Enhanced Knowledge in Sciences and Technology Vol. 1 No. 2 (2021) 177-184
© Universiti Tun Hussein Onn Malaysia Publisher's Office

EKST

Homepage: http://publisher.uthm.edu.my/proceeding/index.php/ekst
e-ISSN : 2773-6385

Heart Disease Prediction Using Logistic Regression

Nor Fatihah Zulkiflee[1], Mohd Saifullah Rusiman[1]*

[1]Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, UTHM Pagoh, 86400 Muar, Johor, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2021.01.02.021
Received 11 June 2021; Accepted 19 July 2021; Available online 29 July 2021

Abstract Heart is the main organ of the human body. Heart disease is one of the disease that contribute most death in the world. It occurs when the heart is not functioning with full potential to pump the blood to all parts of the body. In this study, the researcher attempt to determine the significant variables between absence or presence of heart disease with others variables such as age, resting blood pressure, serum cholesterol, maximum heart rate achieved and others variables such as gender, chest pain type, resting electrocardiographic result, fasting blood sugar, oldpeak, slope, number of major vessels and thalassemia. The provided data covers 270 patients' information. The three methods that applied to heart disease data are Binary Logistic Regression (BLR) models, BLR models with Least Quartile Difference (LQD) method and BLR models with Median Absolute Deviation (MAD) method. After comparing among three methods, it was found that the binary logistic with applied MAD model tend to be the best model with the highest percentage of accuracy. According to the final model, x_2(gender), x_3(chest pain type), x_4(Resting Blood Pressure- in mmHg), x_5(Serum Cholesterol- in mg/dl), x_7(Resting Electrocardiographic), x_9(Exercise-Induced Angina), x_{10}(oldpeak), x_{12}(number of major vessels), x_{13}(thalassemia) and x_8 (the maximum heart rate achieved) are significant. This kind of study is to gain public awareness about the most important factor that can lead to heart disease so that they can take prompt action or make a prevention to avoid this disease from occurring.

Keywords: Heart Disease, Binary Logistic Regression, Least Quartile Difference (LQD), Median Absolute Deviation (MAD) And Percentage Of Accuracy

1. Introduction

Human body consists of several organs that perform their own role. The most important organ in the body is the heart that pumps blood into our lungs. Heart is covered by the rib cage and the skin is layered by tissue membrane called Pericardium. It has four spaces organ which separates oxygenated and deoxygenated blood. Heart disease occurs when plaque is formed in the arteries and blood vessels. Plaque is a waxy material produced by cholesterol, fat molecules, and minerals. High blood pressure, cigarette smoking, or high cholesterol or triglycerides harm an artery's inner lining [1].

**Corresponding Email: saifulah@uthm.edu.my*

Between 1990 and 2013, the number of cardiovascular disease deaths increased by 41%. Half the deaths are attributed to the same issue in the United States [2]. 17.5 million people died from CVD in 2012, representing 31 percent of all global deaths. Coronary artery disease (CAD) is a type of CVD [3]. Machine learning has been widely used to diagnose and predict the existence of diseases using medical data models. Logistic regression is one of the relatively utilized machine learning algorithms [2]. Heart disease has remained the number one cause of death in the world. An estimated 17.7 million deaths worldwide from cardiovascular diseases in 20 years [4]. 17.9 million peoples died because of cardiovascular heart disease in 2016. 85 percent of these deaths are because of heart attacks and stroke. Heart disorders impact both women and men equally [5].

Logistic regression was used to test hypotheses about the relationships of outcome variable with predictor variables [6]. Logistic regression does not require normally distributed data compared with discriminant analysis [7]. Logistic regression helps one to predict the discrete outcome from a variety of variables. For example, group size can be permanent, discrete, dichotomous, or mixed [8]. Researchers used a binary logistic regression method to predict the pattern of antidepressants in a tertiary care center. They found that female patients suffer more from depression than male patients [9].

Least Squares is one of the most common regression methods. The objective of robust plane regression is to fit a straight line through a set of two-dimensional points so the outliers is not affected the fit. To measure the robustness of an estimator [10]. The breakdown point is highest proportion of measurements that can be corrupted without forcing estimator to generate a false value. If a single observation has an infinite value, the mean of all observations is infinite; hence the breakdown point of the mean is 0.5. The median value, by contrast, remains the same. Only when more than 50% of results are infinite, the mean is insane. The Median Absolute Deviation is the "single most accurate ancillary scale calculation" [11].

The purpose of this study to identify the significant factors that affect the absence or presence of heart disease. In this study, three methods will be used which are binary logistic regression model, binary logistic regression with LQD method and binary logistic regression with MAD method. Lastly, the model from these methods will be compared by using percentage of accuracy in order to find the best model..

2. **Materials and Methods**

2.1 Data sources and data set

The heart disease data were obtained from the website of the UCI Machine Learning Repository and consists of 271 samples. One response variable and twelve independent variables were included where the response variable indicates that '0' and '1' are the results of the absence or presence of heart disease. The predictor variables are x_1(age), x_2(gender), x_3(chest pain type), x_4(Resting Blood Pressure- in mmHg), x_5(Serum Cholesterol- in mg/dl), x_6(Fasting Blood Sugar->120mg/dl), x_7(Resting Electrocardiographic), x_8(Maximum Heart Rate), x_9(Exercise-Induced Angina), x_{10}(oldpeak), x_{11}(slope), x_{12}(number of major vessels) and x_{13}(thalassemia). While the dependent variable y(presence of heart disease). This study should be able to identify the significant factors that affect the presence of heart disease and to find the prediction of the probability the presence of heart disease. Table 3.1 shows that the data description of the absence or presence of heart disease's dataset.

Table 1. Data description of absence or presence of heart disease

Variable	Description	Type of variable
Y	Absence or presence heart disease: 1 represents absence 2 represents presence	Qualitative
x_1	Age (Patient's Age)	Quantitative
x_2	Gender: 0 represents female 1 represents male	Qualitative
x_3	Chest pain type: 1 represents typical angina 2 represents atypical angina 3 represents non-anginal pain 4 represents asymptomatic	Qualitative
x_4	Resting Blood Pressure - in mmHg	Quantitative
x_5	Serum Cholesterol- in mg/dl	Quantitative
x_6	Fasting Blood Sugar- > 120 mg/dl: 0 represents false 1 represents true	Qualitative
x_7	Resting Electrocardiographic Result: 0 represents normal 1 represents having ST-T wave abnormality 2 represents probable or definite left ventricular hypertrophy by Estes' criteria	Qualitative
x_8	Maximum Heart Rate	Quantitative
x_9	Exercise-Induced Angina: 0 represents no 1 represents yes	Qualitative
x_{10}	Oldpeak (ST depression induced by exercise relative to rest)	Quantitative
x_{11}	Slope (the slope of the peak exercise ST segment): 1 represents upsloping 2 represents flat 3 represents downsloping	Qualitative
x_{12}	Number of Major Vessels -(0-3) colored by fluoroscopy	Qualitative
x_{13}	Thalassemia: 3 represents normal 6 represents a fixed defect 7 represents a reversible defect	Qualitative

2.2 Methods

In this study, there are two methods will be used which are binary logistic regression analysis and robust method. Logistic regression referred to as a logistic model or a logit model. It analyzes the relationship between multiple independent variables and a categorical [12]. Logistic regression is widely used to analyze data involving a dependent, dichotomous, or binary outcome variable against the independent variable. Normality data with an equal variance and covariance for all variables is not required to perform logistic regression. The general model of explanatory variable as suggested by [13],

$$logit(y) = \ln\left(\frac{p}{1-p}\right) = \alpha + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + e \qquad Eq.1$$

There are two robust methods that will be used which are least quartile difference (LQD) method and median absolute deviation (MAD). The first step is test the multicollinearity test between the independent variables using spearman's correlation. Next is testing the significance of variable. The binary logistic regression method will be done if the data fulfil the assumptions. Lastly, the model from both binary logistic regression analysis and robust method analysis will be compared by using the percentage of accuracy in order to find the best model and significant factors for the presence of heart disease.

Next, the model of binary logistic regression (BLR) is run in Statistical Packages for Social Sciences (SPSS). Then the value of percentage accuracy is obtained. The regression estimator for the least quartile difference (LQD) is highly robust. It can increase up to nearly 50% of largely deviant data values without becoming extremely biased [10]. The method of Least Quartile Difference (LQD) and Mean Absolute Deviation (MAD) run in Microsoft Excel. Then the model of MLR with applied LQD and model of MLR with applied MAD are obtained as in Eq.2 and Eq.3 [10, 11].

$$\text{LQD model: } r_i(L) = y_i - \alpha \text{ - } b_ix_i \qquad Eq.2$$

$$MAD = median(|x_i - median(x)|\; i = 1,2,\ldots,n) \qquad Eq.3$$

Finally, there are comparison between all the percentage of accuracy value of the model in order to find the best model. To compare the accuracy of the method, the overall percentage accuracy can be utilized. The formula to calculate the percentage of accuracy is stated in the Eq. 4 [13],

$$\text{Percentage of predictive frequency} = \frac{\text{number of accurate data}}{\text{total data}} \times 100 \qquad Eq.\ 4$$

1. Results and Discussion

3.1 Binary Logistic Regression

In this study, the Spearman's correlation was performed as in Table 2. Spearman's Rank Correlation Coefficient is used to discover the strength of a link between two sets of data. A Spearman's correlation coefficient of greater than 0.9 indicate the existence of multicollinearity. However, based on the analysis, none of the variables have the value of Spearman's correlation coefficient exceeding 0.9. This shown that multicollinearity does not exist.

Table 2. Spearman's Correlation

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	1	-.099	.121	.277	.211	.118	.128	-.400	.095	.258	.174	.375	.109
x_2	-.099	1	.067	-.055	-.158	.042	.039	-.076	.180	.112	.051	.104	.382
x_3	.121	.067	1	-.002	.100	-.108	.110	-.349	.402	.212	.192	.251	.305
x_4	.277	-.055	-.002	1	.190	.134	.117	-.043	.070	.183	.107	.058	.107
x_5	.211	-.158	.100	.190	1	.033	.164	-.056	.104	.014	-.001	.155	.020
x_6	.118	.042	-.108	.134	.033	1	.054	.015	-.004	-.009	.021	.116	.046
x_7	.128	.039	.110	.117	.164	.054	1	-.115	.095	.110	.162	.111	.005
x_8	-.400	-.076	-.349	-.043	-.056	.015	-.115	1	-.394	-.423	-.427	-.290	-.269
x_9	.095	.180	.402	.070	.104	-.004	.095	-.394	1	.276	.271	.186	.321
x_{10}	.258	.112	.212	.183	.014	-.009	.110	-.423	.276	1	.625	.242	.339
x_{11}	.174	.051	.192	.107	-.001	.021	.162	-.427	.271	.625	1	.122	.288
x_{12}	.375	.104	.251	.058	.155	.116	.111	-.290	.186	.242	.122	1	.266
x_{13}	.109	.382	.305	.107	.020	.046	.005	-.269	.321	.339	.288	.266	1

Next, logistic regression was run with 4 steps in backward logistic regression to provide an overview of the significant of each variable by using SPSS software as in Table 3. Ten independent variables are significant with a *p*-value less than 0.10 whereas 3 independent variables are not significant due to the *p*-value of these variables were more than 0.10 which are x_1(age), x_6(Fasting Blood Sugar->120mg/dl) and x_{11}(slope). The analysis shows that the heart disease is directly proportional with x_2(gender), x_3(chest pain type), x_4(Resting Blood Pressure- in mmHg), x_5(Serum Cholesterol- in mg/dl), x_7(Resting Electrocardiographic), x_9(Exercise-Induced Angina), x_{10}(oldpeak), x_{12}(number of major vessels) and x_{13}(thalassemia) whereas the heart disease inversely proportional to x_8 (the maximum heart rate achieved).

Table 3. The BLR model with coefficients

		B	S.E.	Wald	df	Sig.	Exp(*B*)
Model Step 4	x_2	1.492	.526	8.031	1	.005	4.445
	x_3	.730	.213	11.788	1	.001	2.075
	x_4	.020	.011	3.543	1	.060	1.021
	x_5	.006	.004	2.709	1	.100	1.006
	x_7	.329	.196	2.810	1	.094	1.389
	x_8	-.021	.009	5.032	1	.025	.979
	x_9	.823	.429	3.685	1	.055	2.278
	x_{10}	.496	.195	6.487	1	.011	1.642
	x_{12}	.990	.241	16.936	1	.000	2.692
	x_{13}	.733	.211	12.121	1	.000	2.082
	Constant	-7.833	2.450	10.218	1	.001	.000

From Table 2, the BLR model is shown as in (Eq. 5),

$$\ln\left(\frac{p}{1-p}\right) = -7.833 + 1.492x_2 + 0.730x_3 + 0.02x_4 + 0.006x_5 + 0.329x_7 - 0.21x_8 + 0.823x_9 + 0.496x_{10} + 0.990x_{12} + 0.733x_{13} \qquad Eq.\ 5$$

3.2 Binary Logistic Regression with applied LQD model

The data will use BLR model with least quartile difference method where the model is shown in Eq. 6.

$$\ln\left(\frac{p}{1+p}\right) = -5.958 + 1.325x_2 + 0.630x_3 + 0.018x_4 + 0.005x_5 + 0.209x_7 - 0.025x_8 + 0.688x_9 + 0.275x_{10} + 1.136x_{12} + 0.795x_{13} \qquad Eq.\ 6$$

3.3 Binary Logistic Regression with applied MAD model.

The MLR model with mean absolute deviation method is shown in Eq. 7.

$$\ln\left(\frac{p}{1-p}\right) = -46.954 + 21.099x_2 + 0.354x_3 + 0.309x_4 + 0.012x_5 + 5.662x_7 - 0.333x_8 + 4.316x_9 + 4.648x_{10} + 18.048x_{12} + 5.271x_{13} \qquad Eq.\ 7$$

3.3 Comparison all models

In order to find the best model, the comparison of percentage accuracy within all models has been made as in Table 4.

Table 4. Comparison Model

Model	Percentage Accuracy Value
Binary Logistic Model	85.9%
Binary Logistic Model with applied LQD method	84.4%
Binary Logistic Model with applied MAD method	86.6%

Based on Table 4, the binary logistic model with the applied MAD model has the highest value of percentage accuracy with a value of 86.6%. This can be concluded that the binary logistic with applied MAD model was the best model among others model. Therefore, the best chosen model is indicated as in Eq. 7.

4. Conclusion

In conclusion, according to the final model where BLR with MAD method, The analysis shows that the heart disease is directly proportional with x_2(gender), x_3(chest pain type), x_4(Resting Blood Pressure- in mmHg), x_5(Serum Cholesterol- in mg/dl), x_7(Resting Electrocardiographic), x_9(Exercise-Induced Angina), x_{10}(oldpeak), x_{12}(number of major vessels) and x_{13}(thalassemia) whereas the heart disease inversely proportional to x_8 (the maximum heart rate achieved). Based on the comparison of percentage accuracy values for the three models, the binary logistic model with the applied MAD model has the highest value of percentage accuracy with a value of 86.2%. This can be concluded that the binary logistic with applied MAD model was the best model among others model. This binary logistic regression with applied MAD model can be used for future modelling for the heart disease prediction in hospital management.

Acknowledgement

I would like to express my deep and sincere gratitude to my research supervisor Prof Madya Dr Mohd Saifullah bin Rusiman for his advice, patience, support and guidance throughout the completion of my project. The research work is supported by Ministry of Higher Education, Malaysia with Fundamental Research Grant Scheme (FRGS) grant (Vot K297), reference number FRGS/1/2020/STG06/UTHM/02/4.

References

[1] Prasad, R., Anjali, P., Adil, S., & Deepa, N. (2019). *Heart disease prediction using logistic regression algorithm using machine learning. International Journal of Engineering and Advanced Technology*, *8*(3 Special Issue), 659–662.

[2] Nishadi, A. S. T. (2019). *Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab*. *3*(8), 69–74.

[3] Dharani, M. M. K., & Poovitha, C. (2017). *a Data Mining Model To Predict the Risk of Heart Disease Using Multinomial Logistic Regression (Mlr)*. *8*(7), 66–70. http://www.ijser.org.

[4] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics, 36*, 82–93.

[5] Varun, S. A., Mounika, G., Sahoo, P. K., & Eswaran, K. (2019). *Efficient System for Heart Disease Prediction by applying Logistic Regression*. *8491*, 13–16.

[6] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3–14.

[7] Dutta, A., Bandopadhyay, G., & Sengupta, S. (2012). Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression. *International Journal of Business and Information*, *7*(1), 105–136.

[8] Diaz, A. A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M. J., & Harrison, R. G. (2010). *Prediction of Protein Solubility in Escherichia coli Using Logistic Regression*. *105*(2), 374–383.

[9] Banerjee, I., Banerjee, I., Roy, B., & Sathian, B. (2013). Application of Binary Regression Analysis in the Prescription Pattern of Antidepressants. *Medical Science, 1*(1), 19.

[10] Bernholt, T., Nunkesser, R., & Schettlinger, K. (2007). Computing the least quartile difference estimator in the plane. *Computational Statistics and Data Analysis, 52*(2), 763–772.

[11] Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*(424), 1273–1283.

[12] Ae, H. (2013). *An Introduction to Logistic Regression : From Basic Concepts to Interpretation with Particular Attention to Nursing Domain*. *43*(2), 154–164.

[13] Rusli, N. M., Ibrahim, Z., Janor, R. M., & Teknologi, U. (2008). *Predicting Students ' Academic Achievement : Comparison between Logistic Regression , Artificial Neural Network , and Neuro-Fuzzy*. *2005*, 1–6.

[14] IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969 (Example for a standard).

[15] Williams, J. O., “Narrow-band analyzer,” PhD dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993 (Example for a thesis).