Classification using nonparametric logistic regression for predicting working status

Cite as: AIP Conference Proceedings **2329**, 060032 (2021); https://doi.org/10.1063/5.0043598
Published Online: 26 February 2021

Wahyu Wibowo, Rahmi Amelia, Fanny Ayu Octavia, and Regina Niken Wilantari

View Online Export Citation

AIP Conference Proceedings **2329**, 060032 (2021); https://doi.org/10.1063/5.0043598 **2329**, 060032

Classification Using Nonparametric Logistic Regression for Predicting Working Status

Wahyu Wibowo[1,a)], Rahmi Amelia[1,2], Fanny Ayu Octavia[1] and Regina Niken Wilantari[3]

[1] *Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2] *Regional Economic Development Institute, Surabaya, Indonesia*
[3] *University of Jember, Jember, Indonesia*

[a)] Corresponding author: wahyu_w@statistika.its.ac.id

Abstract. Logistic regression is classical and prominent method for classification and it is used as benchmark for comparing the alternative methods. However, logistic regression is not always superior compared to the other methods. The accuracy of logistic regression could be improved by incorporating nonparametric model. The response variable used in this study is working status of housewife that categorized as working or not-working. Meanwhile the predictor variables consists of three variables, they are highest education level, age, and household expenditure. The result of fitting model shows that by incorporating nonparametric model to the binary logistic regression model can improve the classification accuracy. This is indicated not only by accuracy percentage, but also by area under Receiving Operating Characteristic (ROC) curve. The dataset will be divided into two parts, 80% as training data and 20% as testing data. The classification accuracy resulted by the binary logistic regression model is 60.36% for training data and 59.30% for testing data. Meanwhile, the classification accuracy of nonparametric logistic model is 63.43% for training data and 64.94%. for testing data. The classification accuracy and area under curve of nonparametric logistic regression is higher than those of binary logistic regression.

INTRODUCTION

One of problems in the machine learning is classification. Classification is one of the tasks in the machine learning, especially deal with the supervised learning and the main purpose is to obtain the fittest model to predict the categorical target feature stand on a set of input feature. Machine learning is vibrant field of research topic and there are such a significant number of techniques and algorithms to create arrangement model from the uncomplicated until the advance model [1]. It motivates some authors to assess performance of many machine learning methods [2].

Logistic regression is one of popular methods for classification and widely use in the practice. It is also classical and prominent method for classification and it is used as benchmark for comparing the alternative methods. This method is used to model the relationship between the categorical target feature and a set of input feature by using the logistics function. This function will predict the probability of one of event of categorical response. Logistic regression will not only predict the probability of event, but also interpretability of relationship model between response and predictor variable. This will lead researchers have more comprehensive understanding connections among variables.

In term of classification accuracy, logistic regression is not always superior compared to the other methods. However, the accuracy could be improved by incorporating nonparametric model for continuous predictor variables. Nonparametric model is one approach of statistical modeling that relaxes functional form assumption between response and predictor variable. Hence, the model will be more flexible to capture any form relationship between response and predictor.This study compares the accuracy of the nonparametric logistic regression model and binary

International Conference on Mathematics, Computational Sciences and Statistics 2020
AIP Conf. Proc. 2329, 060032-1–060032-8; https://doi.org/10.1063/5.0043598
Published by AIP Publishing. 978-0-7354-4073-9/$30.00

logistic model by using empirical data from Indonesian Family Life Survey. The target is defined as working status of housewife either working or not-working. Many cases in the households, the housewives have prominent role to support household economy by producing income through self-working or to be employed by another people. However, labor statistic exhibits that women productivity is lower than men [3].

MATERIALS AND METHODS

Nonparametric Regression is a family of regression models with the infinite number of parameter. Nonparametric regression assumes that the functional form between response and predictor is unknown. Therefore, many approximation functions are required to estimate the unknown function. Some of them are spline, kernel, wavelet, and fourier. The key performance of nonparametric regression is its flexibility to model the response and predictor variables. Flexibility means that the functional form of regression can be defined such that the best model is the most appropriate for the data. For any continuous target feature y and input feature x, the nonparametric regression model specification is

$$y = f(x) + \varepsilon \tag{1}$$

where $f(x)$ is unknown function and is assumed smooth and absolutely continuous on interval $[a,b]$. ε is independent and identical random error with mean zero and variance σ^2. Estimator of $f(x)$ can be obtained by minimizing penalized residual sum of squares as follow

$$RSS(f,\lambda) = \sum \{y - f(x)\}^2 + \lambda \int \{f''(x)\}^2 dx \tag{2}$$

where the smoothing parameter λ controls the goodness of fit and smoothness level of f and the best smoothing parameter can be chosen by applying generalized cross validation. It also should be noted that the solution of penalized residual sum of squares as Equation (2) is natural spline.

Logistic Regression is a family of regressions to model the relationship between categorical target with one or more input variables either metric or non-metric. Binary logistic regression is the simplest form of logistic regression. If the response variable y consists of two categories: "success" and "failed" denoted by $y = 1$ (success) and $y = 0$ (failed), and x is the predictor variable, then the model specification for logistic regression is

$$\Pr(y = 1 \mid x) = \pi(x) = \frac{\exp^{(\beta_0 + \beta_1 x)}}{1 + \exp^{(\beta_0 + \beta_1 x)}} \tag{3}$$

where $\pi(x)$ denotes the probability of success and it usually is transformed by logit transformation for estimation parameter purposes. By logit transformation, it implies,

$$\ln\left[\frac{\Pr(y = 1 \mid x)}{\Pr(y = 0 \mid x)}\right] = \beta_0 + \beta_1 x \tag{4}$$

The parameters of logistic regression model are estimated by using maximum likelihood estimation method. This model will perform using `glm` function available in R software [4].

Nonparametric Logistic Regression is developed by combining logistic and nonparametric model such that the model specification is

$$\Pr(y = 1 \mid x) = \pi(x) = \frac{\exp^{f(x)}}{1 + \exp^{f(x)}} \tag{5}$$

which implies,

$$\ln\left[\frac{\Pr(y = 1 \mid x)}{\Pr(y = 0 \mid x)}\right] = f(x) \tag{6}$$

Fitting $f(x)$ as smooth function could be done by minimizing penalized log-likelihood:

$$\begin{aligned} l(f,l) &= \sum_{i=1}^{n}\left[y_i \log \pi(x) + (1-y_i)\log(1-\pi(x))\right] - \frac{1}{2}\lambda\int\{f''(x)\}^2 dx \\ &= \sum_{i=1}^{n}\left[y_i f(x_i) - \log\left(1+e^{f(x)}\right)\right] - \frac{1}{2}\lambda\int\{f''(x)\}^2 dx \end{aligned} \tag{7}$$

Minimizing (7) is about how to choose the optimal smoothing parameter λ. It can be solved by applying Unbiased Risk Estimator (UBRE) criteria. This model performs by using `mgcv` package available in R [5].

The goodness of classifier will be evaluated using accuracy and ROC Curve. Confusion matrix in Table 1 is used to calculate the accuracy.

TABLE 1. Confusion Matrix

Actual Group		**Predicted Group**	
		1	**0**
Y	**1**	True Positive (n_{11})	False Positive (n_{10})
	0	False Negative (n_{01})	True Negative (n_{00})

$$Accuracy = \frac{n_{00}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}} \tag{8}$$

To predict group membership, either observation is in group 0 or 1, and the optimal cut off value will be used as in library `InformationValue` [6]. The accuracy will be between 0 and 1, the greater is the better. Plotting true positive rate and false positive rate will produce ROC curve and the performance of binary classifier can be seen from area under the curve. Area under this curve can be calculated by applying library `ROCR` [7].

To demonstrate the nonparametric logistic regression we use the secondary data from the survey of Indonesian Family Life Survey (IFLS) phase 5 that was conducted in 2014 or referred to as IFLS-5 held by RAND Labor and Population (available at www.rand.org). It was a household survey that was implemented in 13 locations in Indonesia. The locations are Jakarta, West Java, East Java, South Kalimantan, South Sulawesi, South Sumatera, West Nusa Tenggara, Central Java, Yogyakarta, Bali, North Sumatera, West Sumatera and Lampung. The map of survey location is showed in Fig. 1.

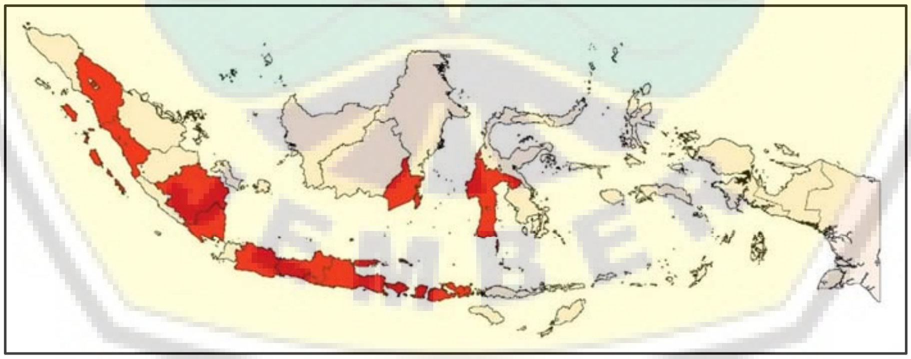

FIGURE 1. Survey Location of IFLS

There are 4,431 marriage women respondents. For modeling purposes, the data is splitted into two groups, training and testing data. The training data is used to reach the best model and the testing data will be used for evaluating the model. The dataset will be divided by 80% as the training data dan 20% as testing data. Meanwhile, the variables are presented in Table 2.

TABLE 2. Variables in the research

Indicator	Description	Scale
Working Status of Housewife (Y)	0: Not-working 1: Working	Nominal
Highest Education (X_1)	0 : No School 1 : Elementary School 2 : Junior High School 3 : Senior High School 4 : Higher Education	Ordinal
Age (X_2)	-	Ratio
Household Expenditure (X_3)	-	Ratio

RESULTS AND DISCUSSION

The exploration of training data is presented in Fig. 2. There are 41% (1,428) of total respondents who are classified in not-working group and 59% (2,116) of total respondents are classified in working group. Most of respondents' education level are elementary, junior, and senior high school. In term of education, the number of respondent in working group is higher than not-working group for all education level. In addition, there is a sizeable gap of frequencies between working and not-working group in elementary and higher education.

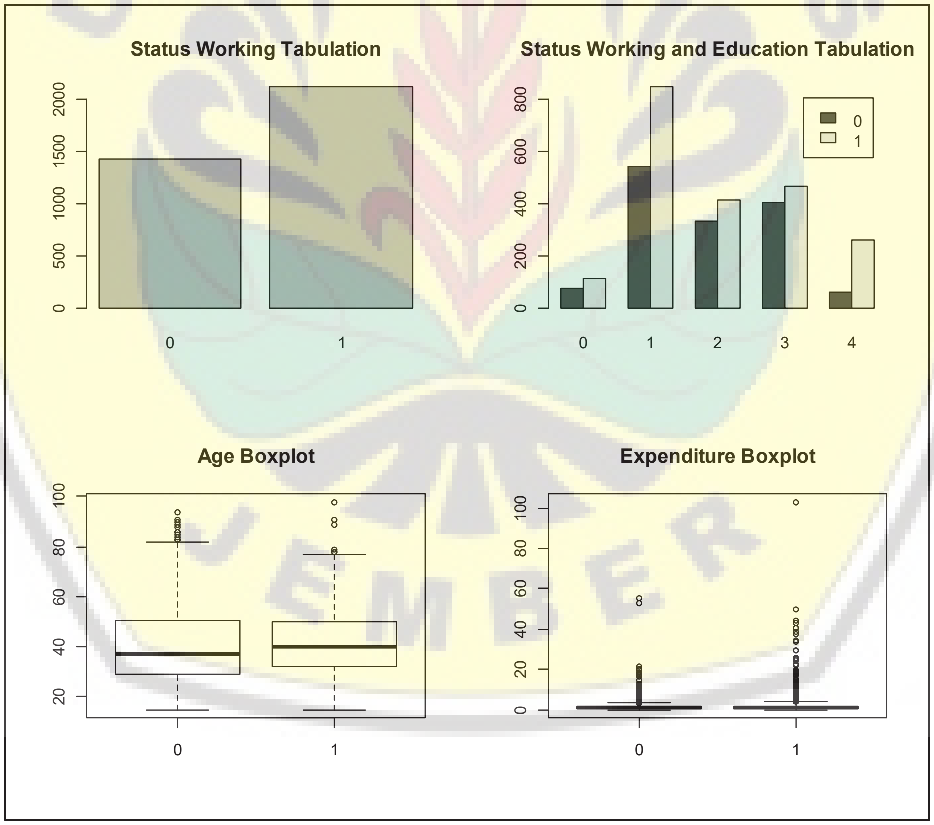

FIGURE 2. Graph summary of variables

Boxplot of age variable in Fig. 2 exhibits that the two groups tend to have the same centrality with some outliers. Conversely, the boxplot of expenditure variable is distributed differently between two groups. Furthermore, it shows that the tail of working group is longer than not-working group. Table 3 shows the descriptive statistics of age and expenditure for both groups. Then, it clears that the average of age between working group and not-working group is almost the same each other. However, the working group have the greater average and standard deviation expenditure than not-working group.

TABLE 3. Statistics Summary

	Age		Expenditure	
Status	**Mean**	**SD**	**Mean**	**SD**
Not Working	41	15	1,505,000	2,916,286
Working	42	12	1,885,000	4,242,965

The result of logistic regression fitting model is presented in the appendix. Using $\alpha = 0.05$, there is only one coefficient that is significant and the rest are not significant. However, either significant or not, all coefficient will be included to the model due to the main purpose of the model is for prediction. The logistic regression model is

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = 0.250 + 0.031X_1(1) - 0.168X_1(2) - 0.256X_1(3) + 0.951X_1(4) + 0.0023X_2 + 0.020X_3 \quad (9)$$

There are four coefficients for education variable, which refer to five categories of education. One of categories will be reference category for four another categories of education.

Evaluation of this model using confusion matrix for the training and testing data is shown in Table 4. From Table 4, it could be computed the classification accuracy of binary logistic regression model for both training and testing data.

TABLE 4. Confusion Matrix of Binary Logistic Regression

Actual Group		**Training Prediction**		**Testing Prediction**	
		0	**1**	**0**	**1**
Y	**0**	50	1,378	13	359
	1	27	2,089	2	513

$$Accuracy\ (training\ data) = \frac{50+2089}{50+1378+27+2089}\ x\,100\% = 60.36\%$$

$$Accuracy\ (testing\ data) = \frac{13+513}{13+359+2+513}\ x\,100\% = 59.30\%$$

The accuracy is 60.36% for training data and 59.30% for testing data. The accuracy is not good enough and almost the same between training data and testing data. The ROC curve of fitted logistic regression model is shown in Fig. 3 for both training and testing data. Area under curve is 0.5771 for training data and 0.5729 for testing data. It is noted that the Area Under Curve (AUC) is almost the same.

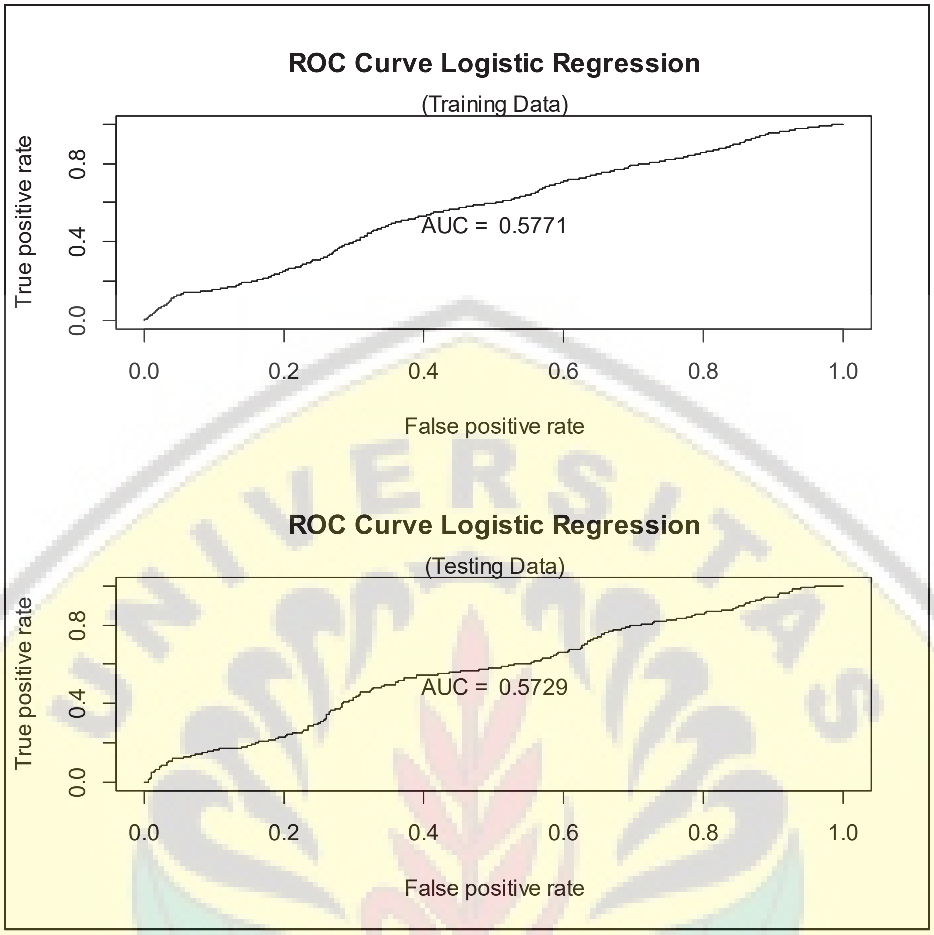

FIGURE 3. ROC Curve of Logistic Regression Model

The result of fitting nonparametric logistic model is shown in appendix. The predictor variables in nonparametric regression are classified as parametric and nonparametric variables. Education is treated as parametric as well as age and household expenditure as nonparametric variables. Inference to the fitted model shows that some coefficients are significant and the others are not significant. However, due to the main purpose of the model is for prediction, so all coefficients will be included to the model. The nonparametric logistic model can be written as,

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right)=0.714-0.3237X_1(1)-0.416X_1(2)-0.544X_1(3)+0.6428X_1(4)+s(X_2)+s(X_3) \qquad (10)$$

The notation $s(X_2)$ and $s(X_3)$ are spline smoothing of age and household expenditure. The nonparametric model is only used for continuous predictor variables – i.e. age (X_2) and household expenditure (X_3). The fitted values of nonparametric model is presented in Fig. 4. It shows that the smoothing parameter of age is 0.5079 and the smoothing parameter of expenditure is 0.00043. These optimal smoothing parameters are chosen by UBRE. In addition, the effective degrees of freedom (edf) of age and household expenditure variables are bigger than 1, which mean that both variables have non-linear relationship with the logit. The curves in Fig. 4 also strengthen that both age and household expenditure form nonlinear patterns.

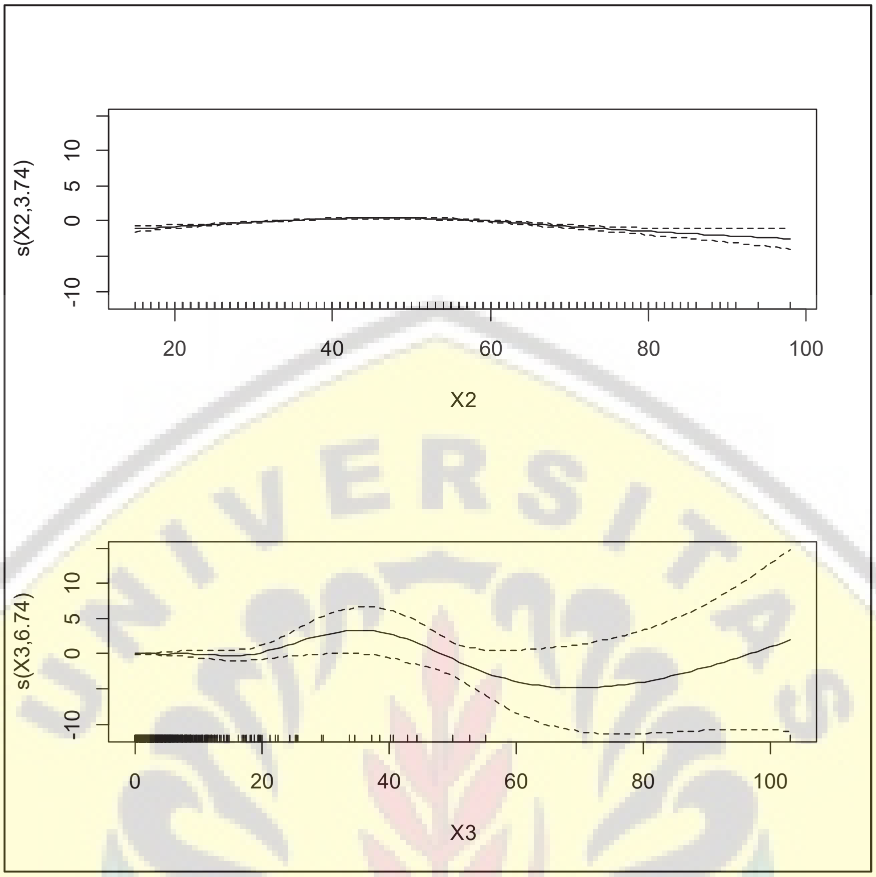

FIGURE 4. Graph of smoothness pattern of age and household expenditure

Confusion matrix of training data and testing data in Table 5 is used to assess the nonparametric logistic model. It shows that the accuracy of nonparametric logistic model of training data is 63.43% and the accuracy of testing data is 64.94%. It is almost the same between training data and testing data, but the accuracy of testing data is higher than training data.

TABLE 5. Confusion Matrix of Nonparametric Logistic Regression

Actual Group		**Training Prediction**		**Testing Prediction**	
		0	**1**	**0**	**1**
Y	**0**	526	902	141	231
	1	394	1722	80	435

$$Accuracy\ (training\ data) = \frac{526+902}{526+902+394+1722}\ x\,100\% = 63.43\%$$

$$Accuracy\ (testing\ data) = \frac{141+435}{141+231+80+435}\ x\,100\% = 64.94\%$$

The ROC curve of nonparametric logistic regression model is shown in Fig. 5 for both training and testing data. Area under curve is 0.6515 for training data and 0.6699 for testing data. It is noted that the AUC is almost the same.

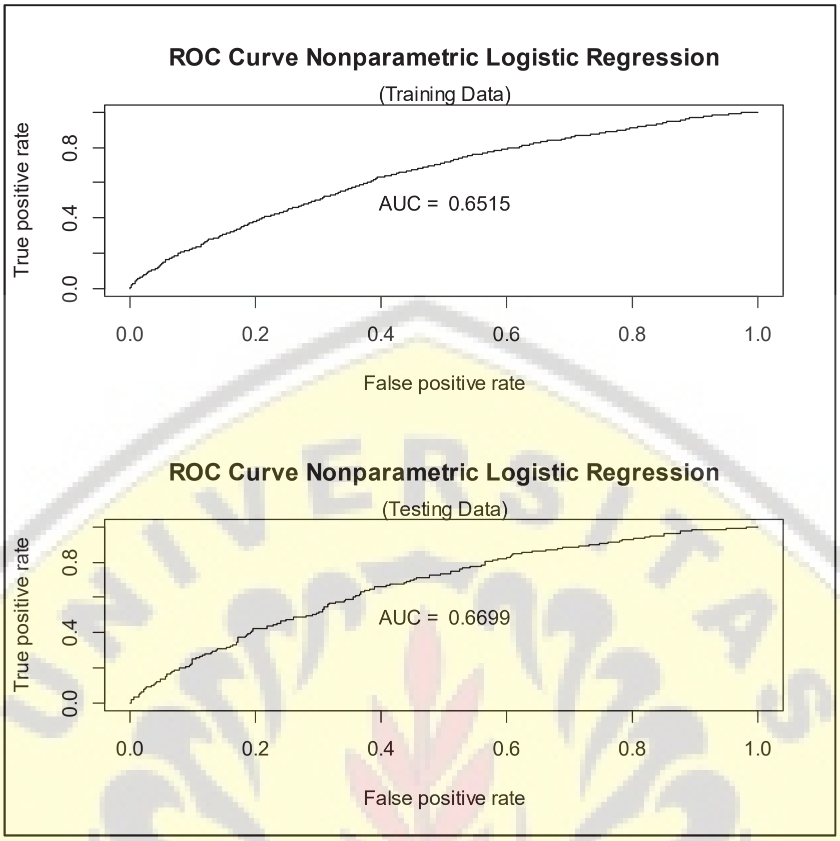

FIGURE 5. ROC Curve of nonparametric logistic regression

CONCLUSION

The result of fitting model shows that by incorporating nonparametric model to the binary logistic regression model can improve the classification accuracy. It is indicated not only by accuracy percentage, but also by area under ROC curve. Accuracy and area under curve of nonparametric logistic regression is higher than those of binary logistic regression. It happens because the main property of nonparametric model approach follows any pattern of the data, especially between response and predictor. However, incorporating nonparametric model appropriate only if there is any continuous predictor variables in the model. The classification accuracy of the binary logistic regression model is 60.36% for training data and 59.30% for testing data. Meanwhile, the classification accuracy of nonparametric logistic model is 63.43% for training data and 64.94%. for testing data. It can be concluded that nonparametric logistic model yields better classification accuracy for both training and testing data.

REFERENCES

1. T. Hastie, R. Tibshirani,, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer New York Inc., New York, NY, USA, 2001), pp. 9-35.
2. M.F. Delgado, E. Cernadas, S. Barro, D. Amorim, J.Mach.Learn.Research **15**, 3133-3181 (2014)
3. Statistics Indonesia, Labor Market Indicators Indonesia February 2017, available at www.bps.go.id
4. R Core Team , R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017, see https://www.R-project.org/
5. S.N.Wood, J. Royal. Stat. Soc 73, 3-36 (2011)
6. S. Prabhakaran, InformationValue: Performance Analysis and Companion Functions for Binary Classification Models. R package version 1.2.3, 2016, see https://CRAN.R-project.org/package=InformationValue
7. T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, Bioinformatics.Appl.Note 21, 3940-3941 (2005)