Int. J. Advance Soft Compu. Appl, Vol. 10, No. 3, November 2018
ISSN 2074-8523

The Logistic Regression Analysis with Nonparametric Approach based on Local Scoring Algorithm (Case Study: Diabetes Mellitus Type II Cases in Surabaya of Indonesia)

Marisa Rifada[1], Suliyanto[1], Eko Tjahjono[1] and Ayundyah Kesumawati[2]
[1]Department of Mathematics, Universitas Airlangga, Indonesia
e-mail: marisa.rifada@fst.unair.ac.id, suliyanto@fst.unair.ac.id , eko-t@fst.unair.ac.id
[2]Department of Statistics, Islamic University of Indonesia
e-mail: ayundyah.k@uii.ac.id

Abstract

One of the statistical methods used to analyze the relationship between categorical scale response variable and categorical or continuous variable of predictors is logistic regression analysis. There are two ways of approaching regression model. The global approach assumes that the regression model for each individual observation has the same parameters, whereas the local approach assumes not all individuals have the same parameters. Regression model with local approach is often called nonparametric regression. The purpose of this research will be to develop the risk model of Diabetes Mellitus Type II incidence in patients of The Hajj General Hospital of Surabaya with nonparametric logistic regression approach by the local scoring algorithm. Based on the result of this research concluded that the probability of a person's risk of Diabetes Mellitus Type II increases with age up to about 60 years and after that tends to decrease, the more the increase in one's BMI there is a significant increase in the prevalence of Diabetes Mellitus Type II. The validity of the estimation of this obtained model is 88.89%.

Keywords: *Diabetes Mellitus Type II, Local Scoring Algorithm, Logistic regression.*

1 Introduction

The model of association between the risk of Diabetes Mellitus Type II incidence in patient and the factors influencing it would be more useful if it could be formulated mathematically, i.e to see how much the factors significantly influence the odds of the patient being exposed to Diabetes Mellitus Type II. One of the statistical methods used to analyze the relationship between categorical scale response variable and categorical or continuous variable of predictors is logistic regression analysis. Binary logistic regression is a logistic regression with dichotomous response variable consisting of two categories [1].

There are differences assumption in regression model that divided by global approach or local approach. If each individual observation assumed has the same parameter so it used the global approach. If the parameter assumes at least there are one individual do not have same parameter it used local approach. Because in general each individual (patient) has varying physical conditions, this allows the characteristics of each patient to be different so that a more suitable regression approach to be applied is the local approach. Regression model with local approach is often called nonparametric regression.

The purpose of this research will be to develop the risk model of Diabetes Mellitus Type II incidence in patients of The Hajj General Hospital of Surabaya with nonparametric logistic regression approach based on Kernel estimator using Generalized Additive Models (GAM) method. To get an estimation of logistic regression model in nonparametric additive used Local scoring algorithm. Local scoring algorithm is best suited for nonparametric additive regression models whose distribution of response variables is included in exponential family members, one of which is the Bernoulli distribution [2].

2 Problem Formulations or Methodology

The logistic regression model with nonparametric approach assumes that the response variable is binary categorical (0,1), whereas the predictor variables are continuous with the logit function is expressed as the sum of nonparametric regression functions $f_j(x_{ji})$ of the *j-th* predictor variable to *i-th* observation with unknown function form and known link function G. If the selected logit link function G, then the logistic regression model with nonparametric approach can be obtained below:

$$\ln\left[\frac{E\left(Y \mid X = x_i\right)}{1 - E\left(Y \mid X = x_i\right)}\right] = \sum_{j=1}^{p} f_j\left(x_{ji}\right) \quad (1)$$

for $E\left(Y \mid X = x_i\right) = P\left(Y = 1 \mid X = x_i\right) = \rho_i$ is the probability of success in *i-th* observation corresponding to the predictor variable X.

The regression function of $f_j(x_{ji})$ is unknown form and it will be estimated by Kernel estimator approach. According to Demir and Toktamis [3], Kernel estimator for the single predictor variable is as follows:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)} \tag{2}$$

with h is the optimal bandwidth and K is a Kernel function. According to Hardle [4], the Kernel function with bandwidth h is defined as :

$$K_h\left(x\right) = \frac{1}{h} K\left(\frac{x}{h}\right), \text{ for } -\infty < x < \infty, h > 0 \tag{3}$$

Since the predictor variable *j-th* has the bandwidth h_j, then from equation (2) is obtained Kernel estimator for n observations with predictor variable *j-th* to the observations *i-th* as follow :

$$\hat{f}_j(x_{ji}) = \frac{\sum_{k=1}^{n} K_{h_j}(x_{ji} - x_{jk}) y_k}{\sum_{k=1}^{n} K_{h_j}(x_{ji} - x_{jk})}, \quad i = 1, 2, \ldots, n \tag{4}$$

Equation (4) can be expressed as :

$$\hat{f}_j(x_{ji}) = \frac{L' W_{ji}(h_j) Y}{L' W_{ji}(h_j) L} \tag{5}$$

where

$$W_{ji}(h_j) = diag\left[K_{h_j}(x_{ji} - x_{j1}), K_{h_j}(x_{ji} - x_{j2}), \ldots, K_{h_j}(x_{ji} - x_{jn})\right] \tag{6}$$

$Y = (y_1, y_2, \ldots, y_n)'$ and $L = (1, 1, \ldots, 1)'$

Equation (5) can be expressed as follows:

$$\hat{f}_j(x_j) = A(h_j) Y \tag{7}$$

where

$$\hat{f}_j(x_j) = \begin{bmatrix} \hat{f}_j(x_{j1}) \\ \hat{f}_j(x_{j2}) \\ \vdots \\ \hat{f}_j(x_{jn}) \end{bmatrix} \quad ; \quad A(h_j) = \begin{bmatrix} \frac{L' W_{j1}(h_j)}{L' W_{j1}(h_j) L} \\ \frac{L' W_{j2}(h_j)}{L' W_{j2}(h_j) L} \\ \vdots \\ \frac{L' W_{jn}(h_j)}{L' W_{jn}(h_j) L} \end{bmatrix} \tag{8}$$

In this study, the binary nonparametric regression model has the response variable which binary categorical (0,1) and distributed Bernoulli. We estimate the binary nonparametric regression model based on Kernel estimator used the Local scoring algorithm because this algorithm very suitable usage in additive nonparametric regression model with the distribution of the response variable was included in the exponential family members, one of which is a Bernoulli distribution. This algorithm consists of two loops that are Scoring step (outer loop) is iterated until the average value of deviance convergent and weighted Backfitting step (inner loop) is iterated until the average value of the Residual Sum of Squares (RSS) convergent. Scoring step is done iteratively by determining the adjusted value of the response variable (z) which is formulated as follows :

$$z_i = m_i^{(s)} + \left(y_i - \pi_i^{(s)}\right)\left(\frac{\partial m_i}{\partial \pi_i}\right)_{(s)} \text{ , i = 1, 2, ..., n} \tag{9}$$

$$\text{where } m_i^{(s)} = \sum_{j=1}^{p} f_j^{(s)}(x_{ji}) = \ln\left(\frac{p_i^{(s)}}{1 - p_i^{(s)}}\right) \text{, s = 0,1,2, ...} \tag{10}$$

The weighting matrix B is a diagonal matrix with the main diagonal elements are :

$$b_i = \left(\frac{\partial \pi_i}{\partial m_i}\right)^2_{(s)} \left(V_i^{(s)}\right)^{-1} \text{ , i = 1, 2, ..., n} \tag{11}$$

Since the response variable Y has Bernoulli distribution, then we have :

$$V_i = \text{Var}(Y_i) = \pi_i(1-\pi_i) \tag{12}$$

Based on equation (9) we get :

$$\frac{\partial m_i}{\partial \pi_i} = \frac{1}{\pi_i} + \frac{1}{1-\pi_i} = \frac{1}{\pi_i(1-\pi_i)} \tag{13}$$

So, from equation (12) we can obtain :

$$\frac{\partial \pi_i}{\partial m_i} = \pi_i(1-\pi_i) \tag{14}$$

Then, by substituting equation (12) into equation (8) we get :

$$z_i = m_i^{(s)} + \frac{\left(y_i - \pi_i^{(s)}\right)}{\pi_i^{(s)}\left(1-\pi_i^{(s)}\right)} \tag{15}$$

Next, by substituting equations (11) and (13) into equation (10) we obtain :

$$b_i = \pi_i(1-\pi_i) \tag{16}$$

Subsequently to step weighted Backfitting is done iteratively by determining the estimation of nonparametric regression functions $\hat{f}_j(x_j)$ in the model which is formulated as follows:

$$\hat{f}_j(x_j) = A(h_j)\left\{ z - \sum_{k=1}^{j-1} \hat{f}_k^{(s+1)}(x_k) - \sum_{k=j+1}^{p} \hat{f}_k^{(s)}(x_k) \right\} \quad \text{, j = 1, 2, ..., p} \tag{17}$$

with s = 0,1,2, ...

Finally, we obtain estimation of binary nonparametric regression model based on the Kernel estimator using Local scoring algorithm as follows:

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \sum_{j=1}^{p} \hat{f}_j(x_{ji})$$

$$= \sum_{j=1}^{p} \frac{L'W_{ji}(h_j)}{L'W_{ji}(h_j)L} \left\{ z - \sum_{k=1}^{j-1} \hat{f}_k^{(s+1)}(x_k) - \sum_{k=j+1}^{p} \hat{f}_k^{(s)}(x_k) \right\} \tag{18}$$

3 Algorithm

To estimate the logistic regression model with nonparametric approach, we used the following algorithm :

A. The algorithm for determining the optimal bandwidth value for each predictor variable.

1. Input pair data $(y_i, x_{1i}, x_{2i}, ..., x_{pi})$; $i = 1, 2, ..., n$
2. Defining the Kernel function used is Gaussian Kernel function, i.e :
3. Determining the initial bandwidth value (h_j)
4. Determining a diagonal weighting matrix $W_{ji}(h_j)$
5. Determining matrix $A(h_j)$
6. Calculating the value of $\hat{f}_j\left(x_j\right) = A(h_j)Y$
7. Calculating the value of $MSE(h_j) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{f}_j(x_{ji}))^2$ (19)

8. Calculating the value of $GCV(h_j) = \dfrac{MSE(h_j)}{(n^{-1}tr[I - A(h_j)])^2}$ (20)
9. Repeat steps 4 to 8 until we get minimum value of $GCV(h_j)$ for optimal bandwidth (h_j).

B. The algorithm for initial estimation of nonparametric regression function $\hat{f}_j(x_j)$ for each predictor variable.

1. Input pair data $(y_i, x_{1i}, x_{2i}, ..., x_{pi})$; $i = 1,2,...,n$
2. Defining the kernel function used is Gaussian kernel function
3. Input the optimal bandwidth value (h_j) obtained from the algorithm (A).
4. Determining a diagonal weighting matrix $W_{ji}(h_j)$
5. Determining matrix $A(h_j)$
6. Calculating the value of $\hat{f}_j(x_j) = A(h_j)Y$

C. The Local scoring algorithm to estimate the binary nonparametric regression model.

1. Input pair data $(y_i, x_{1i}, x_{2i}, ..., x_{pi})$; $i = 1,2,...,n$
2. Input the initial estimation value of $\hat{f}_j^{(0)}(x_j)$; $j = 1,2,..., p$ obtained from the algorithm (B)
3. Iterating Scoring step (outer loop) as follows:
   a. Determining the adjusted value of the response variable (z) with elements of row i-th is

$$z_i = m_i^{(s)} + \frac{\left(y_i - \pi_i^{(s)}\right)}{\pi_i^{(s)}\left(1-\pi_i^{(s)}\right)} \quad , \; i = 1,2,...,n \tag{21}$$

   with $m_i^{(s)} = \sum_{j=1}^{p} \hat{f}_j^{(s)}(x_{ji})$, $\pi_i^{(s)} = \dfrac{\exp(m_i^{(s)})}{1+\exp(m_i^{(s)})}$ and determining the weighting matrix (B) in the form of a diagonal matrix with the main diagonal elements are

$$b_i = \pi_i^{(s)}\left(1-\pi_i^{(s)}\right) \quad , \; i = 1,2, \; ... \; , n \tag{22}$$

   b. Iterating weighted Backfitting step (inner loop) as follows:
      (i) For the initial iteration ($s = 0$), Defining $\hat{m}_j^{(s)}(X_j) = \hat{m}_j^{(r)}(X_j)$ and $z = (z_1, z_2, \text{K}, z_n)^T$
      (ii) Estimating the nonparametric regression functions in the model for $j = 1,2, \; ... \; , p$, that is

$$\hat{f}_j^{(s+1)}(X_j) = A(h_j)\left\{ z - \sum_{k=1}^{j-1} \hat{f}_k^{(s+1)}(X_k) - \sum_{k=j+1}^{p} \hat{f}_k^{(s)}(X_k) \right\} \quad (23)$$

(iii) Calculating the average value of the weighted Residual Sum of Squares (RSS)

$$Avg(RSS)^{(s+1)} = \frac{1}{n}\left\{ \left(z^{(s+1)} - m^{(s+1)} \right)^T B^{(s+1)} \left(z^{(s+1)} - m^{(s+1)} \right) \right\} \quad (24)$$

(iv) Repeat steps (ii) to (iii) for $s = s+1$ until the average value of RSS convergent, i.e

$$abs\left(Avg(RSS)^{(s+1)} - Avg(RSS)^{(s)} \right) < \varepsilon \text{, for } \varepsilon = 0.0001 \quad (25)$$

c. Defining $\hat{f}_j^{(r+1)}(x_j) = \hat{f}_j^{(s+1)}(x_j)$ and then determining the average value of deviance, i.e

$$Avg\left(D(y_i;\pi_i) \right)^{(r+1)} \approx \frac{-2}{n} \sum_{i=1}^{n} \left\{ y_i \ln \pi_i^{(r+1)} + (1-y_i) \ln(1-\pi_i^{(r+1)}) \right\} \quad (26)$$

d. Repeat step a) to step c) for $r = r+1$ until the average value of deviance convergent, i.e

$$abs\left(Avg\left(D(y_i;\pi_i) \right)^{(r+1)} - Avg\left(D(y_i;\pi_i) \right)^{(r)} \right) < \varepsilon \text{ , with } \varepsilon = 0.0001 \quad (27)$$

4 Data

The data used in this research is the incidence of Diabetes Mellitus Type II disease. The data were obtained from the sample of patients in Poly Disease in The Hajj General Hospital Surabaya at 2016 with the total sample of 81 people, consisting of 31 people classified as Diabetes Mellitus Type II patients and 50 people with Diabetes Mellitus Type II patients. The response variable of the data is the condition of the patient that diagnosed by the doctor suffering from Diabetes Mellitus Type II disease (Y = 1) or not suffering from Diabetes Mellitus Type II (Y = 0). Whereas the predictor variables (X) were used to determine the risk of Diabetes Mellitus Type II, i.e Age (X_1), BMI (Body Mass Index) (X_2), Waist Circumference (X_3), and Systolic Blood Pressure (X_4).

5 Results, Analysis and Discussions

The first step in logistic regression modeling with nonparamteric approach is to determine the optimal parameter of the term (bandwidth). The bandwidth selection is very important in obtaining regression function estimators based on nonparametric approaches. Bandwidth is a balance control between the functionality of the data. If bandwidth is very small then the estimated function obtained will be very rough and go to the data whereas if bandwidth is very large then the estimated function will be very smooth and towards the average of the

response variable. Therefore, in choosing bandwidth is expected optimal value. Using the GCV method, the optimal bandwidth for each predictor variable is presented in the following Table 1:

Table 1 : The optimal bandwidth for each predictor variable

Variable	Optimal bandwitdh	GCV minimum
X_1	6.09	0.219563341957588
X_2	5.85	0.236885365100267
X_3	2.06	0.201452655133131
X_4	2.77	0.1550234791469

Furthermore, the optimal bandwidth is used to determine the initial value of nonparametric regression function $\hat{f}_j(X_j)$ in each predictor variable. After obtaining the optimum $\hat{f}_j(X_j)$ for each predictor variable, the next step is to iterate using Local scoring algorithm to get the estimation result of logistic regression model with nonparametric approach. The plot of probability estimate of a patient at risk for Diabetes Mellitus Type II disease based on each predictor variable is shown in Figure 1 until Figure 4 as follows:

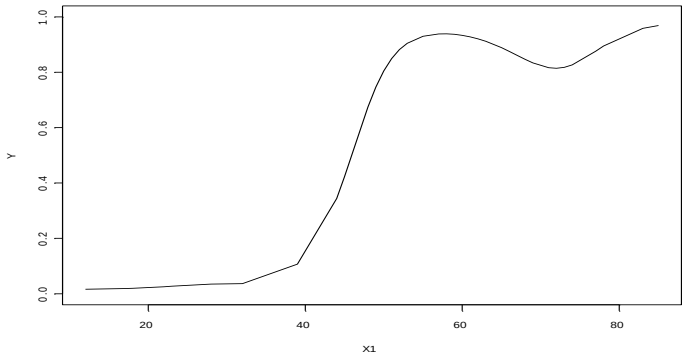

Fig. 1 Plot of probability estimate at risk for Diabetes Mellitus Type II disease based on Age (X_1)

Based on Fig.1 it can be seen that the probability of a person's risk of suffering Diabetes Mellitus Type II increases with age up to about 60 years, after which it tends to decrease. This is in line with the results of a study by M. Sue Kirkman [5] which states that the incidence of diabetes increases with age until around age 65, after which both incidence and prevalence seem to decrease. Diabetes is often found in older adults because at that age the body functions physiologically decreased and decreased secretion or insulin resistance so that the body's ability to control high blood glucose is less than optimal.

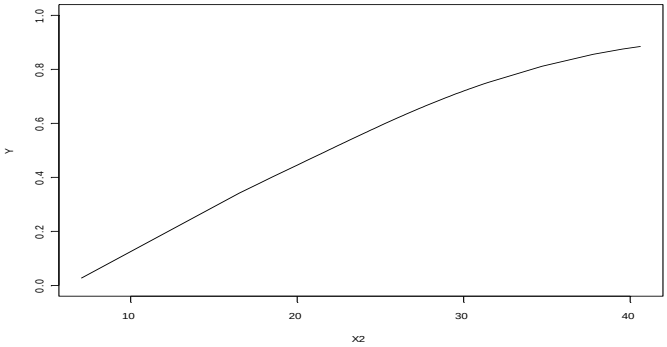

Fig. 2 Plot of probability estimate at risk for Diabetes Mellitus Type II disease based on Body Mass Index (X_2)

Fig. 2 shows that the association between BMI and the prevalence of Diabetes Mellitus Type II forms a linear trend that indicates that an increase in a person's BMI there is a significant increase in the risk probability of developing Diabetes Mellitus Type II. This is in line with a study by Michael L Ganz [6] which concluded that the change in the magnitude of the Odds Ratios for risk Diabetes (ORs) from one BMI category to the next was larger for individuals in higher BMI categories than individuals in lower BMI categories, as illustrated by the increasing slope of the lines connecting the ORs.

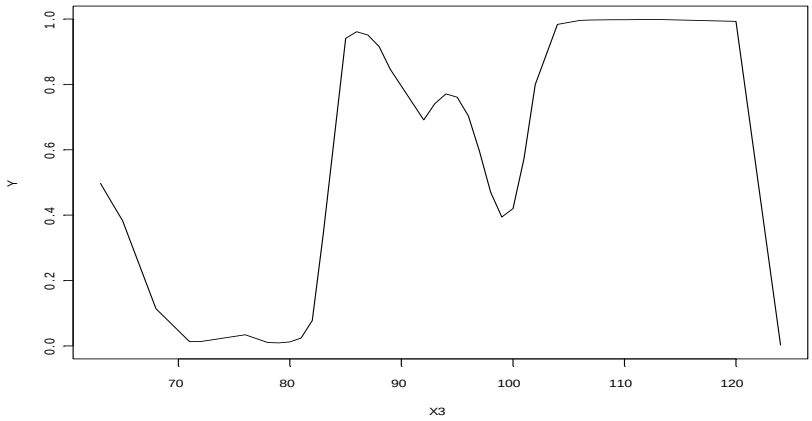

Fig. 3 Plot of probability estimate at risk for Diabetes Mellitus Type II disease based on Waist Circumference (X_3)

The association between waist circumference and the probability of a person's risk of suffering Diabetes Mellitus Type II is shown by Fig.3. The result forms fluctuate trend. This is because in this research we not distinguish by gender. The normal waist circumference limits for both men and women are different. The values of 102 cm for men and 88 cm for women, recommended as cutoff points by National Heart Lung and Blood Institute (NHLBI) of waist circumference for healthy people [7]. This is in line with a study by Dagan [8] which concluded that differences between the gender were statistically significant for waist

circumference. So, in the future research we emphasize the need to investigate men and women separately.

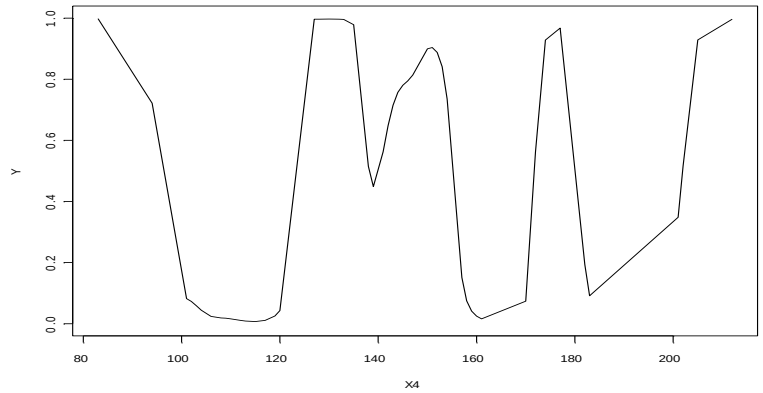

Fig. 4 : Plot of probability estimate at risk for Diabetes Mellitus Type II disease based on Systolic Blood Pressure (X_4)

Fig. 4 shows that the association between Systolic Blood Pressure and the prevalence of Diabetes Mellitus Type II forms fluctuate trend. This is because blood pressure at the time of measurement depends on the patient's condition. No matter which device is used to measure blood pressure, it must be recognised that blood pressure is a variable haemodynamic phenomenon, which is influenced by many factors, not least being the circumstances of measurement itself. These influences on blood pressure can be significant, often accounting for rises in systolic blood pressure greater than 20 mm Hg, and if they are ignored, or unrecognised, hypertension will be diagnosed erroneously and inappropriate management instituted. These factors have to be carefully considered in all circumstances of blood pressure measurement—self measurement by patients, conventional measurement, measurement with automated devices whether in a doctor's surgery, an ambulance, a pharmacy, or in hospital using sophisticated technology [9]. Furthermore, after obtaining the estimation result of logistic regression model with nonparametric approach, we calculate the accuracy of the classification in this research as follows:

Table 2: The accuracy of the classification for logistic regression model with nonparametric approach

		Prediction Y		True Percentage
		0	1	
Observation Y	0	27	5	84.4
	1	4	45	91.8
	Total	31	50	88.89

Based on Table 2, it is found that the patients classified into categories not sufferers of Diabetes Mellitus Type II correctly are as many as 27 persons. While the patients classified into categories of patients with Diabetes Mellitus Type II correctly are as many as 45 persons. The rest of 5 patients who are not sufferers of Diabetes Mellitus Type II are classified into categories of patients with Diabetes Mellitus Type II and 4 patients diagnosed with Diabetes Mellitus Type II are classified into non- Diabetes Mellitus Type II category. The accuracy classification of this method as follow:

$$Accuracy = \frac{\sum number\ of\ correct\ prediction}{total\ \ prediction} = \frac{27 + 45}{81} = 0{,}889$$

So, the estimation of logistic regression model with nonparametric approach in the incidence of Diabetes Mellitus Type II disease has a validity of 88.89%. it can be conclude that the method was able to predict correctly 88.9%. This suggests that the model obtained is valid enough to calculate the probability of a person suffering from Diabetes Mellitus Type II.

6 Conclusion

Based on the result of applying logistic regression model with nonparametric approach to the incidence of Diabetes Mellitus Type II disease at The Hajj General Hospital of Surabaya in 2016, it is concluded that the probability of a person's risk of Diabetes Mellitus Type II increases with age up to about 60 years and after that tends to decrease, the more the increase in one's BMI there is a significant increase in the prevalence of Diabetes Mellitus Type II. The estimation of this obtained model is valid enough to calculate the probability of suffering Diabetes Mellitus Type II with the validity of 88.89%.

ACKNOWLEDGEMENTS
The authors thanks to the Ministry of Research, Technology and Higher Education of the Republic of Indonesia for financial support of this research through The Fundamental Research University Grant 2018. The authors also thank to the anonymous referees for their valuable suggestions which let to the improvement on the manuscript.

References

[1] Agresti, A. (2002). Categorical Data Analysis. *Second Edition*. John Wiley and Sons. New York

[2] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall. London

[3] Demir, S. and Toktamis O. (2010). On The Adaptive Nadaraya-Watson Kernel Regression Estimators. *Hacettepe Journal of Mathematics and Statistics*, 39(3), 429-437

[4] Hardle, W. (1990). *Applied Nonparametrik Regression*. Cambridge University Press.New York

[5] M. Sue Kirkman, et al. (2012). Diabetes in Older Adults. *Diabetes Care*, 35(12), 2650-2664

[6] Michael L Ganz, Neil Wintfeld, Qian Li, Veronica Alas, Jakob Langer and Mette Hammer. (2014). The association of body mass index with the risk of type 2 diabetes: a case control study nested in an electronic health records system in the United States. *Diabetology & Metabolic Syndrome*, 6 :50

[7] Flegal, K, M. (2007). Waist circumference of healthy men and women in the United States. *International Journal of Obesity*, 31, 1134-1139

[8] Dagan, S.S, Segev, S., Novikov, I., and Dankner, R. (2013). Waist circumference vs body mass index in association with cardiorespiratory fitness in healthy men and women: a cross sectional analysis of 403 subjects. *Nutrition Journal*, 12 :12

[9] Beevers, G., Lip G, Y, H., and O'Brien, E. (2001). Blood pressure measurement. *BMJ*, 322 (7292), 981-985