ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2, Issue 1, April 2022, pp. 47-55

Coronary Heart Disease Risk Prediction Using Binary Logistic Regression Based on Principal Component Analysis

M. Fauzan Azhari [a,1,*], Farah Ayu Fitriani [b,2,*]

[a] Department of Informatics, Universitas Islam Indonesia, Jalan Kaliurang km 14.5, Yogyakarta 55584, Indonesia
[b] Department of Statistics, Universitas Islam Indonesia, Jalan Kaliurang km 14.5, Yogyakarta 55584, Indonesia
[1] fauzan.azharie@gmail.com*; [2] 18611002@students.uii.ac.id*

ARTICLE INFO

Article history
Received February 27, 2022
Revised April 20, 2022
Accepted April 27, 2022

Keywords
Coronary Heart
Multicollinearity
Principal Component Analysis
Binary Logistic Regression

ABSTRACT

Based on data from the World Health Organization (WHO), one type of heart disease namely coronary heart disease is the deadliest disease in the world. In 2016 at least 9,4 million people died caused by coronary heart disease. In Indonesia, deaths caused by heart disease, blood vessel (CVD), and respiratory disorders are the fourth highest in ASEAN (23,1%). Because of the danger of coronary heart disease, we need a system or model that can predict heart disease early, so that it can be treated early and can reduce the death rate caused by heart disease. This study uses principal component analysis (PCA) to make a linear combination of variables that have a high correlation so that the assumption of multicollinearity in the data can be resolved. For the prediction, this study uses binary logistic regression to predict heart disease based on existing factors. The result of the PCA there is 7 component variables with a total variance that can be explained as much as 72,9%. From the Bartlett test of the PCA data, the obtained p-value is 1 which means that there is no multicollinearity in the data. Predictive analysis using binary logistic regression based on PCA's data was proven to increase the accuracy to 85%.

1. Introduction

The heart is one of the vital organs in the human body. The heart has the function of circulating oxygenated blood to all parts of the body, but there are conditions where the heart is damaged so that oxygenated blood cannot flow to all aspects of the body. This condition is called heart disease and can interfere with our daily activities and even cause death. Heart disease is one of the deadliest diseases that generally affects older people, but heart disease also can affect adults or even teenagers. Heart disease is when our heart cannot function correctly and gets disturbed [1]. People who suffer from heart disease usually feel chest pain and find it hard to breathe, but in some cases, it can occur without any symptoms at all.

Generally, heart disease is caused by an unhealthy lifestyle, consuming too many foods that are high in carbohydrates, smoking, consuming alcohol or caffeine and rarely doing physical activities [2]. In addition, heart disease can also be caused by other factors such as inherited from the family, infection by viruses or bacteria such as group A beta-hemolytic streptococcus, and congenital abnormalities such as dead heart muscle or heart valve abnormalities [3].

Based on data from the World Health Organization (WHO), one type of heart disease, namely coronary heart disease, is the deadliest disease in the world. In 2016, at least 9.4 million people died from coronary heart disease. In Indonesia, deaths from heart disease, blood vessels (CVD), and respiratory disorders are in the 4th highest rank in ASEAN, reaching 23.1%, even this figure is higher than the world average, which only got 19.4% [4].

Seeing how dangerous heart disease, we need a system or model that can predict heart disease in advance, so that it can be treated early and can reduce the death rate from heart disease. Machine learning is one of methods that can be used for classification and prediction. Its algorithms work by studying existing historical data sets so that patterns can be found to predict new data [5]. Unfortunately, the machine learning method does not always get accurate results. The level of accuracy of the machine learning model can be influenced by one of the conditions of the existing dataset. If the existing factor variables have high correlation, then multicollinearity will be occurred, this assumption must be satisfied. Therefore, the author will try to use one of the methods in unsupervised learning, namely principal component analysis to overcome the problem of multicollinearity in the data. After the multicollinearity assumption is met, then supervised learning analysis is carried out using binary logistic regression to perform modeling that can predict whether a person is at risk of heart disease or not from the existing factors.

The purpose of this research is to try to overcome the problem of multicollinearity in the data so that the assumption test to perform binary logistic regression analysis can be fulfilled. Then from the modeling that has been done and the factors that are known to have a significant effect, it is hoped that the relevant agencies can use the relevant agencies to predict the risk of heart disease in a person so that early treatment can be carried out and help reduce mortality due to heart disease, and can be used as a reference in making programs or policies to address the problem of heart disease in Indonesia.

2. Method

2.1. Data and Data Sources

The data used in this study is secondary data from the https://raw.githubusercontent.com/sta210-sp20/datasets/master/framingham.csv site. The dataset contains variables that are indicated to be associated with the risk of heart disease. The variables included in the feature variables are gender, age, smoking (yes/no), cigarette consumption per day, blood pressure medication (yes/no), family history of stroke (yes/no), family history of hypertension (yes/ no), diabetes (yes/no), total cholesterol (mg/dL), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), body mass index (kg/m2), heart rate (beats/minute) and glucose levels (mg/dL). The target variable is TenYearCHD (0 = patient does not have a 10-year risk of future coronary heart disease; 1 = Patient has a 10-year risk of future coronary heart disease).

2.2. Spearman Correlation (Multicollinearity)

Multicollinearity is a condition where there is a fairly strong linear relationship between the independent variables in a regression model [6]. Just like multiple regression, logistic regression also requires the assumption of multicollinearity. This must be considered to affect the predictive value of the independent variable. To check whether there is a linear relationship between the independent variables, you can use the correlation test. One of the most widely used correlation tests is the Spearman correlation test. The formula used to calculate the Spearman correlation is written in the following equation [7]:

$$r_s=1-\frac{6\Sigma d_i^2}{(n^3-n)} \qquad (1)$$

where:
r_s = spearman correlation coefficient
d_i = difference between the two paired observations

n = total observations

2.3. Principal Component Analysis

PCA or principal component analysis is an analysis that can be used to reduce data dimensions horizontally (data variables) without significantly reducing the characteristics of the data [8]. The way principal component analysis works is by combining the original variables in the collapsing data into a new set of independent variables, so that principal component analysis can be used to solve the problem of multicollinearity in the data.

The steps for reducing variables are carried out using the eigenvalue and eigenvector values obtained from the data covariance matrix. The eigenvector can be calculated to the equation:

$$Y_i = \sum_{i=1}^{p} a_i x_i \tag{2}$$

where:
Y_i = eigenvector
a_i = eigenvalue
x_i = observation data
p = observation variable

In principal component analysis, the total number of initial variables is usually denoted by p and the optimal number of components (variables) that can be formed is denoted by m. So that the ideal output resulting from the principal component analysis is m<p [9].

2.4. Logistic Regression

Logistic regression analysis is an analysis that can be used to find a causal relationship or a relationship between the independent variable and the dependent variable if the dependent variable on the data is categorical. The basic difference between linear regression and logistic regression lies in the dependent variable, if in linear regression it has a dependent variable with numerical properties (interval scale & ratio), then in logistic regression, it has a dependent variable with categorical properties (nominal scale). The categories contained in the dependent variable usually consist of 2 labels such as "yes or no", "pass or fail", and others.

The model equation in logistic regression is slightly different from linear regression, wherein in the logistic regression model there is an exponential function [10]. The logistic regression model can be written with the following equation:

$$\ln\left(\frac{\hat{p}}{1+\hat{p}}\right) = \beta_0 + \beta_1 x \tag{3}$$

$$\hat{p} = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{4}$$

where:
ln = natural logarithm
$\beta_0+\beta_1 x$ = regression equation
$\hat{p}$ = logistics probabilistic

2.5. Research Stage

The flowchart of the research to be conducted is as follows:

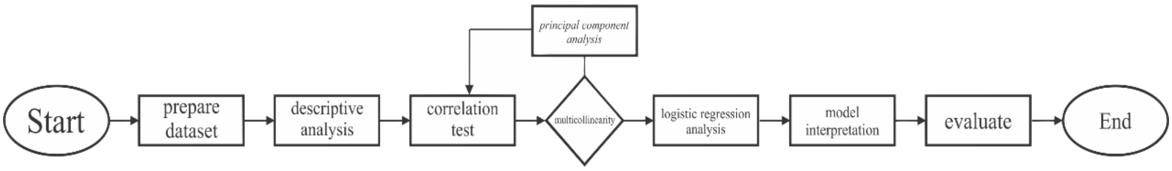


Fig. 1. Research stage.

Three methods of statistical analysis will be used in this study, namely the correlation test to check whether there is multicollinearity between the independent variables, the correlation test that will be used is the Spearman correlation test. If it is proven that there is multicollinearity in the data, then proceed to the second analysis, namely principal component analysis to reduce variables that have a high enough correlation so that the multicollinearity assumption can be fulfilled. If the assumption test on the data has been met, then proceed to predictive analysis to create a model based on the new variables.

3. Result and Discussion

3.1 Correlation Test

The correlation test used is the spearman correlation because the existing data is categorical and is not normally distributed . The results of the correlation test, shown in Figure 2, indicate that there is multicollinearity in the existing data. It can be seen that there is correlation formed between the variables. Therefore, principal component analysis must be conducted.

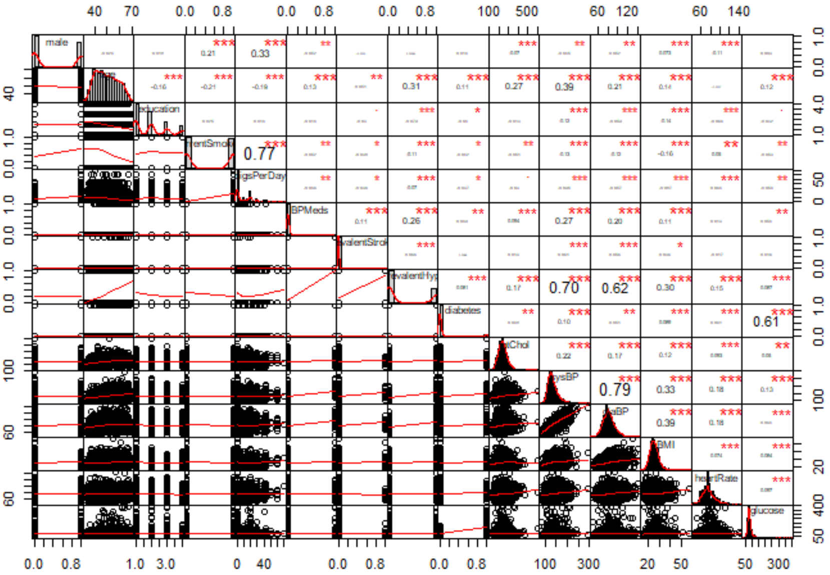


Fig. 2. Correlation test chart.

3.2. Principal Component Analysis

Process of PCA can reduce variables by combining variables that have multicollinearity into new variables. Variable reduction process is carried out based on the eigenvalues and eigenvectors. Eigenvalues are obtained from the calculation of the covariance matrix. After the eigenvalues are obtained, it will be sorted from the largest to the smallest. The eigenvectors will be selected based on the eigenvalues to reduce the variables contained the data [9]. From the results of PCA analysis, the eigenvalues are summarized in Table 1.

Table 1. Eigenvector

PC1	PC2	PC3	PC4	PC5
3.23	1.88	1.57	1.12	1.06
PC6	**PC7**	**PC8**	**PC9**	**PC10**
1.04	1.01	0.87	0.79	0.69
PC11	**PC12**	**PC13**	**PC15**	**PC15**
0.58	0.39	0.38	0.21	0.17

Determination of the optimal number of factors or components can be seen from the eigenvalue above 1, from the results of the eigenvector value above it is known that the components that have an eigenvalue above 1 are component 1 to component 7, therefore the number of new components to be used is as much as 7 components.

In addition, the determination of the optimal number of components can also be seen from the scree plot, if the determination of the optimal number of components is seen from the eigenvalues, 7 components will be taken. This can be seen through the scree plot in Figure 3, it is known that if seven components are selected to be used as new variables based on the results of the eigenvector values, then the seven components can explain the cumulative data variance of 72.9%.

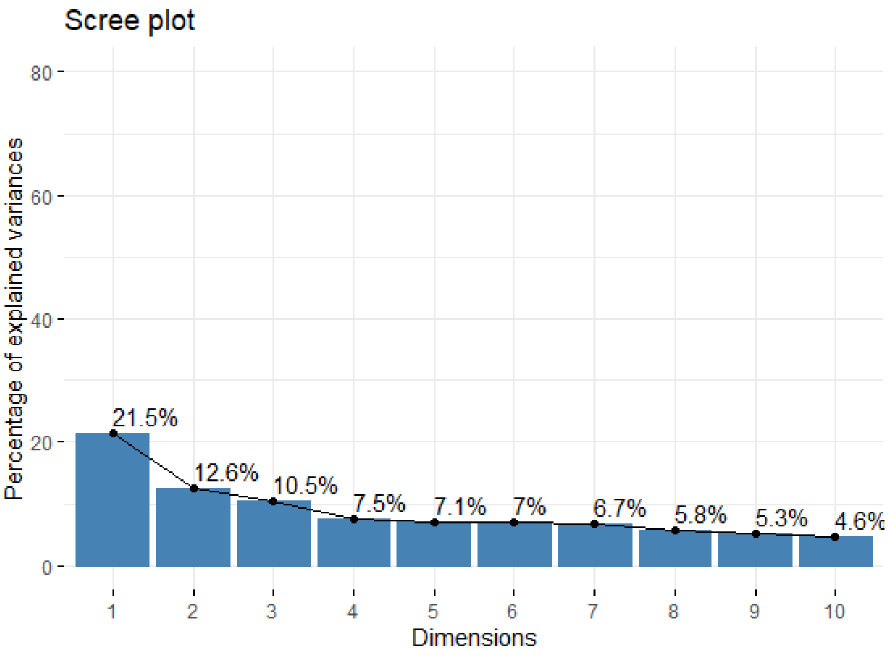
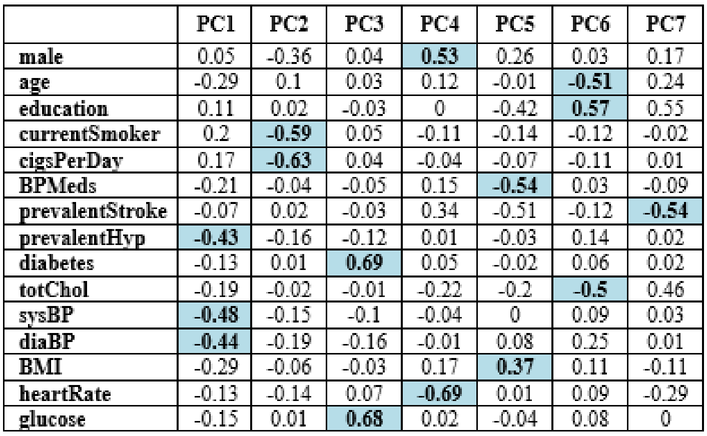


Fig. 3. Scree Plot.

Determination of members in each component can be seen from the value of the highest factor loading. From Table 2 it can be seen that the male variable is a member of the PC4 component because it has the highest factor loading value compared to other factors, which is 0.53. If there is a negative factor loading, the absolute value will be calculated, for example, the age variable will be a member of the PC6 factor because it has a factor loading value of |-0.51| = 0.51 which is the highest value compared to other factors. For other variables, the same steps were also carried out so that the members in each component can be seen in Table 4.

Table 2. Factor Loading



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
male	0.05	-0.36	0.04	**0.53**	0.26	0.03	0.17
age	-0.29	0.1	0.03	0.12	-0.01	**-0.51**	0.24
education	0.11	0.02	-0.03	0	-0.42	**0.57**	0.55
currentSmoker	0.2	**-0.59**	0.05	-0.11	-0.14	-0.12	-0.02
cigsPerDay	0.17	**-0.63**	0.04	-0.04	-0.07	-0.11	0.01
BPMeds	-0.21	-0.04	-0.05	0.15	**-0.54**	0.03	-0.09
prevalentStroke	-0.07	0.02	-0.03	0.34	-0.51	-0.12	**-0.54**
prevalentHyp	**-0.43**	-0.16	-0.12	0.01	-0.03	0.14	0.02
diabetes	-0.13	0.01	**0.69**	0.05	-0.02	0.06	0.02
totChol	-0.19	-0.02	-0.01	-0.22	-0.2	**-0.5**	0.46
sysBP	**-0.48**	-0.15	-0.1	-0.04	0	0.09	0.03
diaBP	**-0.44**	-0.19	-0.16	-0.01	0.08	0.25	0.01
BMI	-0.29	-0.06	-0.03	0.17	**0.37**	0.11	-0.11
heartRate	-0.13	-0.14	0.07	**-0.69**	0.01	0.09	-0.29
glucose	-0.15	0.01	**0.68**	0.02	-0.04	0.08	0

The final output of the principal component analysis will form a new dataset resulting from the linear combination of each member in each component. Each component and its members are as follows:

PC1 = PrevalentHyp (hypertension prevalent in family history), SysBP (systolic blood pressure), DiaBP (diastolic blood pressure)
PC2 = Current smoker, CigsPerDay (number of cigarettes smoked per day)
PC3 = Diabetes, Glucose
PC4 = Male, HeartRate
PC5 = BPMeds (not on blood pressure medications), BMI (body mass index)
PC6 = Age, Education, TotChol (total cholestrol)
PC7 = PrevalentStroke (stroke prevalent in family history)

After getting a new dataset from the results of the PCA analysis, the correlation test was carried out again to see if there was still multicollinearity in each of the new components or variables.

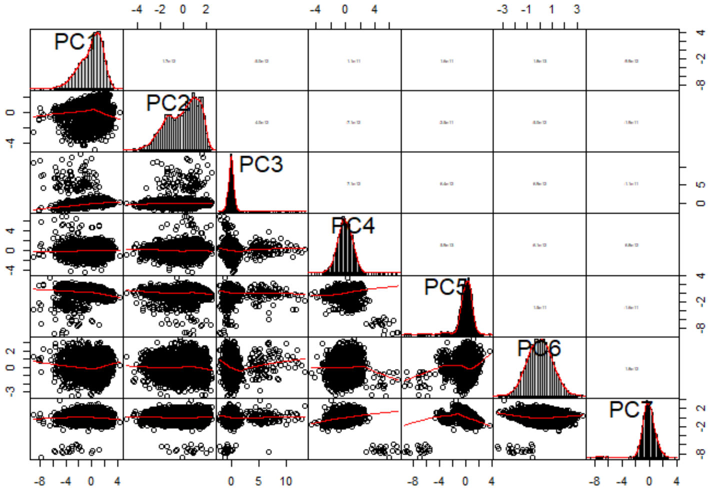


Fig. 4. Correlation test chart using PCA data.

From the correlation test using the principal component analysis data in Figure 4, it can be seen that there is no longer any correlation formed between the new variables. So it can be said that there is no more multicollinearity in the new dataset and the assumption test for logistic regression analysis has been fulfilled.

3.3. Logistics Regression Analysis

From the modeling results obtained using logistic regression analysis, it is known that of the 7 independent variables or existing components, 6 components have a significant effect in determining whether a person is at risk of heart disease or not, namely components PC1, PC2, PC3, PC4, PC6, and PC7, as evidenced by the p-value of the partial test less than (0.05). The PC5 component does not have a significant effect because the p-value is > 0.05.

Table 3. Regression Model Partial Test

Coefficients	Estimate	P-value	Description
Intercept*	-1.95182	< 2e-16	Take effect
PC1	-0.34511	< 2e-16	Take effect
PC2	-0.25770	6.75e-14	Take effect
PC3	0.10011	0.000757	Take effect
PC4	0.22301	7.91e-07	Take effect
PC5	0.01589	0.696743	No effect
PC6	-0.34189	8.45e-13	Take effect
PC7	0.17630	0.000118	Take effect

Table 3 shows the results of the partial test which shows the value of and the value of p-value of each component. So from the table, the equation of the logistic regression model can be written as follows:

$$\hat{p} = \frac{\exp\left(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6\right)}{1+\exp\left(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6\right)}$$

$$= \frac{\exp\boxed{f0}((-1,95)+(-0,34)x_1+(-0,25)x_2+0,1x_3+0,22x_4+(-0,34)x_5+0,17x_6)}{1+\exp\boxed{f0}((-1,95)+(-0,34)x_1+(-0,25)x_2+0,1x_3+0,22x_4+(-0,34)x_5+0,17x_6)} \quad (5)$$

To see the amount of accuracy obtained from the model that has been formed, a confusion matrix will be used which shows the amount of data that is predicted correctly or incorrectly. In the confusion matrix, the x-axis represents the actual data and the y-axis represents the predicted data. In this study, there are only 2 labels that will be predicted (Yes/No) so that the confusion matrix will show the amount of data with the label "Yes" which is predicted correctly (true positive), data with the label "Yes" which is predicted incorrectly (false positive), data with the label “No” which was predicted correctly (true negative), and data with the label “No” which was predicted incorrectly (false negative).

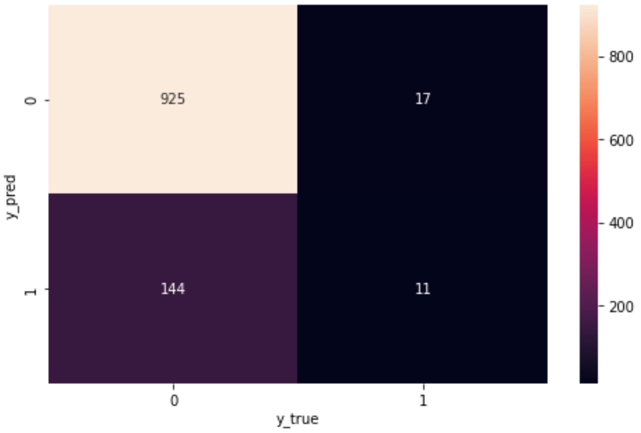


Fig. 5. Confusion matrix.

From the confusion matrix in Figure 5, it can be seen that 925 data are correctly predicted to have the label "No" (0) and as many as 11 data that are predicted to be correct have the label 1 "Yes" (1). In addition, there are as many as 144 data that should have the label "No" but are predicted to go into the "Yes" label and there are 17 data that should have the label "Yes" but are predicted to go into the "No" label. From the confusion matrix, calculations can then be made. To get the accuracy of the model, the accuracy calculation uses the following formula:

$$\text{accuracy}=\frac{\text{the amount of data predicted is correct}}{\text{the total number of data}}\times 100\%$$

$$\text{accuracy}=\frac{925+11}{1097}\times 100\%$$

$$\text{accuracy}=85\%$$

The accuracy results obtained from the logistic regression model are 85%, this shows that the model using the dataset from the principal component analysis produces a higher accuracy value when compared to the initial dataset, which is 83%. In addition, the evaluation can also be seen through the ROC (receiver operating characteristics) curve. On the ROC curve, the x-axis represents the percentage of the amount of data with a positive label (“Yes”) that is predicted to be wrong, while for the y-axis it represents the percentage of the number of data with a positive label that is predicted to be correct. If the curve line formed is close to the point (0,1) or leans towards the upper left, it can be said that the resulting model is quite good.

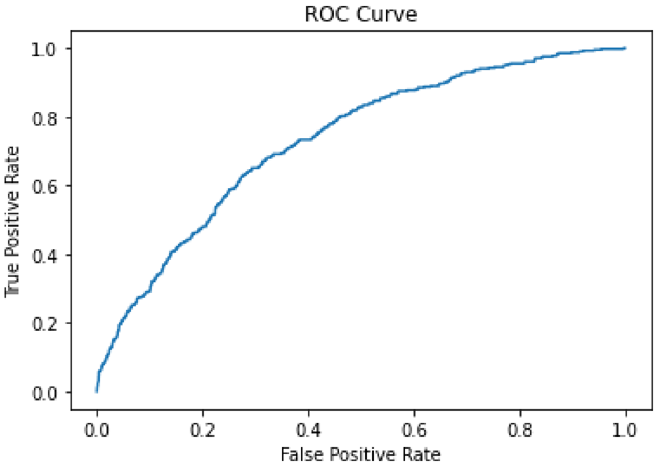


Fig. 6. ROC curve.

In Figure 6 it can be seen that the resulting ROC curve line is skewed towards the upper left, approaching the point (0.1) with an aggregate AUC (area under the curve) value of 76%.

4. Conclusion

Based on the research which has been conducted, it can be concluded:

- Principal component analysis can be used to overcome the occurrence of multicollinearity in the data so that the assumptions to proceed to logistic regression analysis can be fulfilled. Of the 15 variables contained in the initial dataset, it was reduced to only 7 variables (components) that were mutually independent. Of the 7 components that were formed managed to explain the variance of the initial data of 72.9%.
- From the logistic regression model formed, it is known that the components PC1 (prevalentHyp, sysBP, dyaBP), PC2 (currentSmoker, cigsPerDay), PC3 (diabetes, glucose), PC4 (male, heartrate), PC6 (age, education, totChol) and PC7 (prevalent stroke) has a significant effect in determining whether a person can be at risk of heart disease or not. From the logistic regression model that was formed, it managed to get an accuracy rate of 85%, higher than the regression model when using the initial dataset, which was 83%.
- The results of this study are proven to have a higher accuracy value when compared to previous studies because the assumption test before carrying out the classification process has been fulfilled using principal component analysis. So from here, suggestions for further research are to be able to use and compare the performance of methods for other feature extraction such as factor analysis, partial least squares, confirmatory factor analysis, and others and then see the effect on the results of the model in the predictive analysis whether it will make performance the model is getting better or worse.

References

[1] D. Maulana and R. Yahya, "Implementasi Algoritma Naïve Bayes untuk Klasifikasi Penderita Penyakit Jantung di Indonesia Menggunakan Rapid Miner," *SIGMA Information Technology Journal,* 2019.

[2] B. M. Metisen dan H. L. Sari, “Analisis Clustering Menggunakan Metode K-Means dalam Mengelompokan Penjualan Produk pada Swayalan Fadhila,” 2015.

[3] J. Fitriany and I. Annisa, "Demam Reumatik Akut," *Jurnal Averrous,* vol. 5, p. 2, 2019.

[4] "World Health Organization," 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[5] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *Journal of Applied Informatics and Computing,* 2020.

[6] E. Supriyadi, S. Mariani and Sugiman. , "Perbandingan Metode Partial Least Square (PLS) dan Principal Component Regression (PCR) untuk Mengatasi Multikolinearitas pada Model Regresi Linear Berganda," *UNNES Journal of Mathematics,* 2017.

[7] N. R. Sari and W. f. Mahmudy, "Fuzzy Inference System Tsukamoto untuk Menentukan Kelayakan Calon Pegawai," in *Seminar Nasional Sistem Informasi Indonesia 2015,* 2015.

[8] P. R, A. A and R. A.A., "Analisis Ekstraksi Fitur Principal Component Analysis pada Klasifikasi Microarray dan Menggunakan Classification and Regression Trees," *eProceedings of Engineering,* 2019.

[9] D. T. C. Sirait, A. and W. Astuti, "Analisis Perbandingan Reduksi Dimensi Principal Component Analysis (PCA) dan," *e-Proceeding of Engineering : Vol.6,* 2019.

[10] O. Haloho, P. Sembiring and A. Manurung, "Penerapan Analisis Regresilogistik pada Pemakaian Alat Kontrasepsiwanita," *Saintia Matematika,* vol. 1, p. 53, 2013.

[11] A. Alahmad, A. I. S. Azis, B. Santoso and S. Taliki, "Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote," *JEPIN (Jurnal dukasi dan Penelitian Informatika),* 2019.

[12] M. Khairani, A. Susanta and N. A. Y. B, "Analisis Tingkat Kognitif Soal Modul Pengayaan Kelas VIII Materi Persamaan Garis Lurus dan Sistem Persamaan Linear Dua Variabel Berdasarkan Taksonomi Bloom Revisi," *Jurnal Edukasi Matematika dan Sains,* 2021.

[13] F. Yuliani, F. Oenzil and D. Iryani, "Hubungan Berbagai Faktor Risiko Terhadap Kejadian Penyakit Jantung Koroner Pada Penderita Diabetes Melitus Tipe 2," *Jurnal Kesehatan Andalas,* vol. 3, 2014.

[14] D. Zahrawardani, K. S. Herlambang and H. D. Anggraheny, "Analisis Faktor Risiko Kejadian Penyakit Jantung Koroner di RSUP Dr Kariadi Semarang," *Jurnal Kedokteran Muhammadiyah,* vol. 1, 2012.

[15] D. L. Mihardja and H. Siswoyo, "Prevalensi dan Faktor Determinan Penyakit Jantung di Indonesia," *Buletin Penelitian Kesehatan,* vol. 44, 2016.