Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:51

https://doi.org/10.28919/cmbn/9209

ISSN: 2052-2541

TRUNCATED SPLINE REGRESSION FOR BINARY RESPONSE: A COMPARATIVE STUDY OF NONPARAMETRIC AND SEMIPARAMETRIC APPROACHES

AFIQAH SAFFA SURIASLAN[1], I NYOMAN BUDIANTARA[2,*], VITA RATNASARI[2]

[1]Doctoral Study of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Kampus ITS- Sukolilo, Surabaya 60111, Indonesia

[2]Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Kampus ITS- Sukolilo, Surabaya 60111, Indonesia

Abstract: Type 2 diabetes mellitus is a chronic metabolic disorder with a growing global prevalence, especially in developing countries. According to the International Diabetes Federation, Indonesia ranks fifth in the world for the highest number of diabetes cases, with 19.5 million adults affected by 2021. Early detection and intervention strategies are critical in managing this disease, and predictive models play a vital role in identifying individuals at high risk. Recent advances in regression analysis have introduced nonparametric and semiparametric regression methods, particularly truncated spline-based regression, which offer greater flexibility in capturing complex relationships in data. This study compares the performance of nonparametric and semiparametric truncated spline regression models in classifying binary response variables, specifically in predicting type 2 diabetes mellitus status. The models were evaluated using deviance values and classification accuracy metrics, including sensitivity, specificity, and precision. The results showed that the semiparametric truncated spline regression model outperformed the nonparametric approach, with lower deviance values (42.46 vs 52.94) and higher classification accuracy (86.67% vs 76.67%). In addition, the semiparametric model showed better sensitivity (97.44% vs 92.31%), specificity (66.67% vs 47.62%),

*Corresponding author

E-mail address: i_nyoman_b@statistika.its.ac.id

Received February 22, 2025

and precision (84.44% vs 76.60%), indicating a greater ability to correctly classify diabetic and non-diabetic individuals.

Keywords: type 2 diabates mellitus; nonparametric; semiparametric; truncated spline; binary response.

2020 AMS Subject Classification: 62J02, 62G05.

1. INTRODUCTION

Type 2 diabetes mellitus is a chronic metabolic disorder whose prevalence continues to increase worldwide. The condition is rapidly evolving into an epidemic in many countries, with the number of sufferers expected to double in the coming decades due to an increase in the ageing population, further burdening healthcare systems, especially in developing countries [1]. According to the Diabetes Atlas 2021 published by the international diabetes federation, there are an estimated 537 million people living with diabetes worldwide [2]. Based on the latest data from the international diabetes federation in 2021, Indonesia ranks fifth as the country with the highest number of people with diabetes, which is 19.5 million adults with a prevalence of 10.8%, where most cases of diabetes in Indonesia are type 2 Diabetes Mellitus. Type 2 diabetes mellitus is a metabolic disorder characterized by elevated blood sugar levels due to reduced insulin secretion by pancreatic beta cells and/or impaired insulin function. Insulin resistance commonly occurs in obese individuals and is marked by diminished insulin action in the liver and decreased glucose uptake in adipose and muscle tissues [3]. Although lifestyle modifications and pharmacological interventions have been recommended as preventive measures, they have not been entirely effective in curbing the rising prevalence of diabetes. Therefore, gaining a deeper understanding of the factors contributing to type 2 diabetes mellitus is essential for developing more effective treatment strategies. According to a World Heatlh Organization, in 2016, at least 41 million children under the age of five were overweight or obese (BMI ≥ 35 kg/m^2). If this trend continues, an estimated 60% of the global population will be overweight or obese by 2030 [4].

With the increasing prevalence of type 2 diabetes mellitus, the development of more effective early detection and intervention strategies is crucial. Prediciton models play a vital role in identifying individuals at high risk of developing diabetes, thereby supporting informed clinical decision-making. Numerous predictive models have been proposed to assess and quantify diabetes risk factors. For example, Razavian et al. [5] developed a logistic regression based prediction model to estimate the incidence of type 2 diabetes, the developed model serves as a screening tool to identify individuals at high risk of the disease. Meanwhile, Zou et al. [6] applied machine

learning methods in predicting diabetes in Luzhou, China, with five-fold cross-validation to evaluate the performance of their model. In addition, Joshi and Dhakal [7] used the decision tree algorithm, one of the machine learning methods, to model the incidence of type 2 diabetes.

Truncated spline based nonparametric and semiparametric regression has been widely applied in data analysis in recent years. These methods leverage advances in estimation techniques to capture more flexible relationship patterns compared to parametric approaches. Nonparametric regression using truncated splines is particularly advantageous in identifying varying relationship structures within specific sub-intervals [8]. Meanwhile, semiparametric regression integrates both parametric and nonparametric components, striking a balance between model interpretability and estimation flexibility. Several studies [9], [10] have applied truncated spline estimators in nonparametric regression. However, in many cases, the data may exhibit a partially known curve pattern, and also partially undergo changes that occur at certain sub-intervals. This condition encourages the development of semiparametric truncated spline regression, which combines the flexibility of nonparametric with parametric approaches. While a number of studies [11], [12], [13] have applied semiparametric truncated spline regression, but still focused on cases with quantitative response variables. In fact, in practice, there are often situations where the response variable is binary. As a result, the semiparametric truncated spline regression models that have been developed have not been able to fully handle cases with binary response variables. As scientific advancements continue, study [14] has explored a nonparametric approach based on truncated spline to analyze data with binary response variables, opening wider opportunities in the application of this method in various fields of study.

Although there have been many studies comparing parametric and nonparametric regression, research on comparing the performance of truncated spline based nonparametric and semiparametric regression in categorical data analysis is still limited. Therefore, this study aims to compare the performance of the two approaches in modeling data with binary response variables, which is applied to data of status type 2 diabetes mellitus at Haji General Hospital Surabaya, Indonesia. The evaluation is done by using the deviance value and also the classification accuracy. Through this research, it is hoped that an overview of the advantages and limitations of nonparametric and semiparametric truncated spline regression in categorical data analysis can be obtained, as well as providing recommendations regarding the selection of models that are more suitable for the characteristics of the data used.

2. PRELIMINARIES

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

2.1. Dataset

The dasatet of type 2 diabetes mellitus, age, body mass index, and waist circumference are presented in Table 1.

Table 1. Dataset of Research

Patient Number	Type 2 Diabetes Milletus Status (Y)	Age (X_1)	Body Mass Index (X_2)	Waist Circumference (X_3)
1	1	61	23.42	91
2	1	51	22.66	90
3	1	52	29.33	82
⋮	⋮	⋮	⋮	⋮
60	0	71	24.02	90

Table 1 presents an overview of the dataset utilized in this study. The dataset comprises type 2 diabetes milletus status as the response variable, along with age, body mass index, and waist circumference as predictor variables, which are considered to influence diabetes. The data were collected from 60 patients. The response variable, type 2 diabetes mellitus status (Y), is measured on an binary scale, where category 0 represents non diabetes and category 1 represents diabetes. While the age variable (X_1), body mass index (X_2) and waist circumference (X_3) has a ratio scale.

2.2. Nonparametric Truncated Spline Regression

Spline regression is a polynomial regression analysis method that has continuous segmented properties. In nonparametric regression, spline has high flexibility so that it has the ability to estimate data behavior that tends to be different [8]. In a truncated spline, there are two components, namely the polynomial component and the truncated component. In this case, the polynomial has the property of being divided into several intervals formed by knots.

If the variable function is approximated using a truncated spline function of degree m and knot points $K_{1j}, K_{2j,\dots,}K_{rj}$, where j is 1, 2, ... , p. Then it can be written into the following equation [15]:

$$f(x_{1i}, x_{2i}, \dots, x_{pi}) = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}(x_{ji} - K_{ju})_{+}^{m} \quad (1)$$

with $i = 1,2, \dots, n$

The Truncated function is given by:

$$(x_{ji} - K_{ju})_+^m = \begin{cases} (x_{ji} - K_{ju})^m, x_{ji} \geq K_{ju} \\ 0 \qquad\qquad\quad , x_{ji} < K_{ju} \end{cases} \tag{2}$$

Where β_0, β_{jk}, and $\beta_{j(m+u)}$, $j = 1,2,\dots,p$, $k = 1,2,\dots,m$, $u = 1,2,\dots,r$ are the model parameters in the Truncated Spline function.

2.3. Semiparametric Truncated Spline Regression

Semiparametric regression merges both nonparametric and parametric elements. Given paired data (v_{ji}, w_{ji}, y_i) and the relationship between x_{ji}, m_{ji}, and y_i is assumed to follow a semiparametric regression model [16].

$$y_i = g(v_{ji}) + f(x_{ji}) + \varepsilon_i \, , i = 1,2,\dots,n \tag{3}$$

where y_i represents the response variable, $g(v_{ji})$ is a parametric function, and $f(x_{ji})$ is a nonparametric function and ε_i is the error or unexplained variability not accounted for by the model components.

In nonparametric regression, splines provide flexibility, enabling them to model data with varying patterns. In this case, this polynomial has the property of being divided into several intervals formed by knot points. The knots will indicate the truncated function attached to the estimator [17]. Based on equation (3) Truncated Spline semiparametric regression can generally be formulated as follows:

$$g(v_{ji}) = \delta_0 + \delta_1 v_{1i} + \delta_2 v_{2i} + \cdots + \delta_q v_{qi}$$

$$f(x_{ji}) = \beta_0 + \sum_{j=1}^{p} \sum_{k=1}^{m} \beta_{jk} x_{ji}^k + \sum_{j=1}^{p} \sum_{u=1}^{r} \beta_{j(m+u)} (x_{ji} - K_{ju})_+^m$$

where the Truncated function is given by:

$$(x_{ji} - K_{ju})_+^m = \begin{cases} (x_{ji} - K_{ju})^m, x_{ji} \geq K_{ju} \\ 0 \qquad\qquad\quad , x_{ji} < K_{ju} \end{cases} \tag{4}$$

So that the Truncated Spline Semiparametric Regression can be written as:

$$y_i = \delta_0 + \delta_1 v_{1i} + \cdots + \delta_q v_{qi} + \sum_{j=1}^{p} \sum_{k=1}^{m} \beta_{jk} x_{ji}^k + \sum_{j=1}^{p} \sum_{u=1}^{r} \beta_{j(m+u)} (x_{ji} - K_{ju})_+^m + \varepsilon_i \tag{5}$$

2.4. Truncated Spline Regression for Binary Response

Given $x_{1i}, x_{2i}, \dots, x_{pi}$; $i = 1, 2, \dots, n$, are as many as p predictor variables. Furthermore, response variable (Y_i) is bernoulli distributed, with a probability distribution of

[18]:

$$Y_i \sim B\left(1, \pi\left(x_{1i},\ x_{2i},\ \dots,\ x_{pi}\right)\right),\ i = 1,2,\dots,n$$

with the probability function [18]:

$$P(Y_i\ =\ y_i) =\ \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i};\ y_i = 0,1\ ;\ i = 1,2,\dots,n \tag{6}$$

Where $\pi(x_i)$ is defined in the probability distribution function $P(Y_i\ =\ y_i)$ as follows:

$$P(Y_i\ =\ y_i) =\ \ = \left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i}(1-\pi(x_i)) \tag{7}$$

In the context of regression for binary response data, the logit function is a tool to transform the nonlinear relationship between predictor variables and probabilities into a linear relationship. From equation (7), then we made in the natural logarithm function (ln)

$$\ln P(Y_i\ =\ y_i) =\ y_i\ \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) + \ln(1-\pi(x_i)) \tag{8}$$

When made in exponential form, equation (8) forms an exponential family distribution function

$$\exp(\ln P(Y_i\ =\ y_i))\ = \exp\left(y_i\ \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) + \ln(1-\pi(x_i))\right) \tag{9}$$

where, the distribution function of the exponential family is defined as follows:

$$f(y_i, z) = \exp\left(\frac{y_i.h - b(h)}{a(\emptyset)} + c(z,\emptyset)\right)$$

Thus,

$$P(Y_i\ =\ y_i) = \exp\left(\frac{y_i\ \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) - (-\ln(1-\pi(x_i))\ )}{1}\right) \tag{10}$$

where, the logit function is obtained

$$h = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \tag{11}$$

Equation (11) is a link function used to simplify the logistic regression model to facilitate parameter estimation. To achieve this goal, logit transformation is used.

$$h = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$$

$$\ln(\exp(h)) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$$

$$\exp(h) = \frac{\pi(x_i)}{1-\pi(x_i)}$$

$$\exp(h) = \pi(x_i) + \exp(h)\,\pi(x_i)$$

$$\pi(x_i) = \frac{\exp(h)}{1+\exp(h)} \tag{12}$$

The Logistic Regression model can be written as follows equation (12) and logit transformation of $\pi(x_i)$ is defined as follows:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = f(x_{1i},\, x_{2i}, \dots, x_{pi}) \tag{13}$$

$f(x_{1i},\, x_{2i}, \dots, x_{pi})$ is approximated by nonparametric and semiparametric truncated spline function with knot points $K_{1j}, K_{2j, \dots, }K_{rj}$, where j is 1, 2, …, p.

The logit equation of onparametric truncated spline for binary response is obtained as follows:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m} \tag{14}$$

The logit equation of semiparametric truncated spline for binary response is obtained as follows:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \delta_0 + \delta_1 v_{1i} + \cdots + \delta_q v_{qi} + \sum_{j=1}^{p}\sum_{k=1}^{m}\beta_{jk}x_{ji}^{k} + \sum_{j=1}^{p}\sum_{u=1}^{r}\beta_{j(m+u)}\left(x_{ji} - K_{ju}\right)_{+}^{m} + \varepsilon_i \tag{15}$$

Truncated function for Spline is defined as follows:

$$\left(x_{ji} - K_{ju}\right)_{+} = \begin{cases}\left(x_{ji} - K_{ju}\right) & , x_{ji} \geq K_{ju} \\ 0 & , x_{ji} < K_{ju}\end{cases}$$

The logit function of truncated spline function can be presented in matrix form:

$$\begin{bmatrix} 1 & x_{11} & \cdots & {x_{11}}^{m} & (x_{11} - K_{11})^{m}{}_{+} & \cdots & {x_{p1}}^{m} & \cdots & (x_{p1} - K_{pr})_{+}^{m} \\ 1 & x_{12} & \cdots & {x_{12}}^{m} & (x_{12} - K_{11})^{m}{}_{+} & \cdots & {x_{p2}}^{m} & \cdots & (x_{p2} - K_{pr})_{+}^{m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & {x_{1n}}^{m} & (x_{1n} - K_{11})^{m}{}_{+} & \cdots & {x_{pn}}^{m} & \cdots & (x_{pn} - K_{pr})_{+}^{m} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{p(m+r)} \end{bmatrix}$$

2.5. Evaluation of Classification Criteria

To assess the performance of the classification method, various metrics such as Sensitivity, Specificity, Accuracy, and F1-Score are utilized. These metrics are derived from the classification results, which are summarized in the following Confusion Matrix table:

Tabel 2. Tabel *Confusion Matrix*

Actual Group	**Prediction Group**		**Total**
	0	**1**	
0	A	B	A+B
1	C	D	C+D
Total	A+C	B+D	A+B+C+D

Accuracy measures how often the model gives correct predictions for both positive and negative cases overall. It is the ratio of the number of correct predictions to the total number of predictions. Specificity measures the ability of the model to identify true negative cases, i.e. the proportion of all negative cases that are correctly classified by the model. Sensitivity measures how well the model can identify true positive cases. In other words, it is the proportion of all positive cases that are correctly classified by the model. Precision is a measure that indicates how precise the model is in classifying positive cases. Formula for calculating the case of classification criterial can be found in equation as follows:

$$accuracy = \frac{A + D}{A + B + C + D} \tag{16}$$

$$specificity = \frac{D}{C + D} \tag{17}$$

$$sensitivity = \frac{A}{A + B} \tag{18}$$

$$precision = \frac{D}{D + B} \tag{19}$$

3. MAIN RESULTS

This section provides results and discussion of the implementation of methods on type 2 diabetes mellitus data at the Haji general hospital Surabaya.

3.1. Characteristics of Data

In the following, we provide the boxplots for each predictors varible against with the type 2 diabetes mellitus status.

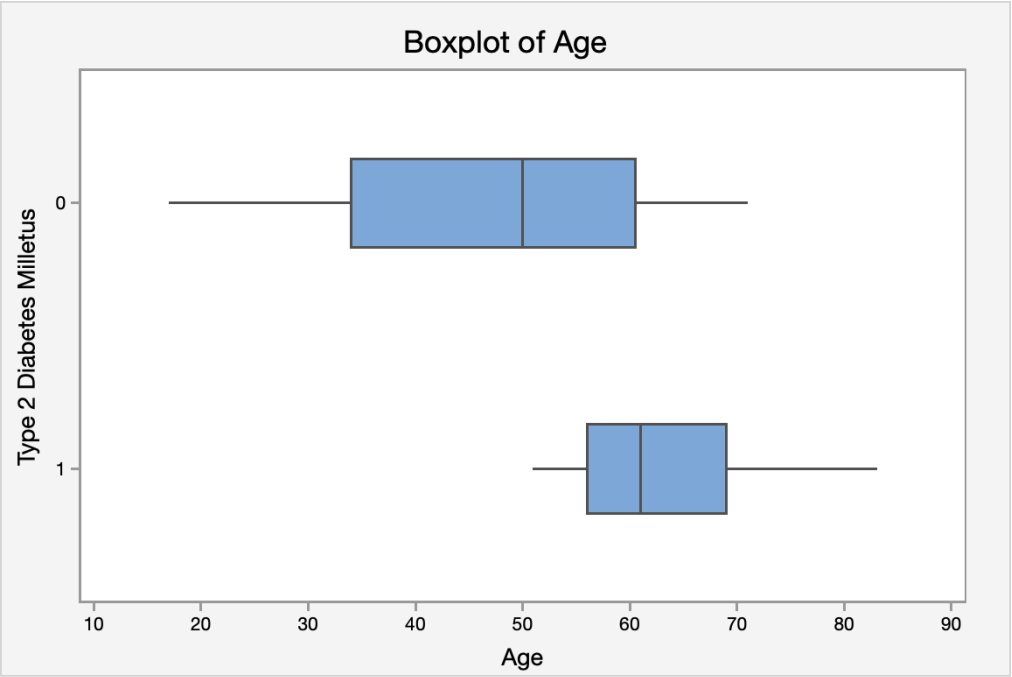

Figure 1. Boxplot for The Age Againt The Type 2 Diabetes Mellitus

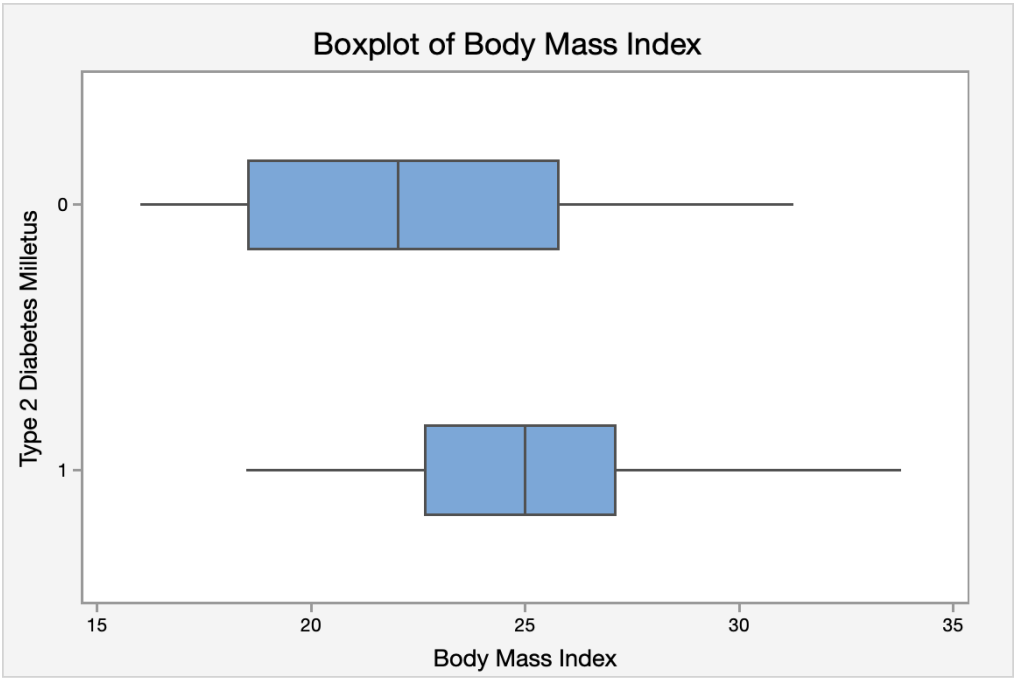

Figure 2. Boxplot for The Body Mass Index Againt The Type 2 Diabetes Mellitus

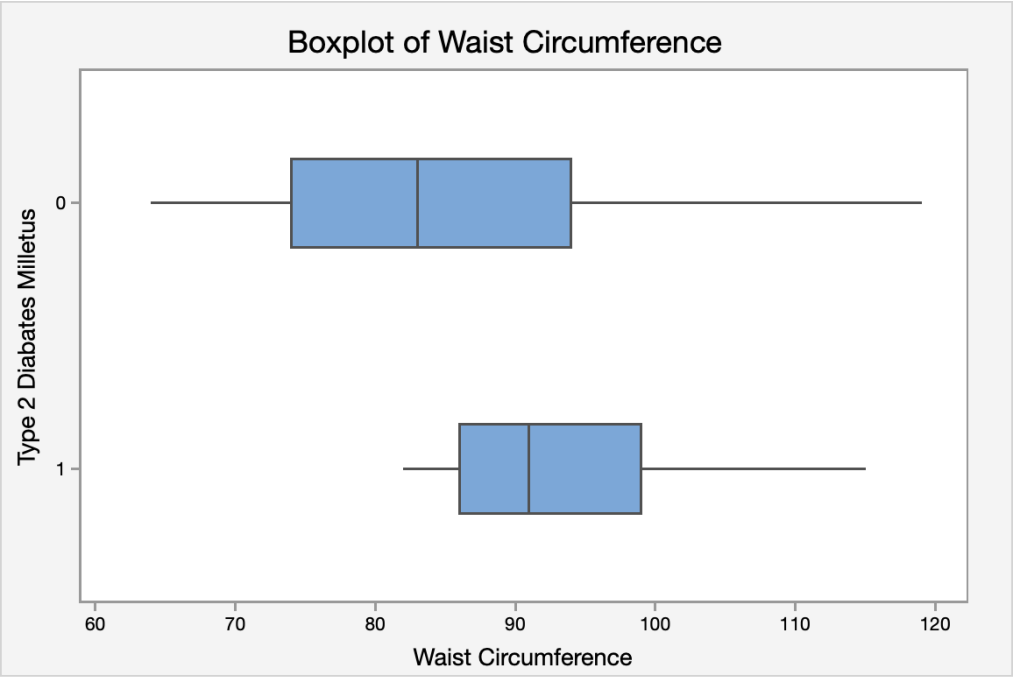

Figure 3. Boxplot for The Waist Circumeference Agaist The Type 2 Diabetes Mellitus

Figure 1 to Figure 3 show boxplots that can provide an indication of the effect of predictor variables on type 2 diabetes mellitus. Figure 1 shows that the interquartile range in category 1 (diabetes) is longer than category 0 (non diabetes), which indicates that age data in individuals with diabetes is more dispersed than individuals non diabetes. Figure 2 also shows that the interquartile range in category 1 (diabetes) is longer than category 0 (non diabetes), so it can be concluded that body mass index data in individuals with diabetes is more spread out than individuals non diabetes. Then Figure 3 shows that the interquartile range in category 0 (non diabates) is longer than category 1 (diabetes), so it can be concluded that waist circumference data in individuals non diabetes is more spread out than individuals with diabetes. Based on this, it can be seen that there are differences in the distribution of age, body mass index, and waist circumference levels between categories of type 2 diabetes mellitus. This can be an indication that the variables of age, body mass index, and waist circumference have an effects on the incidence of diabetes. Furthermore, a general description of the predictor variables is known through descriptive statistics shown in Table 3.

Table 3. Descriptive statistics of age, body mass index, and waist circumference based on type 2 diabetes mellitus

Variable	Type 2 Diabates Mellitus	Total Count	Mean	Variance	Minimum	Maximum	Range
Age	Diabates	39	62.95	76.05	51	83	32
	Non Diabates	21	47.09	27.29	17	71	54
Body Mass Index	Diabates	39	25.50	12.57	18.49	25	15.29
	Non Diabates	21	22.37	20.018	16.02	31.25	15.23
Waits Circumference	Diabates	39	93.39	73.45	82	115	33
	Non Diabates	21	84.95	173.85	64	119	55

Table 3 shows more information about the predictor variables. The Age variable explains that patients with type 2 diabetes have an average age of 62.95 years with a variation of 76.05, and an age range of 32 years from a minimum of 51 years to a maximum of 83 years. Meanwhile, patients non diabetes have an average age of 47.09 years with a variation of 27.29, and an age range of 54 years from a minimum of 17 years to a maximum of 71 years. Thus, it can be concluded that patients with type 2 diabetes have a higher average age than patients non diabetes. The Body Mass Index variable shows that the average body mass index of patients with type 2 diabetes is 25.50, with a variation of 12.57, and a range of 15.29 from a minimum value of 18.49 to a maximum of 25. Meanwhile, the average body mass index of patients non diabetes is 22.37, with a variation of 20.018, and a range of 15.23 from a minimum value of 16.02 to a maximum of 31.25. Thus, patients with type 2 diabetes tend to have a higher body mass index than patients non diabetes. The waist circumference variable showed that patients with type 2 diabetes had an average waist circumference of 93.39, with a variation of 73.45, and a range of 33 from a minimum of 82 to a maximum of 115. Meanwhile, patients non diabetes had a mean waist circumference of 84.95, with a variation of 173.85, and a range of 55 from a minimum of 64 to a maximum of 119. Thus, patients with type 2 diabetes have a larger waist circumference than patients non diabetes. In addition, Figure 4 to Figure 6 provides an overview of the relationship pattern (scatterplot) between

the response variable and the predictor variables, where the predictor variables have been categorized into several interval groups [18].

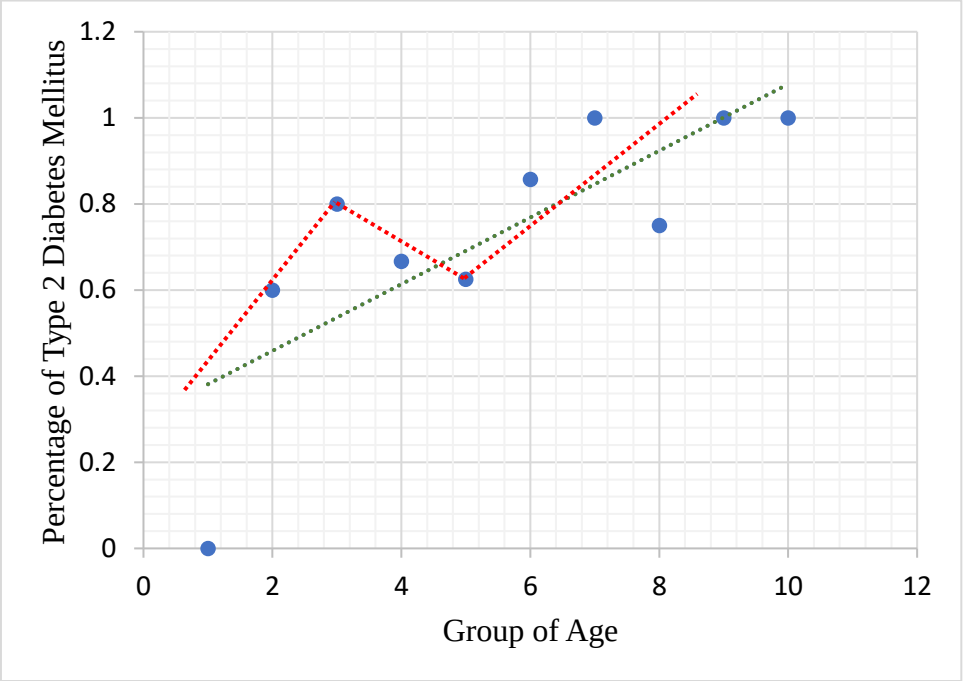

Figure 4. Scatterplot for The Group of Age Agaist The Type 2 Diabetes Mellitus

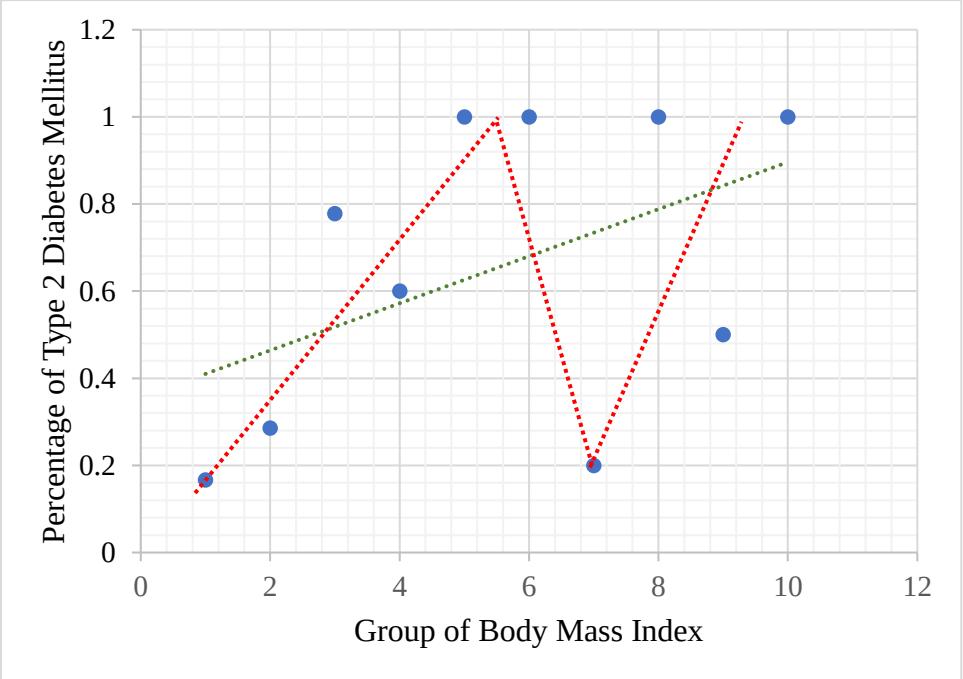

Figure 5. Scatterplot for The Group of Body Mass Index Agaist The Type 2 Diabetes Mellitus

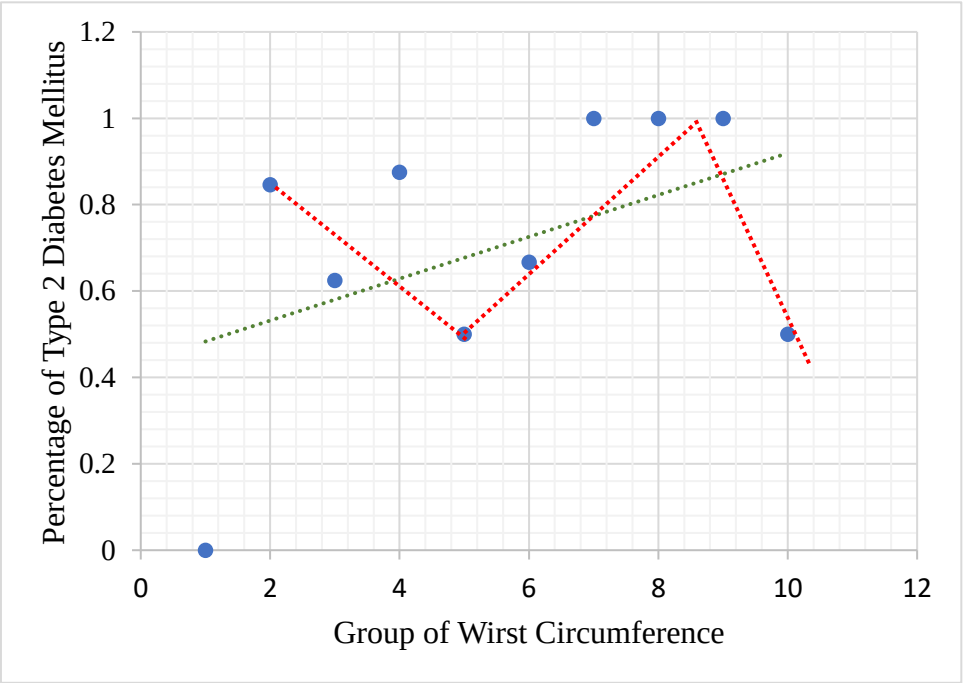

Figure 6. Scatterplot for The Group of Wirst Circumference Agaist The Type 2 Diabetes Mellitus

Figure 4 shows a relationship pattern that most resembles a linear pattern compared to Figure 5 and Figure 6, so the age variable (X_1) is selected as a parametric variable in the semiparametric model. Furthermore, as a nonparametric variable, the data pattern of the age variable in Figure 4 shows an increasing trend in the 3rd data group, then decreases in the 5th data group, and increases again afterwards. Meanwhile, in Figure 5, the relationship between body mass index and response tends to increase until the 5th data group, then decreases in the 7th data group, before showing an increase again. As for Figure 6, the relationship pattern of waist circumference is seen to decrease until the 5th data group, then increase in the 9th data group, and decrease again afterwards. Based on the change in pattern tendency, it can be indicated that the relationship between the variables has changed at certain sub-intervals. Therefore, a suitable model approach is the truncated spline, which is able to capture these pattern changes flexibly.

3.2. Estimation Result of Nonparametric Truncated Spline Regression for Binary Response on Type 2 Diabates Mellitus Data

Furthermore, modeling will be done with a nonparametric approach, which will then be compared with a semiparametric model. Before estimating the parameters of the nonparametric model, it is necessary to determine the knot points to ensure the model reaches the optimal level. In this study, the number of knot points is limited to 3. The list of candidate knot points is presented in the following table.

Table 4. AIC Value based on Knot Point Candidate of Nonparametric Model

Number of Knot Points	K_{ju}		Knot Point Value	AIC
1,1,1	K_{1u}	K_{11}	50	
		K_{12}		
		K_{13}		
	K_{2u}	K_{21}	21.94	
		K_{22}		55.79
		K_{23}		
	K_{3u}	K_{31}	100.67	
		K_{32}		
		K_{33}		
1,1,2	$\boldsymbol{K_{1u}}$	$\boldsymbol{K_{11}}$	**50**	
		$\boldsymbol{K_{12}}$		
		$\boldsymbol{K_{13}}$		
	$\boldsymbol{K_{2u}}$	$\boldsymbol{K_{21}}$	**27.86**	
		$\boldsymbol{K_{22}}$		**55.23***
		$\boldsymbol{K_{23}}$		
	$\boldsymbol{K_{3u}}$	$\boldsymbol{K_{31}}$	**82.34**	
		$\boldsymbol{K_{32}}$	**100.67**	
		$\boldsymbol{K_{33}}$		
1,2,2	K_{1u}	K_{11}	50	
		K_{12}		
		K_{13}		
	K_{2u}	K_{21}	24.90	
		K_{22}	27.86	56.52
		K_{23}		
	K_{3u}	K_{31}	82.34	
		K_{32}	100.67	
		K_{33}		
		⋮		
3,3,3	K_{1u}	K_{11}	50	
		K_{12}	61	
		K_{13}	72	
	K_{2u}	K_{21}	21.94	
		K_{22}	24.90	58.99
		K_{23}	27.86	
	K_{3u}	K_{31}	82.34	
		K_{32}	91.5	
		K_{33}	100.67	

Based on the table 4, the optimal knot points obtained are 50 for X_1, 27.86 for X_2, then 82.34, and 100.67 for X_3 is a model with an minimum AIC value of 55.23. The truncated spline nonparametric regression model based on the optimal knot points is given as follows:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(h)}{1 + \exp(h)}$$

Where

$$h = \beta_0 + \sum_{j=1}^{3} \beta_{j1} x_{ji} + \beta_{12}(x_{1i} - 50)_+ + \beta_{22}(x_{2i} - 27.86)_+ + \beta_{32}(x_{3i} - 82.34)_+ + \beta_{33}(x_{3i} - 100.67)_+$$

With the truncated spline nonparametric regression for binary response method, the model parameter estimation results for the data of status type 2 diabetes mellitus are given in the following table 5.

Table 5. Parameter Estimation Results of Nonparametric Model

Parameter	Estimation
β_0	-0.901
β_{11}	0.305
β_{12}	0.207
β_{21}	0.176
β_{22}	-0.062
β_{31}	-0.231
β_{32}	0.290
β_{33}	-0.134

Based on the estimation results in Table 5, the model is given as follows:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(h)}{1 + \exp(h)} \tag{20}$$

Where

$$h = -0.901 + 0.305x_{1i} + 0.207(x_{1i} - 50)_+ + 0.176x_{2i} - 0.062(x_{2i} - 27.86)_+ - 0.231x_{3i_+} + 0.290(x_{3i} - 82.34)_+ - 0.134(x_{3i} - 100.67)_+$$

The model in equation 20 shows how the predictor variables affect the probability of type 2 diabetes mellitus. The coefficient of 0.305 indicates that before the age of 50 years, every 1 year increase in age increases diabetes by 0.305. After the age of 50 years, there is a change in the pattern of influence with a coefficient of 0.207. In other words, after passing the age of 50 years,

the effect of age on diabetes risk becomes stronger. Furthermore, before the body mass index of 27.86 has a coefficient value of 0.176, meaning that every 1 unit increase in body mass index increases diabetes by 0.176. After a body mass index of 27.86, the effect is reduced by 0.062, which means that an increase in body mass index after this point has less influence on diabetes risk. This illustrates that the relationship between body mass index and diabetes risk is not always linear, and there is a point where the effect of body mass index on diabetes risk becomes smaller. The coefficient -0.231 indicates that before the knot, an increase in waist circumference reduces diabetes. After a waist circumference of 82.34, the effect changes with a coefficient of 0.290, which means that an increase in waist circumference after this point increases the risk of diabetes. After waist circumference 100.67, the effect decreases again by 0.134.

3.3. Estimation Result of Semiparemetric Truncated Spline Regression for Binary Response on Type 2 Diabates Mellitus Data

This section discusses modeling type 2 diabetes mellitus data using a semiparametric truncated spline model for binary response data. In this semiparametric approach, the age variable (Denoted as V_1 to differentiate it from the symbols representing nonparametric variables.) is treated as a parametric variable and other variable are treated as nonparametric variables. Meanwhile, the number of knot points is limited to 3. The list of candidate knot points used is presented in Table 6 to ensure the model reaches the optimal level.

Table 6. AIC Value based on Knot Point Candidate of Semiparametric Model

Number of Knot Points	K_{ju}		Knot Point Value	AIC
1,1	K_{2u}	K_{21}	18.24	58.44
		K_{22}		
		K_{23}		
	K_{3u}	K_{31}	112.12	
		K_{32}		
		K_{33}		
1,2	K_{2u}	K_{21}	29.34	57.80
		K_{22}		
		K_{23}		
	K_{3u}	K_{31}	84.62	
		K_{32}	11.12	
		K_{33}		

Number of Knot Points	K_{ju}		**Knot Point Value**	**AIC**
2,1	K_{2u}	K_{21}	**20.46**	**56.46***
		K_{22}	**29.34**	
		K_{23}		
	K_{3u}	K_{31}	**112.12**	
		K_{32}		
		K_{33}		
		⋮		
3,3	K_{2u}	K_{21}	18.24	58.59
		K_{22}	24.90	
		K_{23}	29.34	
	K_{3u}	K_{31}	84.62	
		K_{32}	98.37	
		K_{33}	112.12	

Based on the table 6, the optimal knot points obtained are 20.46, and 29.34 for X_2, and 112.12 for X_3 is a model with an minimum AIC value of 56.46. The truncated spline semiparametric regression model based on the optimal knot points is given as follows:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(h)}{1+\exp(h)}$$

Where

$$h = \delta_0 + \delta_1 v_{1i} + \sum_{j=2}^{3} \beta_{j1} x_{ji} + \beta_{22}(x_{2i} - 20.46)_+ + \beta_{23}(x_{2i} - 29.34)_+ + \beta_{32}(x_{3i} - 112.12)_+$$

With the Truncated Spline semiparametric regression for binary response method, the model parameter estimation results for the data of status type 2 diabetes mellitus are given in the following table 7.

Table 7. Parameter Estimation Results of Semiparametric Model

Parameter	**Estimation**
δ_0	-42.793
δ_1	0.138
β_{21}	1.705
β_{22}	-1.838
β_{23}	23.525
β_{31}	0.015
β_{32}	-9.172

Based on the estimation results in Table 7, the model is given as follows:

$$\pi(\boldsymbol{x_i}) = \frac{\exp(h)}{1+\exp(h)} \tag{21}$$

Where

$$h = -42.793 + 0.138v_{1i} + 1.705x_{2i} - 1.838(x_{2i} - 20.46)_+ + 23.525(x_{2i} - 29.34)_+ + 0.015x_{3i} - 9.172(x_{3i} - 20.46)_+$$

The model in equation 21 shows how the predictor variables affect the probability of type 2 diabetes mellitus. Every year increase in age increases the likelihood of developing diabetes by 0.138, which means that the older one gets, the higher the risk of developing diabetes. The relationship between body mass index and diabetes risk shows a non-linear pattern. Before reaching 20.46, every 1 unit increase in body mass index increases the risk of diabetes by 1.705. However, after crossing this figure, the effect decreases by 1.838, so the impact on diabetes risk becomes smaller. Conversely, after body mass index reaches 29.34, the risk of diabetes increases by 23.525, indicating that individuals with a high body mass index have a much greater risk of diabetes. For waist circumference, before reaching 20.46, the increase has almost no significant impact on diabetes risk. However, after crossing this mark, the effect changes drastically to negative with 9.172, meaning that an increase in waist circumference after this point actually decreases the risk of diabetes significantly. This suggests that the relationship between waist circumference and diabetes risk is not always unidirectional, but rather depends on a certain tipping point.

3.4. Comparison of Classification Criteria Evaluation in Nonparametric and Semiparametric Models

The more optimal regression model can be determined based on the smallest deviance value. Based on the deviance statistical test results, the following results are obtained:

Table 8. Comparison of Deviance Values

Model	Deviance Values
Truncated Spline Nonparametric Regression for Binary Response	52.94
Truncated Spline Semiparametric Regression for Binary Response	42.46

Based on Table 8, the deviance value for truncated spline semiparametric regression model (42.46) is smaller than the deviance value for truncated spline nonparametric regression model (52.94). So that the truncated spline semiparametric regression for binary response model is a

better model for status type 2 diabetes mellitus.

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. The confusion matrix can be used to measure classification accuracy for binary response data.

Table 9. Confusion Matrix

Truncated Spline Nonparametric Regression for Binary Response		**Prediction**	
		Non Diabetes	**Diabetes**
Actual	**Non Diabetes**	10	11
	Diabetes	3	36
Truncated Spline Semiparametric Regression for Binary Response		**Prediction**	
		Non Diabetes	**1**
Actual	**Non Diabetes**	14	7
	Diabetes	1	38

Table 9 shows that there are 10 individu classified as a non diabetes and 36 individu classified diabates that are correctly predicted by the truncated spline nonparametric regression model for binary response data. For the truncated spline semiparametric regression model, the table shows that there are 14 individu classified as non diabetes and 38 individu classified as diabetes that are correctly by the model.

In evaluating the performance of regression models for binary response data, evaluation criteria are used to determine how well the model can classify the data. Table 10 below presents the comparison between the two regression model.

Table 10. Confusion Matrix

Criteria	**Model**	
	Truncated Spline Nonparametric Regression for Binary Response	**Truncated Spline Semiparametric Regression for Binary Response**
Accuracy	76.67%	86.67%
Sensitivity	92.31%	97.44%
Specificity	47.62%	66.67%
Precision	76.60%	84.44%

The comparative analysis of the two regression models indicates that the semiparametric approach outperforms the nonparametric approach in classification accuracy. The semiparametric model achieves a higher accuracy of 86.67% compared to 76.67% for the nonparametric model, demonstrating its superior ability to correctly classify individuals. Moreover, the semiparametric model exhibits greater sensitivity (97.44%) than the nonparametric model (92.31%), indicating semiparametric model is better at detecting individuals who actually have diabetes. In terms of specificity, the semiparametric model has a higher value (66.67%) than the nonparametric model (47.62%), indicating that this model is better at identifying individuals who do not have diabetes. Additionally, the semiparametric model demonstrates higher precision (84.44%) compared to the nonparametric model (76.60%), signifying that its positive predictions are more reliable with a lower margin of error. Overall, the semiparametric model proves to be superior across all evaluation metrics, including accuracy, sensitivity, specificity, and precision. This finding suggests that incorporating parametric components in truncated spline regression allows for a more flexible modeling of relationships across different sub-intervals, leading to improved classification performance.

4. CONCLUSION

Based on the estimation results of the semiparametric truncated spline regression model for binary response data, it can be concluded that this model provides better performance than the nonparametric model in predicting type 2 diabetes mellitus status. This conclusion is supported by the lower deviance value in the semiparametric model and its superior evaluation criteria. The analysis results indicate that age has a positive effect on the risk of type 2 diabetes mellitus. That is, the older a person gets, the greater the probability of that individual developing diabetes. This finding aligns with existing literature, which highlights age as a major determinant in diabetes risk, due to metabolic changes and increased insulin resistance in older individuals Additionally, body mass index and waist circumference significantly contribute to the model. Using the truncated spline approach, knots were identified in these variables, indicating a nonlinear relationship with diabetes risk. For example, an increase in body mass index may have a relatively minor effect on diabetes risk up to a certain threshold. A similar pattern is observed for waist circumference, where exceeding a specific threshold is associated with a significantly higher likelihood of developing type 2 diabetes mellitus.

The strength of the semiparametric approach with a truncated spline in this model lies in its

flexibility to capture complex relationships between predictor variables and response variable (binary). It accommodates patterns that are partially linear and patterns that are partially invariant within certain sub-intervals, such as body mass index and waist circumference. The model is expected to provide a more comprehensive understanding of diabetes risk factors, serving as a valuable basis for decision-making in the health sector, especially in the prevention and treatment of type 2 diabetes mellitus.

ACKNOLEDGMENTS

The authors would like to express their gratitude to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding this research under the "Master Program of Education Leading to a Doctoral Degree for Excellent Graduates (PMDSU) Batch VII" through Grant Number 038/E5/PG.02.00.PL/2024 and local Grant Number 1799/PKS/ITS/2024.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

[1] A.B. Olokoba, O.A. Obateru, L.B. Olokoba, Type 2 Diabetes Mellitus: A Review of Current Trends, Oman Med. J. 27 (2012), 269–273. https://doi.org/10.5001/omj.2012.68.

[2] International Diabetes Federation, IDF Diabetes Atlas, https://diabetesatlas.org.

[3] S. Klein, A. Gastaldelli, H. Yki-Järvinen, P.E. Scherer, Why Does Obesity Cause Diabetes?, Cell Metab. 34 (2022), 11–20. https://doi.org/10.1016/j.cmet.2021.12.012.

[4] T. Kelly, W. Yang, C.S. Chen, et al. Global Burden of Obesity in 2005 and Projections to 2030, Int. J. Obes. 32 (2008), 1431–1437. https://doi.org/10.1038/ijo.2008.102.

[5] N. Razavian, S. Blecker, A.M. Schmidt, et al. Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors, Big Data 3 (2015), 277–287. https://doi.org/10.1089/big.2015.0020.

[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Front. Genet. 9 (2018), 515. https://doi.org/10.3389/fgene.2018.00515.

[7] R.D. Joshi, C.K. Dhakal, Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches, Int. J. Environ. Res. Public Health 18 (2021), 7346. https://doi.org/10.3390/ijerph18147346.

[8] L.R. Eubank, Nonparametric Regression and Spline Smoothing, CRS Press, Texas, 1999.

[9] 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.

[10] A.A. Puspitasari, A.A.R. Fernandes, A. Efendi, et al. Development of Nonparametric Truncated Spline at Various Levels of Autocorrelation of Longitudinal Generating Data, J. Stat. Appl. Probab. 12 (2023), 757–766. https://doi.org/10.18576/jsap/120234.

[11] Y. Zhang, L. Hua, J. Huang, A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data: Cox Model with Interval-Censored Data, Scandinavian J. Stat. 37 (2010), 338–354. https://doi.org/10.1111/j.1467-9469.2009.00680.x.

[12] D.D. Prawanti, I.N. Budiantara, J.D.T. Purnomo, Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data, IOP Conf. Ser.: Mater. Sci. Eng. 546 (2019), 052053. https://doi.org/10.1088/1757-899X/546/5/052053.

[13] M. Setyawati, N. Chamidah, A. Kurniawan, Confidence Interval of Parameters in Multiresponse Multipredictor Semiparametric Regression Model for Longitudinal Data Based on Truncated Spline Estimator, Commun. Math. Biol. Neurosci. 2022 (2022), 107. https://doi.org/10.28919/cmbn/7672.

[14] A.S. Suriaslan, I.N. Budiantara, V. Ratnasari, Nonparametric Regression Estimation Using Multivariable Truncated Splines for Binary Response Data, MethodsX 14 (2025), 103084. https://doi.org/10.1016/j.mex.2024.103084.

[15] I.N. Budiantara, Regresi Nonparameterik Spline Truncated, ITS Press, Surabaya, 2019.

[16] V. Ratnasari, Purhadi, I. Calveria et al. Parameter Estimation and Hypothesis Testing the Second Order of Bivariate Binary Logistic Regression (S-BBLR) Model with Berndt Hall-Hall-Hausman (BHHH) Iterations, Commun. Math. Biol. Neurosci. 2022 (2022), 35. https://doi.org/10.28919/cmbn/7258.

[17] Sifriyani, I.N. Budiantara, S.H. Kartiko, et al. A New Method of Hypothesis Test for Truncated Spline Nonparametric Regression Influenced by Spatial Heterogeneity and Application, Abstr. Appl. Anal. 2018 (2018), 9769150. https://doi.org/10.1155/2018/9769150.

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