ISSN 2355-0082

Logistic Regression Prediction Model for Cardiovascular Disease

Tania Ciu[1], Raymond Sunardi Oetama[2]
[1, 2] Department of Information System, Universitas Multimedia Nusantara, Tangerang, Indonesia
tania.ciu@student.umn.ac.id
raymond@umn.ac.id

Accepted on April 10, 2020
Approved on June 15, 2020

***Abstract*—It is undeniable that cardiovascular disease is the number one cause of death in the world. Various factors such as age, cholesterol level, and unhealthy lifestyle can trigger cardiovascular disease. The symptoms of cardiovascular disease are also challenging to identify. It takes careful understanding and analysis related to patient medical record data and identification of the parameters that cause this disease. This study was conducted to predict the main factors causing cardiovascular disease. In this study, a dataset consisting of 14 attributes with class labels was used as the basis for identification as a link between factors that cause cardiovascular disease. The research area used is the area of analysis data where the analyzed data are on factors that influence the presence of cardiovascular disease in the State of Cleveland. In predicting cardiovascular disease, a logistic regression algorithm will be used to see the interrelation between the dependent variable and the independent variables involved. With this research, it is expected to be able to increase readers' knowledge and insight related to how to analyze cardiovascular disease using logistic regression algorithms and the main factors that cause cardiovascular disease.**

***Index Terms*—Decision Tree, K-Means Algorithm, Logistic Regression, Naïve Bayes**

I. INTRODUCTION

The number of cardiovascular disease sufferers is also increasing yearly. This disease occurs due to several factors, such as age, blood pressure, cholesterol levels, diabetes, hypertension, genes, obesity, and unhealthy lifestyles. Various symptoms can be identified through physical signs such as chest pain, shortness of breath, dizziness, and easy feeling of fatigue [1].

Cardiovascular disease identification techniques are complicated to do. It is essential to know the existence of this disease as early as possible because the complication of cardiovascular disease can give an impact on one's life as a whole. The diagnosis and treatment of cardiovascular disease are very complex. While still using invasive-based techniques through analysis of the patient's medical history, reports of physical examinations performed by the medical tend to be less accurate and require a relatively long time. For this reason, a support system is implemented to predict cardiovascular disease through a machine learning model.

Palaniappan and Awang [2] uses classification modelling techniques by digging up the information contained in cardiovascular disease data. The model is created using training data and tested with data testing as a form of evaluation of results. Sellappan Palaniappan, Rafiah Awang uses the lift chart and matrix method to evaluate the effectiveness of the model. Based on the research that has been done, they concluded that the most effective model for prediction of heart disease is Naive Bayes which found that the use of the Naive Bayes algorithm produces a better level of accuracy than the Decision Tree.

Mai Showman, Tim Turner, Rob Stocker [3] applied the K-Means method and decision tree to predict cardiovascular disease. In the research, they implemented Initial centroid selection techniques to improve the accuracy of the model. Meanwhile, to diagnose cardiovascular disease, they use the decision tree classification. To produce the initial centroids based on the actual number of samples in the data set, random rows, random attributes, inliers, outliers and range methods are used. The result, researchers can compare the performance of the decision tree that was applied previously with the K-Means method applied using the same data with the data used in the decision tree. As a result, the K-Means technique used produces better accuracy. The process shows an accuracy of around 83.9%.

II. LOGISTIC REGRESSION

Logistic regression is a predictive model used to evaluate the relationship between the dependent variable (target) which is categorical data with nominal or ordinal scale and the independent variable (predictor) which is categorical data with interval or ratio scale. This algorithm can also be used for time series modelling to find the relationship between the variables involved. Logistic regression is an algorithm used to predict the probability of categorical dependent variables. In logistic regression, the

dependent variable is shown as a binary variable that is valued at 1 (yes) Or 0 (no). The logistic regression model predicts as a function of X. The assumptions used in Logistic regression are as follows: binary logistic regression requires binary dependent variables, for binary regression, the factor 1 level of the dependent variable must represent the desired result, independent variables must be independent of each other. In this case, the model must have little or no multicollinearity and be linearly related to log opportunities [4].

Logistic regression used appropriate regression analysis to be performed when the dependent variable is dichotomous (binary). Logistic regression acts as a predictive analytical model. Logistic regression is applied to describe data and explain the relationship between one dependent binary variable with one or more independent variables at the nominal, ordinal, interval or ratio level. Logistic regression has several advantages and disadvantages. The benefits of logistic regression include the following. First, logistic regression can show a significant relationship between the dependent variable and the independent variable. Second, logistic regression analysis can also be used to compare the effect of variables measured at different scales includings the effect of price changes and the number of promotional activities. This benefit helps market researchers or data analysts to eliminate and evaluate the best set of variables that will be used to build predictive models. Third, the logistic regression model is not only a classification model, but also provides information related to probability. To achieve a better result using Logistic Regression, first all independent variable must contain their valid value. Secondly, logistic regression works well for predicting categorical results and multinomial results. Third, there is no multicollinearity between variables in the dataset [5].

III. METHODOLOGY

Human cardiovascular system is examined in this study using some variables that affect its performance. As shown on Fig. 1, the process is started from retrieve data, analyze the correlation between variables, split data, prediction with logistic regression algorithm, and finished with data validation.

A. Data Retrieval

The first process is Data Retrieval. In this process, Heart Disease UCI Dataset -published by Ronit in Kaggle website (https://www.kaggle.com/ronitf/heart-disease-uci)- will be used. It will be imported into the Rstudio software. The data obtained are categorical data and numerical data. The data in this study contain 14 variables with 76 attributes and 304 responses as the basis for analysis. First variable is age with units in years (age). Second, the gender with value one means male and value 0 means female (sex). Third, the variable type of chest pain (cp). Fourth, the variable trestbps-resting blood pressure in mm Hg at admission to hospital (tresbps). Fifth, chol-serum cholesterol variable in mg/dl (chol). Sixth, the fbs variable, which is blood sugar when fasting with a value of 1, means true, and 0 means false (fbs). The seventh variable is resting electrocardiographic outcome variables (restecg). Eighth, the maximum thalach heart rate variable is reached (thalac). Ninth, the exacting-exercise variable induced angina with value 1 means yes, and value 0 means no (exang). Tenth, oldpeak-ST variable depression caused by exercise relative to rest (oldpeak). Eleventh, the slope variables of the peak training segment ST (slope). Twelfth, ca-number of main vessels with values 0 to 3, colored by fluoroscopy (ca). Thirteenth, thal-3 variable means normal; 6 means permanent disability; 7 means reversible defects (thal). Fourteenth, the target variable with a value of 1 or 0 (target).

B. The Correlation between Variables Analysis

Besides, to facilitate data analysis, all variables in the imported dataset will be visualized in the form of a histogram to facilitate the reading of the data in general. In the process, Analyze the Correlation between Variables; the correlation between variables is examined to prove that the method to be used is the logistic regression model is the right model. Relationships between variables in the available dataset will be plotted in the form of a matrix. This is also done to check whether there is multicollinearity between variables in the dataset.

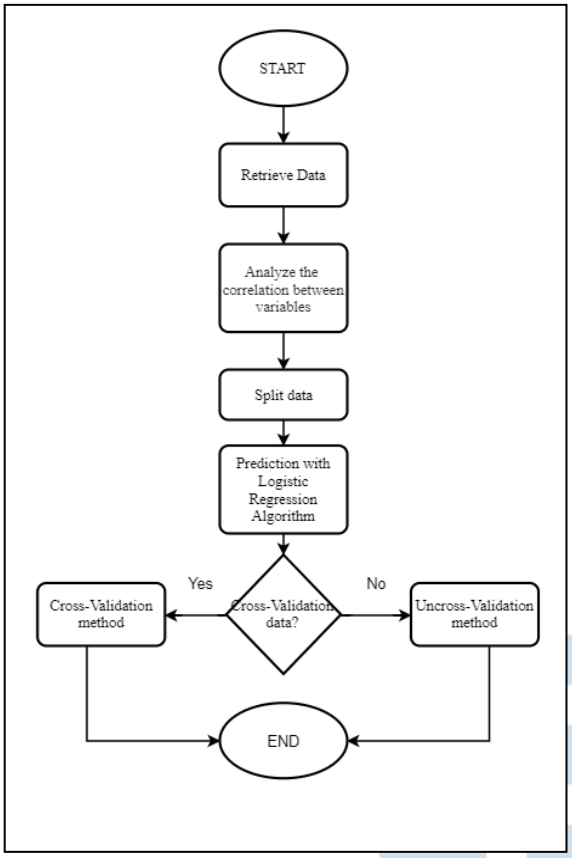

Fig. 1. Framework

C. *Data Preparation*

The dataset imported in Rstudio will be divided into two parts, namely training data and testing data. Training data is used as a basis for building models. Meanwhile, testing data is used as a basis for testing or validating the model. In this data preparation process, 293 data will be sampled. Then the data will be partitioned into train data and test data.

D. *Prediction with Logistic Regression Algorithm*

In this process, the data that has been partitioned in the previous process will be used. Prediction using the logistic regression method will produce several data that can be used as a basis for concluding to make predictions.

E. *Data Validation*

The technique used to validate the results is the method of the confusion matrix and K-fold cross-validation with 10-fold. By using a confusion matrix, the accuracy of the use of the logistic regression model can be known. Besides, the use of the K-fold cross-validation method, produces values of errors that may occur when using a logistic regression model.

IV. Data Analysis and Discussion

A. *Data Analysis*

The dataset obtained by the researcher as a basis for analysis is imported into RStudio. The output of this process can be seen on Fig. 2 and Fig. 3. The data retrieval process is also performed in the data visualization to see the value of each variable involved in the overall research analysis.

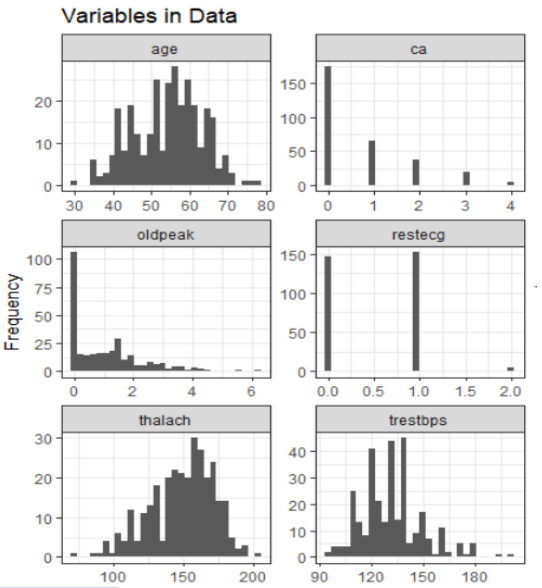

Fig. 2. Plot Variable 1

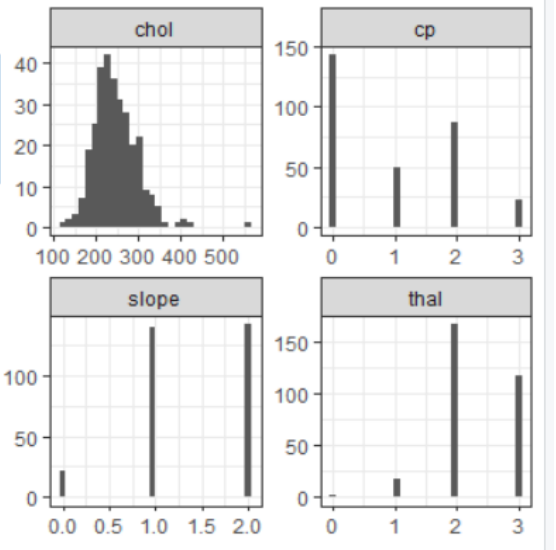

Fig. 3. Plot Variable 2

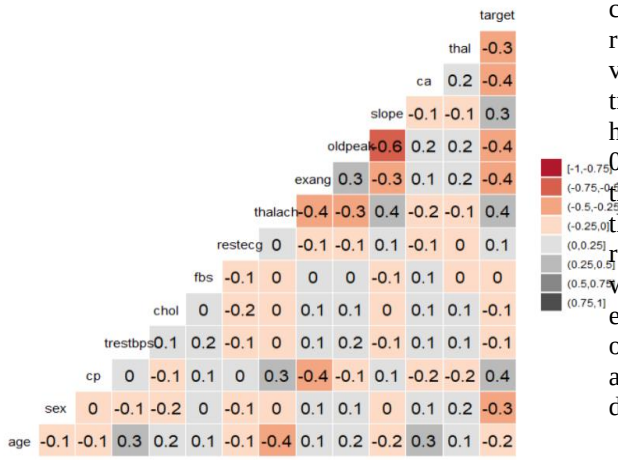

Fig. 4. Correlation Plot

In this process, the correlation between variables will be examined, which will be used as a basis for analysis to predict cardiovascular disease. Based on the matrix in Fig. 4, it was found that the variables induced angia (exang), chest pain type (cp), ST depression induced by exercise relative to rest (oldpeak), maximal heart rate (thalac) had a strong correlation with the target variable. Meanwhile, blood sugar (fbs) and cholesterol (chol) levels do not correlate with the target variable. Meanwhile, among the independent variables, there is a strong correlation between the slope and oldpeak variables. Besides, thalac, exhang, oldpeak, and slope variables are also strongly correlated. Strong correlation also applies to variables Exang, cp, and thalac. It proves that there is no multicollinearity in the relationship between variables where each independent variable does not correlate with each other.

Data that has been imported will be taken as many as 293 random data as a basis for analysis. The data is divided into train data and test data. The data shown in Fig. 5 on the next page is data from train_data and test_data that will be used in this study. The training data is used to build a logistic regression model using the glm () function because logistic regression is included in the generalized linear model with binomial type families. Based on the results of using the logistic regression method, it is predicted that the sex, cp, trestbps, restecg, ca and that variables influence the target variable at an alpha value of 5% significantly. The selected variables are the variables that significantly affect the target variable. In logistic regression, the effect of each variable on the target variable can be seen from the odds ratio value. For example, for the sex variable having a coefficient value of -1.547601 with a reference category with a male value, the odds ratio value is 4.2655 which means that for male patients, the odds of getting heart disease are 4.2655 times the female odds or it can be said the tendency of men to heart disease is higher than women. For the trestbps variable with a coefficient value of -0.029713, it is found that the odds ratio value is 0.0822 which means that for the trestbps variable there will be a significant increase when trestbps enters the value 0.0822 mmHg. On the other hand, the thalach variable with a coefficient of 0.032028 will have an odds of 0.08856 which means that at that value there will be a significant change in the performance of the heart rate or cardiovascular rate. The exang1 variable is exercise-induced angina with an estimated coefficient of -1.05855 so that the exang variable with a reference value of 1 will have an odds of 2.92710 which means that if the value is achieved then cardiovascular performance will decrease.

Next is the variable ca with reference ca values 1, 2, and 3. Ca1 with an estimated coefficient of -1.430110 will have odds of 3.955, while ca2 with an estimated ratio of -3.329874 will have odds of 9.1777 and ca3 with an estimated factor of -0.553711 will have odds in the amount of 1.5261. It proves that when the number of fluoroscopy vessels reaches its value odds, this will have an impact on decreasing cardiac performance which will affect the increased potential for cardiovascular disease.

```
Call:
glm(formula = target ~ age + sex + trestbps + chol + fbs + restecg +

    thalach + exang + oldpeak + slope + ca + thal, family = binomial
(link = "logit"),
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7562  -0.3933   0.1859   0.5167   2.6218

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.647872   3.103802  -0.209  0.83466
age          0.017395   0.029489   0.590  0.55528
sex1        -1.547601   0.612277  -2.528  0.01148 *
trestbps    -0.029713   0.012591  -2.360  0.01828 *
chol        -0.004401   0.004313  -1.020  0.30760
fbs1         0.690600   0.584336   1.182  0.23726
restecg1     0.454490   0.445116   1.021  0.30723
restecg2     0.029025   1.924828   0.015  0.98797
thalach      0.032028   0.012098   2.647  0.00811 **
exang1      -1.058555   0.472189  -2.242  0.02497 *
oldpeak     -0.400718   0.262301  -1.528  0.12659
slope        0.702040   0.418460   1.678  0.09341 .
ca1         -1.430110   0.545908  -2.620  0.00880 **
ca2         -3.329874   0.854150  -3.898 9.68e-05 ***
ca3         -1.972362   0.912773  -2.161  0.03071 *
ca4         -0.553711   1.822086  -0.304  0.76121
thal1        2.136555   1.847010   1.157  0.24737
thal2        1.860416   1.572538   1.183  0.23678
thal3        0.688015   1.611155   0.427  0.66936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 293.58  on 212  degrees of freedom
Residual deviance: 149.90  on 194  degrees of freedom
AIC: 187.9

Number of Fisher Scoring iterations: 6
```

Fig. 5. Output Logistic Regression

The method used to validate the logistic regression model used in this study is the k-fold cross-validation method with k-fold value of 10. Following is the syntax of the k-fold cross-validation method. Based on the k-fold cross-validation data method, it was found that the prediction data using the logistic regression method had an error rate that tended to be lower at 0.1406565. It proves that referring to the two validation methods that have been done, and it can be concluded that the logistic regression model is an appropriate and effective model for this research.

Besides, the composition of value 0 and value 1 on variable target is 97:116, which is still fairly balance, so the result will be reliable and free from any imbalanced dataset problems.

B. Discussion

This study involved fourteen factors that affect cardiovascular performance as variables to build a logistic regression model. Among the variables, it is not found that there is no significant relationship between variables. Therefore, the potential for multicollinearity in this study tends to be smaller. This study uses a logistic regression algorithm as a solution to the problem. With the use of the algorithm, it was found that the logistic regression algorithm was classified as an effective and efficient algorithm in predicting the main factors causing cardiovascular disease as the problem raised in this study. Confusion matrix is shown on Fig. 6. With an accuracy of 85.45% and an error rate that tends to be small at 0.1406565, the logistic regression algorithm can be said to be successful in predicting factors that affect cardiovascular performance significantly. Especially with calculations using specific estimated values, it can be obtained the probability of the potential for cardiovascular disease in a person.

By modelling data and predicting data using a logistic regression algorithm, it was found that not all factors had a significant influence on the performance of the cardiovascular system. The factors that affect cardiovascular performance are gender, trestbps - blood pressure level, thalach - heart rate, and ca- number of vessels affected by fluorosophy. By obtaining an estimated value of these factors, probabilities can be obtained related to the potential for cardiovascular disease in a person.

```
Confusion Matrix and Statistics


LogisticPred   0   1
           0  78  12
           1  19 104

               Accuracy : 0.8545
                 95% CI : (0.7998, 0.8989)
    No Information Rate : 0.5446
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7048

 Mcnemar's Test P-Value : 0.2812

            Sensitivity : 0.8041
            Specificity : 0.8966
         Pos Pred Value : 0.8667
         Neg Pred Value : 0.8455
             Prevalence : 0.4554
         Detection Rate : 0.3662
   Detection Prevalence : 0.4225
      Balanced Accuracy : 0.8503

       'Positive' Class : 0
```

Fig. 6. Model Performance

V. Conclusions

A. Conclusion

By using the Heart Disease UCI dataset consisting of fourteen variables, including age, sex, cp, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target, it was found that the use of the logistic regression algorithm is effective and efficient in predicting cardiovascular disease where based on the results of data validation it is found that the accuracy of the prediction results with the algorithm reaches 85% with an error rate that tends to be small at 0.1406565. It proves that this algorithm is suitable for use as a prediction algorithm in this study.

Based on the results of cardiovascular disease predictions, it can be concluded that cardiovascular disease is significantly affected by gender, trestbps - blood pressure level, thalach - heart rate, and ca- number of vessels affected by fluoroscopy. An increase in the value of these variables will have an impact on overall cardiovascular performance where the cardiovascular performance will decrease, while the potential for cardiovascular disease is predicted to increase. The use of the logistic regression algorithm is successful in predicting the main factors causing cardiovascular disease where the main elements of the disease are gender factors, blood pressure level factors, heart rate level factors, and blood vessel colour factors (vessels).

B. Future Work

For similar studies, it is better to use two or more algorithms. More algorithms create more results, and the best result can be chosen as the decision. In addition, Logistic regression itself is an algorithm that is quite effective in providing predictions related to cardiovascular disease. In analysing using logistic regression, binary data is needed to simplify the development of an algorithm model. Besides, in implementing the logistic regression algorithm, further validation is also required. We recommend that you validate the data and the results of using the algorithm, more than twice. It is suggested that a better level of accuracy is obtained so that the analysis and evaluation of the effects of implementing the logistic regression model can be.

References

[1] A. Felman, Cardiovascular Disease: Types, Symptoms, Prevention, and Causes. The Technical Writer's Handbook. Mill Valley, CA: 1989.

[2] S. Palaniappan, and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE/AAACS International Conference on Computer Systems and Application, Doha, pp.108-115, 20008.

[3] M. Shouman, T. Turner, and R. Stocker, "Integrating Decision Tree and K-Means Clustering With Different Initial Centroid Selection Methods in The Diagnosis of Heart Disease Patients", Proceedings of the International Conference on Data Mining (DMIN). The Steering

Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldCom), 2012.

[4] E. Y. Boateng, and D. A. Abaye, “A Review of the Logistic Regression Model with Emphasis on Medical Research”. Journal of Data Analysis and Information Processing, Vol.7. No.4, pp.190-207, 2019.

[5] J. Harrell and E. Frank, “Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis”. Springer, 2015.

[6] Tyagi, Shivani, and S. Mittal,, "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning." Proceedings of ICRIC 2019. Springer, Cham, pp.209-221, 2019.

[7] Y. Khourdifi, and M. Bahaj, K-Nearest Neighbour Model Optimized by Particle Swarm Optimization and Ant Colony Optimization for Heart Disease Classification. In: Farhaoui Y., Moussaid L. (eds) Big Data and Smart Digital Environment. ICBDSDE 2018. Studies in Big Data, vol 53. Springer, Cham.