$
With critical point where reject H_0 if $t_{score}>F_{\left(\frac{\alpha}{2};N-M\right)}$ or $p-value<\alpha$

2.5 Variable Importance Level Test

The importance level of variables is one of the outputs generated by the MARS application. The importance level of variables is used to rank predictor variables that influence the response variable . The criteria used to estimate the importance of variables in MARS models are nsubsets, GCV (Generalized Cross Validation), and RSS (Residual Sum of Squares). The nsubsets criterion calculates the number of model subsets that include a particular variable, assuming that variables appearing in more subsets are considered more important. With the RSS criterion, the total decrease in RSS is computed for each subset when adding a variable, where variables causing a greater decrease in RSS are deemed more significant. A similar approach applies to the GCV criterion, where an increase in GCV value indicates a detrimental impact on the model. For ease of interpretation, decreases in RSS or GCV values are scaled such that the largest decrease corresponds to a scale of 100.

2.6 Logistic Binary Regression Model Estimation

Binary logistic regression is used to model the probability of a certain event occurring based on the influence of influencing factors. The binary logistic regression model is employed to depict the connection between a binary dependent variable and numerous independent variables, whether they are continuous or categorical [13]. In binary logistic regression, the response variable Y follows a Binomial distribution [14]. The binary logistic regression model used is [15].

$$\pi(x_i) = \frac{\exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)} \tag{4}$$

2.7 Logistic Binary Model Suitability Test

Model fit test is a procedure used to determine whether there is a difference between predictions and observed results (whether the model is appropriate or not) [15].

Hypotheses:
H_0 : Model fits (no difference between observed and predicted results)
H_1 : Model does not fit (there is a difference between observed and predicted results)

Statistics test :

$$\hat{\hat{C}} = \sum_{k=1}^{g} \frac{(o_k - n'_k \overline{\pi_k})^2}{(n'_k \overline{\pi_k})(1 - \overline{\pi_k})} \tag{5}$$

2.8 Logistic Binary Regression Model Signification Test

2.8.1 Model Coefficient Partial Test

Wald Testing is conducted to evaluate the influence of each coefficient β_j either partially or individually [15].

Hypotheses:
H_0 : $\beta_j = 0, j = 1,2,\ldots,p$ (there is no influence between the independent and dependent variable)
H_1 : There is at least one $\beta_j \neq 0,\ j = 1,2,\ldots,p$ (There is at least one independent variable that influence the dependent variable)

Statistics test :

$$W_j = \left[\frac{\widehat{\beta_j}}{se\left(\widehat{\beta_j}\right)}\right]^2 \tag{6}$$

2.8.2 Model Coefficient Simultaneous Test

Likelihood Ratio Testing is conducted to determine and assess the significance of the coefficient β on the dependent variable simultaneously [15].

Hypotheses:
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (there is no influence between the independent and dependent variable)
H_1 : There is at least one $\beta_j \neq 0,\ j = 1,2,\ldots,p$ (There is at least one independent variable that influence the dependent variable)

Statistics test :

$$G^2 = -2\ln\left(\frac{L(\widehat{\omega})}{L(\widehat{\Omega})}\right) \tag{7}$$

2.9 Multicollinearity Test

Multicollinearity is one of the assumptions that must be met in binary logistic regression. The data must be non-multicollinear to be processed into advanced binary logistic regression stages. Multicollinearity detection can be done using the VIF (Variance Inflation Factors) value. The VIF value should be less than 10 to ensure there is no multicollinearity among variables [16]. The formula for VIF is [17]

$$VIF = \frac{1}{1-R^2} \tag{8}$$

2.10 Comparison between MARS and Binary Logistic Regression

In the world of classification, accuracy plays a crucial role in assessing how well a model can identify and predict data. Therefore, it is important to check the classification accuracy (APPER). Apparent Error Rate (APPER) describes the proportion of incorrect classifications relative to the total [18].

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

Before modeling someone suffering from hypertension, we will first describe the variables studied. The general description of the data is as follows

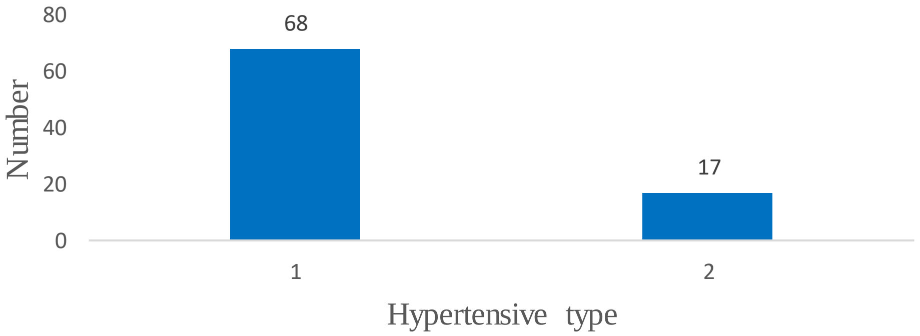

Figure 1. Bar Chart of Hypertension Types

Based on the Figure 1 it can be concluded that of the 85 respondents who took part in the survey, there were 68 respondents who suffered from hypertension and 17 respondents who did not suffer from hypertension

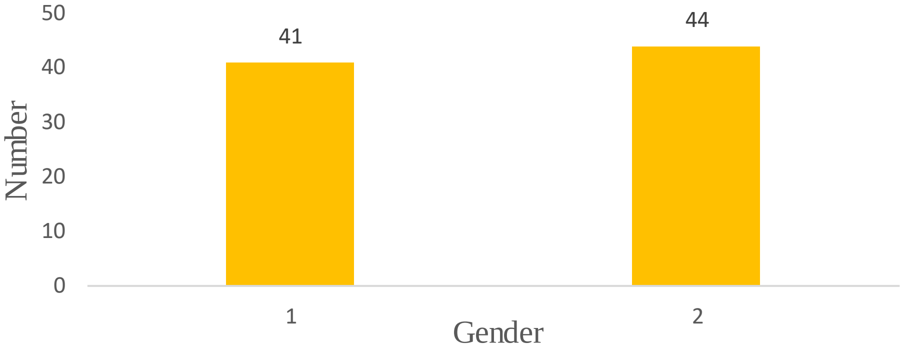

Figure 2. Gender Bar Chart

Based on the **Figure 2**, it can be concluded that of the 85 correspondents who took part in the survey, there were 41 respondents who had gender 1, namely men and 44 respondents who had gender 2, namely women.

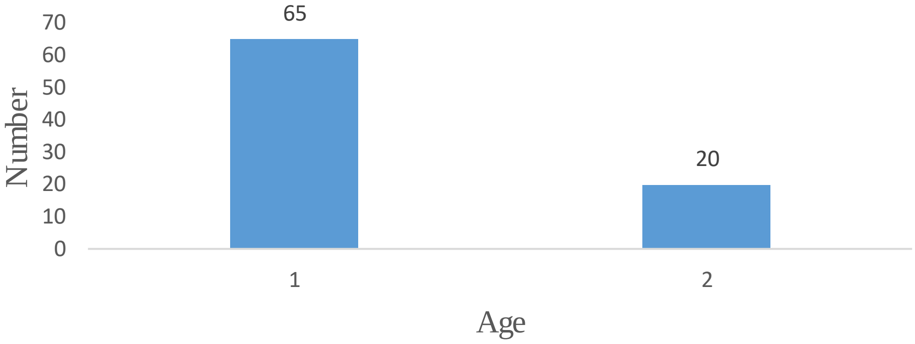

Figure 3. Age Bar Chart

Based on **Figure 3**, it can be concluded that of the 85 respondents who took part in the survey, there were 65 respondents who were included in age category 1 and 20 respondents who were included in age category 2.

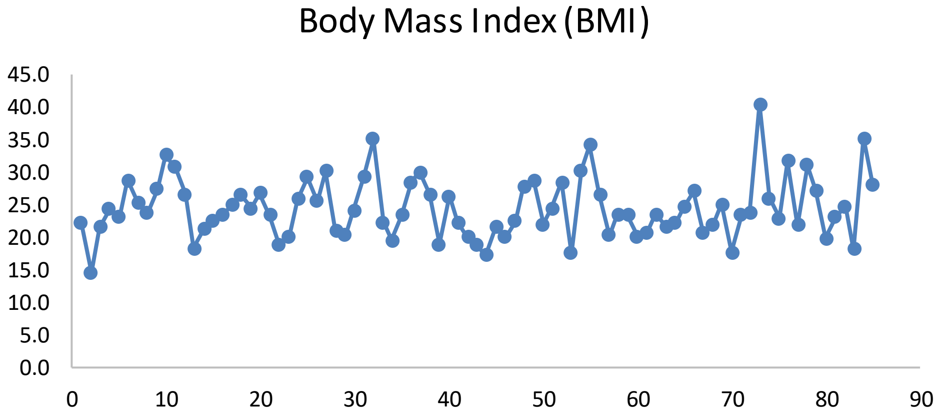

Figure 4. Body Mass Index Scatter Chart

Based on **Figure 4**, it can be concluded that of the 85 respondents who took part in the survey, there were the Body Mass Index (BMI) value obtained from respondents was between 40.5 and 14.6.

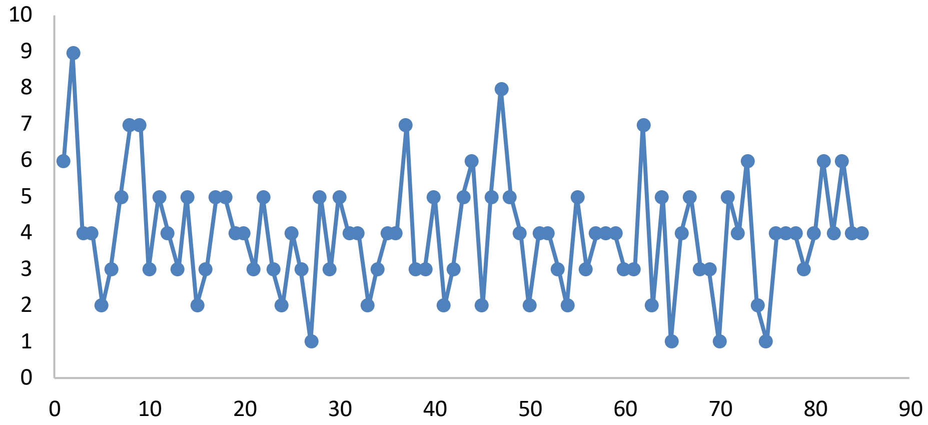

Figure 5. **Physiological Condition Score Scatter Chart**

Based on **Figure 5**, it can be concluded that of the 85 respondents who took part in the survey, there were the Physiological condition value obtained from respondents was between 1 and 9.

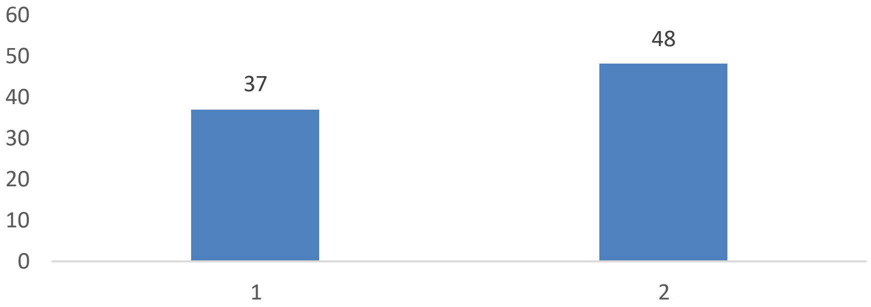

Figure 6. **Physical Activity Score Bar Chart**

Based on **Figure 6**, it can be concluded that of the 85 respondents who took part in the survey, there were 37 respondents who were included in physical activity score category 1 and 48 respondents who were included in age category 2.

3.2 Modeling with MARS

The MARS (Multivariate Adaptive Regression Splines) method has the advantage of handling high-dimensional data by overcoming the "curse of Dimensionality" through selecting and breaking down data into relevant segments. As a non-parametric method, MARS does not require assumptions about variable relationships, so it can capture complex and non-linear relationships that are difficult to handle by linear regression. The flexibility of MARS in using the basis of adaptive spline functions allows the model to adapt well to data patterns, providing accurate and easy-to-interpret results. Therefore, this study uses the MARS method. In this study, the MARS method was used to model the incidence of hypertension based on 5 predictor variables. Calculations are carried out with the help of the MARS application by combining Basis Function (BF), Maximum Interaction (MI), and Minimum Observation between knots (MO). The criterion for the best model is the lowest Generalized Cross Validation (GCV) value. If there is a similarity in the minimum GCV value, it can be seen from the maximum R^2 value and the minimum Mean Square Error (MSE) value. The best model resulting from the combination is at the base ten function value. The results of the combination of base ten functions are as follows.

Table 3. **Base Ten Function Combinations**

BF	MI	MO	R-Square	MSE	GCV
10	1	0	0.22	0.114	0.128
		1	0.23	0.113	0.126
		2	0.222	0.114	0.127

BF	MI	MO	R-Square	MSE	GCV
		3	0.22	0.114	0.128
	2	0	0.124	0.118	0.143
		1	**0.318**	**0.092**	**0.112**
		2	0.266	0.101	0.12
		3	0.254	0.101	0.122

Based on Table 3, it can be seen that the best model is obtained from a combination of base functions of 10, maximum interactions of 2, and minimum observations between knots of 1. The best model estimates for modeling hypertension in Indonesia is showed in Table 4.

Table 4. **Best Model Basis Function Estimation**

Basis Function (BF)	Parameter Estimation
BF_0	1.585
$\mathrm{BF}_1 = (X_2 = 1)$	-0.514
$\mathrm{BF}_6 = \max(0, 19.575 - X_3)\mathrm{BF}_2$ where $\mathrm{BF}_2 = (X_2 = 2)$	-0.344
Basis Function (BF)	**Parameter Estimation**
$\mathrm{BF}_7 = (X_1 = 1)BF4$ where : $\mathrm{BF}_4 = \max(0{,}7.000 - X_4)$	0.801

Based on the Table 4, the following MARS model was obtained to estimate hypertension in Indonesia:

$$\hat{Y} = 1.585 - 0.514\ \mathrm{BF}_1 - 0.344\ \mathrm{BF}_6 + 0.801\ \mathrm{BF}_7$$

In this study. the response variable consists of two values so that a model of the probability of someone getting hypertension is obtained with the following equation.

$$P(Y = 2) = \pi((x) = \frac{\exp(1.585 - 0.514\ \mathrm{BF}_1 - 0.344\ \mathrm{BF}_6 + 0.801\ \mathrm{BF}_7)}{1 + \exp(1.585 - 0.514\ \mathrm{BF}_1 - 0.344\ \mathrm{BF}_6 + 0.801\ \mathrm{BF}_7)}$$

3.2.1 MARS Model Significance Test

1. Simultaneous test

The simultaneous test aims to determine whether all the basis function coefficients in the MARS model simultaneously show an influence on the response variable. The test statistical results from this test were obtained using the help of the MARS application and the results obtained were as follows.

Table 5. **Simultaneous Test of MARS Model Coefficients**

Statistics Test	Value
F-test score	22.197
$P - value$	0.138123×10^{-9}

Based on the Table 5, the resulting $p - value$ 0.138123×10^{-9}, which is less than the significance level ($\alpha = 0.05$). Therefore, the decision obtained is to reject H_0, so the conclusion is that there is at least one a_m that is not equal to zero, with $m = 1, 6$, and 7. This can be interpreted that the model obtained is appropriate and shows that there is a relationship between the coefficients basis functions with response variables.

2. Partial test

The partial test aims to determine whether each basis function coefficient in the MARS model partially shows an influence on the response variable. The test statistical results from this test were obtained using the MARS application and the results obtained were as follows.

Table 6. Partial Test of MARS Model Coefficients

Parameter	Estimate	S.E.	T-Ratio	P-value	Decision
Constant	1.585	0.077	20.519	0.999201×10^{-15}	Reject H_0
BF_1	-0.514	0.085	-6.019	0.488104×10^{-7}	Reject H_0
BF_6	-0.344	0.098	-3.509	0.738362×10^{-3}	Reject H_0
BF_7	0.801	0.156	5.141	0.185867×10^{-5}	Reject H_0

Based on the Table 6, the $p - value$ obtained for each basis function in the model is less than the significance level ($\alpha = 0.05$). Therefore, the decision obtained is to reject H_0, so the conclusion is that the values a_1, a_6 and a_7 are not equal to zero. This can be interpreted that the model obtained shows a relationship between the basis function coefficients and the response variable.

3.2.2 Variable Importance Level

The level of variable importance is used to rank the predictor variables that influence the response variable. The level of importance of variables in modeling Gini ratio data in Indonesia is as follows.

Table 7. Variable Importance Level

Variable	Variable Name	Level of Importance	GCV reduction
X_2	Age	100 %	0.141
X_1	Sex	93.774 %	0.137
X_4	Psychological condition	93.774 %	0.137
X_3	BMI	51.249 %	0.119
X_5	Physical activity	0 %	0.112

Based on the Table 7, it can be seen that the predictor variable that has the most influence on the response variable is the Age variable with an importance level of 100% and BMI can reduce the Generalized Cross Validation (GCV) value by 0.141. Meanwhile, the predictor variable which has an importance level of 0% is the Physical Activity variable. If we look at the contribution of predictor variables in reducing GCV, it is found that the physical activity variable can reduce the GCV value by 0.112.

3.3 Binary Logistic Regression Modeling

Binary logistic regression modeling is carried out according to the steps explained in section 2.6. Based on the analysis results, the binary logistic regression probability equation is obtained as follows.

$$(Y = 2|X_1 = 1, X_2 = 1) = \frac{\exp(0.2460 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}{1 + \exp(0.2460 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}$$

$$P(Y = 2|X_1 = 1, X_2 = 2) = \frac{\exp(-2.266 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}{1 + \exp(-2.266 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}$$

$$P(Y = 2|X_1 = 2, X_2 = 1) = \frac{\exp(0.4380 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}{1 + \exp(0.4380 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}$$

$$P(Y = 2|X_1 = 2, X_2 = 2) = \frac{\exp(-2.074 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}{1 + \exp(-2.074 - 0.1021\,X_3 + 0.3524\,X_4 + 0.1705\,X_5)}$$

3.3.1 Multicollinearity Test

Multicollinearity is one of the assumptions that must be met in binary logistic regression.

Table 8. VIF values for predictors

Variable	VIF
X_1	1.09
X_2	1.25

Variable	VIF
X_3	1.03
X_4	1.27
X_5	1.29

Based on the table above, it is found that there is no VIF value more than 10, so it can be concluded that there is no multicollinearity between variables.

3.3.2 Model Suitability Test

In this test, the Hosmer and Lemeshow test was carried out. The results of the analysis are as follows.

Table 9. Fit Test of the Binary Logit Regression Model

Test	*DF*	*Chi-Square*	*P-value*
Deviance	79	66.16	0.848
Pearson	79	90.28	0.181
Hosmer-Lemeshow	8	4.73	0.786

Based on the **Table 9**, the Hosmer-Lemeshow $p - value$ is 0.786, which is more than 0.05. Then a decision can be made that it fails to reject H_0. So, it can be concluded that the model is declared fit for the data.

3.3.3 Significance Test of the Binary Logistic Regression Model

The results of significance testing are available in the following table.

Table 10. Logit Regression Significance Test

Source	*DF*	*Chi-Square*	*P-Value*
Regression	5	14.59	0.012
X_1	1	0.09	0.767
X_2	1	12.94	0.000
X_3	1	1.40	0.236
X_4	1	3.68	0.055
X_5	1	0.82	0.365

1. Simultaneous Test

Based on the **Table 10**, the $p - value$ is 0.012, which is less than 0.05. Then a decision can be taken to reject H_0. So, it can be concluded that there is at least one variable that influences the model.

2. Partial test

Based on the **Table 10**, the $p - value$ for X_1, then a decision can be made that it fails to reject H_0. So, it can be concluded that the variables X_1, X_3, X_4, X_5 do not have a significant influence. Meanwhile, the variable X_2 has a $p - value$ of 0.000, which is less than 0.05 and means that the age variable has a significant influence on the model.

3.3.4 Model Accuracy

In the MARS modeling, it was found that a total of 13 data points had incorrect predictions, while 85 other data points had predictions that matched the actual data. The APPER value is obtained as follows:

$$APPER = \frac{13}{85} = 0{,}15294$$

Thus, the classification accuracy of the Gompit model is:

$$(1-APPER)x100\% = (1-0{,}15294)\times 100\% = 84.706\ \%$$

In the Binary Logistic Regression modeling, it was found that a total of 17 data points had incorrect predictions, while 68 other data points had predictions that matched the actual data. The APPEAR value is obtained as follows:

$$APPER = \frac{17}{85} = 0.2$$

Thus, the classification accuracy of the Gompit model is:

$$(1-APPER)\times 100\% = (1-0.2)\ x\ 100\% = 80\ \%$$

3.4 Comparison of the Goodness of the MARS Model and Binary Logistic Regression

In the world of classification, accuracy plays an important role in assessing how well a model can identify and predict data. Therefore, it is important to check the accuracy of the classification (APPEAR). In MARS modeling, it was found that the total data that had incorrect predictions was 13 data while 85 other data had predictions that were the same as the actual data. So the classification accuracy of the logit model is 84.706%. Meanwhile, in Binary Logit Regression modeling, it was found that the total data that had incorrect predictions was 17 data while 68 other data had predictions that were the same as the actual data. So, the classification accuracy of the logit model is 80%.

4. CONCLUSIONS

Based on the results of the analysis and discussion that have been written previously, the conclusions of this research are obtained, namely:

1. In the response variable (Y) hypertension, it can be seen that of the 85 total respondents there are 68 respondents who suffer from hypertension and 17 respondents who do not suffer from hypertension.
2. The best MARS model is obtained from a combination of base 10 functions, maximum interaction of 2 and minimum observation between knots of 1. With a Generalized Cross Validation (GCV) value of 0.112, an R2 value of 0.318 and a Mean Square Error (MSE) of 0.092. Based on the estimation of the basis function, the model equation is obtained as follows:
$$\hat{Y} = 1.585 - 0.514\,BF1 - 0.344\,BF6 + 0.801\,BF7$$
3. Binary logistic regression modeling obtained a Hosmer-Lemeshow p-value of 0.119 where this value was more than 0.05, which means that it failed to reject H_0 so that the model was declared suitable for the data. Apart from that, the model equation used is as follows:
$$P(Y=2|X_1=1,X_2=1) = \frac{\exp(-0.2460\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}{1+\exp(-0.2460\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}$$
$$P(Y=2|X_1=1,X_2=2) = \frac{\exp(2.266\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}{1+\exp(2.266\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}$$
$$P(Y=2|X_1=2,X_2=1) = \frac{\exp(-0.4380\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}{1+\exp(-0.4380\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}$$
$$P(Y=2|X_1=2,X_2=2) = \frac{\exp(2.074\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}{1+\exp(2.074\ +\ 0.1021\,X_3\ -\ 0.3524\,X_4 - \ 0.1705\,X_5)}$$
4. After analyzing using the MARS method and binary logistic regression, it was obtained that the MARS Apparent Error Rate (APPER) value was 84.706%, while the APPER value for binary logistic regression was 80%. So statistically the MARS method is better in classifying hypertension data in Indonesia.

REFERENCES

[1] Kementerian Kesehatan, "Tekanan Darah Tinggi (Hipertensi)," 2016. [Online]. Available: https://p2ptm.kemkes.go.id/uploads/2016/10/Tekanan-Darah-Tinggi-Hipertensi.pdf.

[2] World Health Organization, "Hypertension," World Health Organization, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[3] Kementerian Kesehatan, Hasil Riset Kesehatan Dasar Tahun 2018, Jakarta: Kementerian Kesehatan, 2023.

[4] N. Salsabila, N. Indraswari, B. Sujatmiko, "Gambaran Kebiasaan Merokok di Indonesia Berdasarkan Indonesia Family Life Survey 5 (IFLS 5)," *Jurnal Ekonomi Kesehatan Indonesia,* vol. 7, p. 1, 2022.

[5] R. Kurniawan & B. Yuniarto , Analisis Regresi, Jakarta: Prenada Media, 2016.

[6] I. Budiantara & I. Wulandari, "Analisis Faktor-Faktor yang Mempengaruhi Persentase Penduduk Miskin dan Pengeluaran Perkapita Makanan di Jawa Timur menggunakan Regresi Nonparametrik Birespon Spline," *Jurnal Sains dan Seni ITS,* vol. 3, p. 1, 2014.

[7] M. Y., "Pemodelan Multivariate Adaptive Regression Spline (MARS) Pada Faktor-Faktor yang Mempengaruhi Kemiskinan di Provinsi Maluku dan Maluku Utara," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika,* vol. 13, no. 1, pp. 8-14, 2020.

[8] M. E. Y. Naiborhu, "Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Spline (MARS) Pada Kasus Penyakit Hipertensi," 2021.

[9] W. Rumaenda, Y. Wulandari, & D. Safitri, "Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4. 5 (Studi KasusUPT Puskesmas Ponjong I, Gunungkidul)," *Jurnal Gaussian,* vol. 5, no. 2, pp. 299-309, 2016.

[10] Cooper & Schindler, Business Research Methods Tenth edition, New York, 2008.

[11] F. J., "Multivariate Adaptive Regression Splines," *The Annals of Statistics,* vol. 19, no. 1, pp. 1-67, 1991.

[12] S. Hidayati & B. W. Otok, "Parameter Estimation and Statistical Test in Multivariate Adaptive Generalized Poisson Regression Splines," *IOP Conference Series: Materials Science and Engineering* , vol. 546, 2019.

[13] N. A. M. R. Senaviratna & T. M. J. A. Cooray, "Diagnosing Multicollinearity of Logistic Regression Model," *Asian Journal of Probability and Statistics,* vol. 5, no. 2, pp. 1-9, 2019.

[14] D. I. Ruspriyanty & A. Sofro, " Analysis of hypertension disease using logistik and probit regression," *Journal of Physics: Conference Series,* vol. 1108, 2018.

[15] Hosmer & Lemeshow, Applied Logistik Regression Second Edition, New York: John Wiley & Sons, Inc, 2000.

[16] D. Belsley, Conditioning diagnostics: Collinearity and weak data in regression, New York: John Wiley & Sons, Inc, 1991.

[17] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *American Journal of Applied Mathematics and Statistics,* vol. 8, pp. 39-42, 2020.

[18] Schwarzer, Vach, & Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in medicine,* vol. 19, no. 4, pp. 541-561, 2000.