ITC 2/50
Information Technology and Control
Vol. 50 / No. 2 / 2021
pp. 308-318
DOI 10.5755/j01.itc.50.2.27672

A Predictive Model for Heart Disease Diagnosis Based on Multinomial Logistic Regression

Received 2020/09/15 | Accepted after revision 2021/05/18

http://dx.doi.org/10.5755/j01.itc.50.2.27672

HOW TO CITE: Munandar, T.A., Sumiati, S., Rosalina, V. (2021). A Predictive Model for Heart Disease Diagnosis Based on Multinomial Logistic Regression. *Information Technology and Control*, 50(2), 308-318. https://doi.org/10.5755/j01.itc.50.2.27672

A Predictive Model for Heart Disease Diagnosis Based on Multinomial Logistic Regression

Munandar Tb Ai, Sumiati Sumiati

Faculty of Information Technology; Informatics Engineering Department, Universitas Serang Raya, Indonesia, e-mails: tbaimunandar@gmail.com; sumiati@yahoo.com

Vidila Rosalina

Faculty of Information Technology; Computer Engineering Department, Universitas Serang Raya, Indonesia, e-mail: vidila.suhendarsah@gmail.com

Corresponding author: tbaimunandar@gmail.com

Many computational approaches are used to assist the analysis of influencing factors, as well as for the need for prediction and even classification of certain types of disease. In the case of disease classification, the data used are often categorical data, both for dependent variables and for independent variables, which are the results of conversion from numeric data. In other words, the data used are already unnatural. Conversion processes often do not have standard rules, thus affecting the accuracy of the classification results. This research was conducted to form a predictive model for heart disease diagnosis based on the natural data from the patients' medical records, using the multinomial logistic regression approach. The medical record data were taken based on the patients' electrocardiogram information whose data had been cleansed first. Other models were also tested to see the accuracy of the heart disease diagnosis against the same data. The results showed that multinomial logistic regression had the highest level of accuracy compared to other computational techniques, amounting to 75.60%. The highest level of accuracy is obtained by involving all variables (based on the results of the first experiment). This research also produced seven regression equations to predict the heart disease diagnosis based on the patients' electrocardiogram data.

KEYWORDS: computational technique, heart disease, multinomial logistic regression, prediction, diagnosis.

1. Introduction

Computational heart disease types diagnosis can indeed be done by various approaches, one of which is using the decision tree classification technique. Some of them are to identify factors that influence heart disease [2], early detection of patient heart disease [11], [6], also classification and prediction of heart disease [10], [20]. In many cases, implementing the decision tree technique requires training data in the form of categorical data when forming a decision tree. It is very rare for the decision tree technique to use pure numerical data because, in fact, a number of numerical data used at the time of forming a decision tree will be converted into a categorical form before calculating the entropy value and gain information to find the roots, nodes, and leaves. Forming a decision tree does not naturally use the data just what they are, yet it is based on the conversion of the original data into a categorical form. From the researchers' point of view, the process of converting numerical data into a categorical form may have weaknesses, resulting in the decision tree that does not always produce a very good accuracy. One of the weaknesses in the process of converting numeric data into categorical data is that there is no standard rule; it can even be very subjective. Of course, this becomes very problematic and influential if the data used is numerical data that is obtained directly, either from sensors or other devices that directly produce data through certain recording systems, for example, an electrocardiogram machine.

In a real case of a conventional heart disease diagnosis, a specialist doctor reads the patient's medical record data based on the results of a numerical electrocardiogram. In the practice, without realizing it, the conversion process of numeric values into certain categorical forms indeed occurs when a specialist doctor interprets the numerical data from the electrocardiogram results, until finally it refers to a diagnosis conclusion that is considered the most relevant. This process of converting numeric into categorical has a little problem if there are values between the maximum and minimum intervals that should be categorized into which class. In the end, it is determined on subjective decisions. Therefore, it is very important to present scientific studies related to the use of computational techniques to be able to diagnose a disease type based on natural data so that the results obtained are based on the data just the way they are. The regression approach in computation is not something new, however, in the world of health, the results of this study will make a significant contribution to the diagnosis of disease, particularly heart disease. Not only based on the subjectivity and experience of doctors, but by looking at the patterns and relationships between existing medical record data and then entered into a mathematical formulation. The results, of course, can be combined between experience, subjectivity and computational results to produce a better diagnosis.

This research was conducted to predict the heart disease diagnosis based on the existing data naturally without going through the conversion of numerical data into categorical data. Therefore, the prediction of the heart disease diagnosis can be more objective since the data were not naturally done with any conversion. The approach used in this research was the multinomial logistic regression (MLR). MLR allows data processing just the way they are without going through the data type conversion process. MLR also has the ability to form regression models with more than one target class, meaning that it is not only the class with binary values. This is certainly different from other regression models that can only be used in two target classes, such as binary logistic regression. Some studies are using binary logistic regression for simplicity of analysis, such as to determine the nutritional status and stunting of children [17], a factor analysis wait for public transport [4], the determination of the causes of traffic accidents [3], and the prediction of their rapidly developing incidence of liver disease [1]. But unfortunately it does not accommodate more than two target classes. Moreover, the MLR regression model that is formed can be directly tested using the numerical data from the patients' medical records. In this research, the MLR results were also compared with several decision tree classification techniques to see the diagnosis accuracy level produced

2. Multinomial Logistic Regression

Multinomial logistic regression is a logistic regression approach that allows finding relationships between variables, both independent and dependent, where the values of the dependent variable are categorical data with more than two types of categories

[13]. Usually the dependent variable is the nominal scale, and the independent variables are categorical and/or continuous data. There are several purposes for using this method: the first is to determine the probability of whether or not the data fits into certain categories according to the existing categories in the independent variables, of course, by paying attention to the significance of the independent variables; the second is to see the characteristics between groups in the dependent variables based on the distribution of data owned by the independent variables; the third is to determine what factors or variables affect the data group in the independent variables and in the dependent variables [19]. The multinomial logistic regression model is as shown in Equation (1)

$$\ln\left(\frac{\pi(x)_j}{\pi(x)_q}\right)=\beta_0^{(j)}+\sum_{i=1}^{k}\beta_i^{(j)}\cdot x_i,\ j=1,\ldots,q-1, \quad (1)$$

where $\pi(x)$ is the probability of a category vector of the dependent variables, β_0 is the constant value of the equation, β_i is the ith coefficient value of the independent variable x, and xi is the ith independent variable used in the statistical analysis. The number of multinomial equations formed is usually q – 1 category of the dependent variables. As for calculating the probability of data entering a certain category of the dependent variables, it can be calculated by Equation (2)

$$\pi_{(x)}=\frac{e^{\beta_0^{(j)}+\sum_{i=1}^{k}\beta_i^{(j)}\cdot x_i}}{1+e^{\beta_0^{(j)}+\sum_{i=1}^{k}\beta_i^{(j)}\cdot x_i}}. \quad (2)$$

3. Multinomial Logistic Regression in the World of Health

Multinomial logistic regression has been widely used in various branches of science. Some of them are to solve the problem of lending risk assessment [14], applied to the world of tourism [7] as well as to the world of trade to understand the factors that affect the product purchases by customers [12]. In the world of health, the implementation of MLR has also been widely applied, especially in relation to the identification of factors that affect skin allergies [15], investigations into the effect of oral contraceptives and hormone replacement therapy on the risk of lung cancer in female patients [18], predictions of pregnancy and miscarriage based on assessment of the quality of oocytes and embryos [13], analyses of factors that influence excess and underweight during pregnancy [16], choices of contraceptives in women to prevent pregnancy [5], as well as other health fields such as cardiology [9] and genetics [8].

4. Research Method

This study began by collecting the historical data on the patients with heart disease at a cardiac clinic of a government hospital in Banten Province. There were 324 historical data on the patients with heart disease used in this research. Consisting of 16 variables divided into 15 independent variables (obtained from the results of the patients' ECHO and ECG medical records), and one dependent variable in the form of heart disease type category, diagnosed by the doctors. The data used had been cleansed previously from as many as 350 data collected, by sorting and separating the irrelevant data to use. The next step was to conduct a simultaneous test of the data used to see the validity. A partial test was then carried out for each independent variable, consisting of four test experimental conditions. The first experiment was carried out for all data variables, the second experiment was carried out on variables that were only significant in the first experiment, the third experiment was carried out on variables that were stated insignificant based on the first experiment, and the fourth experiment was carried out by combining variables that were stated significant in the second and third experiments. The next stage was to test the model goodness and the accuracy level of the prediction results of the heart disease type classification for each experimental condition, then choosing the best model that would be used later. The Multinomial logistic regression method was used in this research.

5. Results and Discussion

5.1. Test Results with 15 Complete Variables

In this research, several test parameters were determined before the experiment began. Several parameters were determined, among others, the confidence

interval value and the alpha value were 95% and 5% or 0.05, respectively. The determination of this parameter was useful in testing the significance of the independent variables on the dependent. The statistical test results on the research data showed that the data were very suitable to be used (valid), and there was no missing data. This could be seen through the simultaneous test results by seeing at the significance value (sig) of the likelihood ratio test that was below the alpha value.

The partial test was done to find out how significant each independent variable affected the dependent variable. The test results showed that, of the existing fifteen independent variables, there were five variables that significantly affected the dependent variable. This could be seen from the significance value (sig) that exceeded the alpha value, and the calculated chi-square value was greater than the chi-square table value. The five independent variables were: left atrium, heart's functions, EDD (End-diastolic Diameter), ESD (End-systolic Diameter) and PW (posterior wall) Diastole. The five variables had sig values, respectively, 0.000, 0.000, 0.016, 0.005 and 0.002. Table 1 shows the chi-square and Sig values for each of the independent variables tested.

Table 1
Significance test results of fifteen independent variables

Independent Var	Likelihood Ratio Tests		
	Chi-Square	df	Sig.
AORTA	4,215	7	,755
LEFT-ATRIUM	68,604	7	,000
HEART'S FUNCTIONS	49,239	7	,000
EDD	17,299	7	,016
ESD	20,208	7	,005
IVSDiastole	11,764	7	,109
IVSSystole	3,220	7	,864
PWDiastole	23,057	7	,002
PWSystole	11,791	7	,108
HR	2,386	7	,935
PRPQ	7,125	7	,416
QRS	8,303	7	,307
QT	12,831	7	,076
QTC	6,173	7	,520
P	7,570	7	,372

To test the influential significance of these five variables, the data analysis was carried out a second time by removing ten other independent variables to see whether the significance was still the same or changed.

5.2. The Results After 10 Insignificant Independent Variables Were Excluded (Only Five Variables)

The second experiment was conducted to test five independent variables which were considered significant to the dependent variable as a result of the first experiment. The partial analysis results on the influential significance of the independent variables LEFT-ATRIUM, HEART'S FUNCTIONS, EDD, ESD and PWDiastole showed that two variables such as EDD and ESD had the calculated chi-square values below the values of the chi-square table, with the sig values above alpha. The Chi-Square values for the two variables were 6,946 and 10,027, respectively, while the Sig values were 0.435 and 0.187, respectively. Thus, of the five independent variables that were separated in the first experiment, there were three independent variables that significantly influenced the determination of the patients' heart disease. The three variables were LEFT-ATRIUM, HEART'S FUNCTIONS, and PWDiastole. These three variables were considered to have a significant effect on the determination of the heart disease type (dependent variable) because they had the Sig values below alpha and the calculated Chi-Square values above the values of the Chi-Square Table. Table 2 shows the Chi-Square and Sig values of the five variables in the second experiment.

Table 2
Significance test results of five independent variables3

Independent Var	Likelihood Ratio Tests		
	Chi-Square	df	Sig.
LEFT-ATRIUM	81,148	7	,000
HEART'S FUNCTIONS	53,253	7	,000
EDD	6,946	7	,435
ESD	10,027	7	,187
PWDiastole	75,235	7	,000

5.3. The Results for 10 Insignificant Variables that Were Retested

The third experiment was conducted to retest the independent variables that were excluded after the first experiment. Even though in the first experiment, the ten insignificant variables were said to have no significance to the dependent variable, retesting was still carried out to see whether it affected the dependent variable significantly after being separated, or not at all. With the same standard parameters in the first experiment, the test results showed that of the ten independent variables: AORTA, IVSDiastole, IVSSystole, PWSystole, HR, PRPQ, QRS, QT, QTC, and P; there were five variables that did not influence significantly on determining the heart disease type, while the rest significantly influenced on determining the heart disease type. The five variables that significantly influenced were: AORTA, IVSDiastole, PWSystole, and P. This could be seen from the calculated Chi-Square values that exceeded the Chi-Square Table values, as well as the Sig values of the AORTA, IVSDiastole, PWSystole, and P variables which had the Sig values more than the alpha value. Table 3 shows the results of the experiments on ten variables that were stated insignificant in the first experiment.

Table 3
Significance test results for 10 variables that are insignificant in the first experiment

Independent Var	Likelihood Ratio Tests		
	Chi-Square	df	Sig.
AORTA	14,144	7	,049
IVSDiastole	20,759	7	,004
IVSSystole	2,516	7	,926
PWSystole	54,483	7	,000
HR	1,511	7	,982
PRPQ	12,111	7	,097
QRS	7,683	7	,361
QT	6,215	7	,515
QTC	8,934	7	,257
P	20,229	7	,005

5.4. The Combination Between Independent Variable Values of the Separation Results of Five Variables and Ten Variables

The fourth experiment was combining each independent variable which had significance to the dependent variable in the results of the first and second experiments. This was done to see if there was a significant change in combining these variables in determining the heart disease type. By combining the independent variables in the first and second experiments would contain a combination of seven variables, namely AORTA, LEFT-ATRIUM, HEART'S FUNCTIONS, IVSDiastole, PWDiastole, PWSystole, and P. These seven variables were then tested to see their significance for determining the heart disease type (dependent variable).

The test results showed that there were two variables that did not significantly influence the determination of the heart disease type. The two variables were AORTA with a significant value of 0.596, PWSystole with a significant value of 0.303, and P with a sig value of 0.204. These three variables had the sig values above the specified alpha values, which were greater than 0.05. Table 4 shows the Sig and Chi-Square values of the combined variable experimental results for the second and third experiments.

To strengthen the result analysis of the three experiments conducted, a model goodness test was carried

Table 4
Significance test results for the combined variables of the second and third experiments

Independent Var	Likelihood Ratio Tests		
	Chi-Square	df	Sig.
AORTA	5,527	7	,596
LEFT-ATRIUM	68,714	7	,000
HEART'S FUNCTIONS	145,077	7	,000
IVSDiastole	15,501	7	,030
PWDiastole	27,543	7	,000
PWSystole	8,344	7	,303
P	9,740	7	,204

out for each experimental condition by looking at the coefficient of determination of the Pseudo R-Square. The experimental results showed that the first experiment had a Nagelkerke coefficient value of 0.821, or it could be said that the data diversity of the independent variables in the research was able to explain the data diversity in the dependent variables by 82.1%, while the rest could be explained by the variables outside the research model. In the second experiment, the Nagelkerke coefficient value was obtained at 0.753, or the data diversity in the independent variables on the dependent was 75.3%, while the rest was obtained from outside the research model. The third experiment with ten independent variables as the variables separation results from the first experiment showed a significant decrease in the value of the Nagelkerke coefficient, namely 0.443, or 44.3% of the data diversity in the dependent variables was influenced by the independent variables. Whereas in the last experiment which combined variables that were considered significant in the second and third experiments, the Nagelkerke coefficient value was 0.753, or as much as 75.3% of the data diversity of the dependent variables was influenced by the independent variables, the rest was influenced by the data of the models outside this research.

The next analysis was to look at the accuracy of the prediction results of the heart disease type classification based on the logistic regression model, formed from each experiment. In the first experiment, the accuracy rate of the prediction results was 75.6%, while in the second experiment it decreased to 73.5%. A very significant decrease occurred for the predictive accuracy of the disease type classification in the third experiment which was 57.7%, while in the last experiment the accuracy rate of the prediction results was 71.6%. Table 5 shows the comparison of the model goodness based on the Nagelkarke coefficient, and also the accuracy level of the prediction results for all experiments.

Based on Table 5, the best classification prediction accuracy is obtained in the first experiment. The issuance of insignificant independent variables based on the first experiment did not significantly affect the formed model and the accuracy level of the classification results (see the results of the second experiment in Table 5). In fact, after removing ten other independent variables that were considered insignificant in the first experiment, the accuracy rate decreased by 2.1%. Interestingly, even though the ten independent variables were stated insignificant to the dependent on the first experiment, it turned out that after a re-experiment was carried out on the ten independent variables, they still had variables that also had a significant influence on the dependent. However, both in terms of the model and the accuracy of the classification results did not show the better values than the first and second experiments; in fact, it even worse than the results of other experiments. The last experiment was a combination of independent variables which were stated significant in the first experiment, plus the independent variables that were insignificant in the first experiment yet were significant to the dependent on the third experiment. Therefore, a combination of seven independent variables was obtained which was retested to see the significance on the dependent variables. The test results showed that the model formed had the same value as the second experiment (Nagelakerke coefficient of 0.753), but had a lower level of classification accuracy of 71.6%.

By seeing at the Nagelkerke coefficient and the accuracy level of classification prediction, this research took a model that was formed based on the results of the first experiment, and then converted it based on Equation (1). The multinomial logistic regression

Table 5
Comparison of model goodness and prediction accuracy of all experiments

Experiment	Nagelkerke Coef.	Accuration (%)
Experiment 1 (All Variables)	0.821	75.6%
Experiment 2 (Five Independent Variables)	0.753	73.5%
Experiment 3 (Ten Independent Variables)	0.443	57.6%
Experiment 4 (Combined Variables Selected from Experiment 2 and Experiment 3)	0.753	71.6%

model was formed to determine the heart disease type classification as shown in the formulation below. This formed equation was based on the comparison category of Rheumatic Heart Disease as the preference category at the time of the statistical test.

For the Disease of:

1 Atrial Septal Defect

ln((D=1)/(D=8)) = 42.985 + 0.097(AORTA) + 0.138 (LEFTATRIUM) - 0.102 (HEARTFUNCTION) - 0.805 (EDD) + 0.257 (ESD) - 1.757 (IVSDiastole) + 0.953 (IVSSystole) - 1.607 (PWDiastole) - 0.619 (PWSystole) + 0.050 (HR) + 0.007 (PRPQ) + 0.032 (QRS) + 0.018 (QT) - 0.012 (QTC) - 0.009 (P)

2 Cardial Improvement

ln((D=2)/(D=8))= 124.199 - 0.025(AORTA) - 0.410 (LEFTATRIUM) + 0.000 (HEARTFUNCTION) - 4.173 (EDD) + 2.593 (ESD) - 1.097 (IVSDiastole) - 0.781 (IVSSystole) - 7.532 (PWDiastole) - 1.785 (PWSystole) + 0.344 (HR) - 0.278 (PRPQ) + 0.477 (QRS) - 0.247 (QT) + 0.135 (QTC) + 0.434 (P)

3 Coronary Artery Disease

ln((D=3)/(D=8))= -3.561 - 0.021(AORTA) - 0.251 (LEFTATRIUM) + 0.056 (HEARTFUNCTION) - 0.166 (EDD) + 0.490 (ESD) + 0.199 (IVSDiastole) + 0.242 (IVSSystole) - 0.121 (PWDiastole) - 0.016 (PWSystole) + 0.016 (HR) - 0.009 (PRPQ) - 0.010 (QRS) + 0.008 (QT) - 0.004 (QTC) + 0.001 (P)

4 Diastolic Dysfunction

ln((D=4)/(D=8))= 37.536 - 0.256(AORTA) - 0.612 (LEFTATRIUM) - 0.087 (HEARTFUNCTION) + 0.369 (EDD) - 0.364 (ESD) - 0.290 (IVSDiastole) - 0.130 (IVSSystole) - 0.925 (PWDiastole) + 0.646 (PWSystole) - 0.002 (HR) - 0.038 (PRPQ) + 0.017 (QRS) + 0.016 (QT) - 0.020 (QTC) + 0.005 (P)

5 Hypertensive Heart Disease

ln((D=5)/(D=8))= -9.424 - 0.056(AORTA) - 0.340 (LEFTATRIUM) + 0.204 (HEARTFUNCTION) - 0.208 (EDD) + 0.492 (ESD) + 0.366 (IVSDiastole) + 0.132 (IVSSystole) + 0.136 (PWDiastole) + 0.118 (PWSystole) + 0.011 (HR) - 0.010 (PRPQ) - 0.006 (QRS) + 0.010 (QT) - 0.008 (QTC) + 0.004 (P)

6 Left Ventricular Hypertrophy Suspect Hypertensive Heart Disease

ln((D=6)/(D=8))= -5.482 - 0.089(AORTA) - 0.427 (LEFTATRIUM) + 0.195 (HEARTFUNCTION) - 0.204 (EDD) + 0.503 (ESD) + 0.089 (IVSDiastole) + 0.158 (IVSSystole) + 0.375 (PWDiastole) - 0.041 (PWSystole) + 0.016 (HR) - 0.008 (PRPQ) - 0.006 (QRS) + 0.013 (QT) - 0.008 (QTC) - 0.001 (P)

7 Normal Resting Echocardiography

ln((D=7)/(D=8))= 12.428 + 0.086(AORTA) - 0.448 (LEFTATRIUM) + 0.142 (HEARTFUNCTION) + 0.270 (EDD) - 0.094 (ESD) - 0.103 (IVSDiastole) + 0.100 (IVSSystole) - 0.781 (PWDiastole) - 0.343 (PWSystole) + 0.004 (HR) - 0.026 (PRPQ) - 0.025 (QRS) + 0.040 (QT) - 0.030 (QTC) + 0.021 (P)

To discover to what extent the formed model could be implemented properly, a classification test for heart disease types was carried out by taking a random sample of the ten patients' data, and then calculating the indicated probability of the heart disease types which was in accordance with the resulting multinomial regression model. The probability calculation is carried out using Equation (2), and produces a table for the heart disease types identification based on the probability values as shown in Table 6.

The test results of the diagnostic model for heart disease types on ten patients showed that most of the results of the diagnosis showed the same results between the doctors' diagnosis (based on the real data) and the

Table 6
Probability values of heart disease types in ten patients

Types of Heart Disease	Patient (% Probability)									
	1	2	3	4	5	6	7	8	9	10
Atrial Septal Defect	0,001%	0,00%	0,03%	96,39%	0,000%	0,00%	0,00%	1,99%	99,97%	0,00%
Cardial Improvement	0,00%	0,00%	0,00%	0,00%	0,000%	0,00%	0,00%	0,00%	0,00%	0,00%
Coronary Artery Disease	84,75%	96,09%	59,53%	96,13%	99,990%	99,96%	86,51%	71,57%	58,03%	99,92%
Diastolic Dysfunction	0,321%	0,277%	85,04%	99,95%	21,568%	55,26%	13,11%	78,97%	99,76%	0,03%
Hypertensive Heart Disease	93,98%	18,88%	95,27%	99,19%	99,995%	99,97%	89,97%	87,25%	96,88%	73,38%
Left Ventricular Hypertrophy Suspect Hypertensive Heart Disease	79,33%	18,27%	89,13%	99,35%	99,987%	99,96%	66,08%	81,05%	98,50%	70,04%
Normal Resting Echocardiography	0,091%	2,216%	98,05%	99,53%	90,970%	95,66%	49,58%	99,94%	99,87%	0,85%
Rheumatic Heart Disease	0,00%	0,00%	0,000%	0,000%	0,000%	0,00%	0,00%	0,00%	0,00%	0,00%

test of the formed model. There were only two patients whose prediction results were not the same as the real situation. There was something interesting about the probability calculation results against the ten patients that had been done. Some patients had more than one chance of being indicated by the disease type, seen from the opportunity percentage based on their medical record data. Patient number five, for example, both with the factual results of the doctors' diagnosis and the regression model test, the patient suffered from heart disease with the type of Hypertensive Heart Disease. However, on the other hand, if we take a look at the probability percentage in Table 7, the possibility of the patient suffering from other heart disease types is also there, and very possible. For example, patient number five also had a chance of suffering from coronary artery disease, where the probability percentage was only slightly different from 0.01 to the type of Hypertensive Heart Disease.

The sixth patient was the opposite of the fifth patient, in which the doctors' diagnosis showed that the sixth patient was indicated to have coronary artery disease; while based on the prediction model, the sixth patient was diagnosed with Hypertensive Heart Disease. The

Table 7
Comparison of doctors' diagnosis results and predictive model results on samples of 10 patients

Patient	Doctor's Diagnosis	Model Prediction
1	Hypertensive Heart Disease	Hypertensive Heart Disease
2	Coronary Artery Disease	Coronary Artery Disease
3	Normal Resting Echocardiography	Normal Resting Echocardiography
4	Diastolic Dysfunction	Diastolic Dysfunction
5	Hypertensive Heart Disease	Hypertensive Heart Disease
6	Coronary Artery Disease	Hypertensive Heart Disease
7	Left Ventricular Hypertrophy Suspect Hypertensive Heart Disease	Hypertensive Heart Disease
8	Normal Resting Echocardiography	Normal Resting Echocardiography
9	Atrial Septal Defect	Atrial Septal Defect
10	Coronary Artery Disease	Coronary Artery Disease

difference in the probability percentage of the predictive model calculation between Hypertensive Heart Disease and Coronary Artery Disease for the sixth patient was also very small, which was 0.01. The data visualization of the prediction results of the patients' heart disease types using the multinomial logistic regression model is more complete as shown in Table 1. While the comparison between the disease types suffered by the ten patients, diagnosed by doctors and by the predictive model calculation, is shown in Table 7.

Based on the data in Table 7, it can be seen that almost 80% of the prediction results from the multinomial logistic regression model have the same diagnosis of heart disease. Of the ten patients sampled, there were two patients who differed between the doctors' diagnosis and the predictive model results, namely the sixth and seventh patients. Our suspicions and assumptions regarding the difference in the diagnosis of heart disease from two patients between doctors with the predictive model developed are: first, doctors usually diagnose based on direct observation by reading the results of the patient's electrocardiogram medical record without making direct comparisons with previous cases that may have similarities. ; second, the diagnostic results with the model is done by extracting the pattern of patient data and compare it with previous cases then with computational approaches are calculated so as to produce a different diagnosis by a doctor. Nevertheless, these findings certainly be triger to conduct more in-depth advanced research include investigating the factors that influence these differences. However, this study at least have an alternative in the medical world to diagnose heart disease using computational approaches. To reassure the accuracy of the prediction results of the formed regression model, the research data were tested against six classification model approaches to see which one had the best accuracy level among the models offered. The six models were: Naive Bayes, Algorithm J48, Random Forest Tree, REPTree Algorithm, Random Tree, and Hoeffding Tree. From the test results conducted, it turned out that the multinomial logistic regression approach still had the best accuracy level among the other six approaches, which was 75.60%. The comparison of accuracy between multinomial logistic regression and other classification approaches is shown in Table 8.

Table 8

Comparison of accuracy level between multinomial logistic regression and other models

No	Model	Accury (%)
1.	Multinomial Logistic Regression	75.60%
2.	Naive Bayes	61.42 %
3.	Algorithm J48	64.81 %
4.	Random Forest	72.84 %
5.	REPTree Algorithm	68.83 %
6.	Random Tree	60.80 %
7.	HoeffdingTree	62.65 %

6. Conclusion

According to the research conducted, it can be seen that the multinomial logistic regression model can be used to predict the classification of a patient's heart disease based on the medical electrocardiogram data. The comparison results between the accuracy level and other approaches show that the multinomial logistic regression model has the best accuracy level among other comparative approaches, which is 75.60%. Based on the four-time-conducted experiments related to the influential significance of the independent variables on the dependent, it can be said in this research that the reduction of the independent variables which is insignificant to the dependent does not affect the accuracy level of the formed predictive model. Nevertheless, by combining the variables that are considered significant in the several experiments conducted, it does not necessarily make the model very accurate. Still, the model formed in the first experiment without eliminating certain variables, can be used as a predictive model for the heart disease identification; even though some independent variables are stated insignificant, somehow, the accuracy level is still much better than the results of other conducted experiments.

Acknowledgement

Thank you to the Ministry of Research, Technology and Higher Education of Republic of Indonesia for funding this basic research scheme through the Directorate General of Research and Development Strengthening of RISTEK DIKTI 2020.

References

1. Abdalrada, A. S., Yahya, O. H., Alaidi, A. H. M., Hussein, N. A., Alrikabi, H. T. H., Al-Quraishi, T. A Predictive Model for Liver Disease Progression Based on Logistic Regression Algorithm. Periodicals of Engineering and Natural Sciences, 2019, 7(3), 1255-1264. https://doi.org/10.21533/pen.v7i3.667
2. Abdar, M. Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease. Carpathian Journal of Electronic and Computer Engineering, 2015, 8(2), 31-36. http://cjece.ubm.ro/vol/8-2015/n2/1512.21-8207.pdf
3. Ahmed, L. A. Using Logistic Regression in Determining the Effective Variables in Traffic Accidents. Applied Mathematical Sciences, 2017, 11(42), 2047-2058. https://doi.org/10.12988/ams.2017.75179
4. Al Doori, A. Waiting Time Factor in Public Transport by Binary Logistic Regression. Australian Journal of Basic and Applied Sciences, 2017, 11(4), 72-76. http://www.ajbasweb.com/old/ajbas/2017/March/72-76.pdf
5. Al-balushi, M. S., Ahmed, M. S., Mazharul Islam, M., Khan, M. H. R. Contraceptive Method Choices Among Women In Oman: A Multilevel Analysis. Journal of Data Science, 2016, 14, 117-132. https://doi.org/10.6339/JDS.201601_14(1).0007
6. Aziz, A., Rehman, A. U. Detection of Cardiac Disease Using Data Mining Classification Techniques. (IJACSA) International Journal of Advanced Computer Science and Applications, 2017, 8(7), 256-259. https://doi.org/10.14569/IJACSA.2017.080734
7. Can, V. V. Estimation of Travel Mode Choice for Domestic Tourists to Nha Trang Using the Multinomial Probit Model. Transportation Research Part A: Policy and Practice, 2013, 49, 149-159. https://doi.org/10.1016/j.tra.2013.01.025
8. Cleynen, I., Boucher, G., Jostins, L., Schumm, L. P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J. M., Annese, V., Brand, S., Brant, S. R., Cho, J. H., Daly, M. J., Dubinsky, M., Duerr, R. H., Ferguson, L. R., Franke, A., Gearry, R. B., Goyette, P., Hakonarson, H., Lees, C. W. Inherited Determinants of Crohn's Disease and Ulcerative colitis Phenotypes: A Genetic Association Study. Lancet (London, England), 2016, 387(10014), 156-167. https://doi.org/10.1016/S0140-6736(15)00465-1
9. Dolansky, M. A., Xu, F., Zullo, M., Shishehbor, M., Moore, S. M, Rimm, A. A. Post-acute Care Services Received by Older Adults Following a Cardiac Event: A Population-Based Analysis. Journal of Cardiovascular Nursing, 2010, 25(4), 342-349. https://doi.org/10.1097/JCN.0b013e3181c9fbca
10. Joshi, A., Dangra, E. J., Rawat, M. K. A Decision Tree Based Classification Technique for Accurate Heart Disease Classification & Prediction. International Journal of Technology Research and Management, 2016, 3(11), 1-4. http://www.ijtrm.com/PublishedPaper/3Vol/Issue11/2016IJTRM1120167144-13d54ab7-e821-452e-b9ee-928ca52988462319.pdf
11. Karthiga, A. S., Mary, M. S., Yogasini, M. Early Prediction of Heart Disease Using Decision Tree Algorithm. International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 2017, 3(3), 1-17. https://www.ijarbest.com/journal/v3i3/969
12. Kohansal, M. R., Firoozzare, A. Applying Multinomial Logit Model for Determining Socio-Economic Factors Affecting Major Choice of Consumers in Food Purchasing: The Case of Mashhad. Journal of Agricultural Science and Technology, 2013, 15(7), 1307-1317. https://jast.modares.ac.ir/article-23-5984-en.pdf
13. Milewska, A. J., Jankowska, D., Wiesak, T., Acacio, B., Milewksi, R. The Application of Multinomial Logistic Regression Models for the Assessment of Parameters of Oocytes and Embryos Quality in Predicting Pregnancy and Miscarriage. Studies in Logic, Grammar and Rhetoric, 2017, 51(64). https://doi.org/10.1515/slgr-2017-0030
14. Miyamoto, M. Credit Risk Assessment for a Small Bank by Using a Multinomial Logistic Regression Model. International Journal of Finance and Accounting, 2014, 3(5), 327-334. DOI:10.5923/j.ijfa.20140305.07
15. Mukesi, M., Phillipus, I. N., Moyo, S. R., Mtambo O. P. L. Prevalence of Skin Allergies in Adolescents in Namibia. International Journal of Allergy Medications, 2017, 3(1). https://clinmedjournals.org/articles/ijam/international-journal-of-allergy-medications-ijam-3-022.pdf https://doi.org/10.23937/2572-3308.1510022
16. Neupane, B., McDonald, S. D., Beyene, J. Identifying Determinants and Estimating the Risk of Inadequate and Excess Gestational Weight Gain Using a Multinomial Logistic Regression Model. Open Access Medical Statistics, 2015, 5, 1-10. https://doi.org/10.2147/OAMS.S69707
17. Ohyver, M., Moniaga, J. V., Yunidwi, K. R., Setiawan, M. I. Logistic Regression and Growth Charts to Deter-

mine Children Nutritional and Stunting Status: A Review. 2nd International Conference on Computer Science and Computational Intelligence 2017. https://doi.org/10.1016/j.procs.2017.10.045

18. Pesatori, A. C., Carugno, M., Consonni, D., Hung, R. J., Papadoupolos, A., Landi, M. T., Brenner, H., Müller, H., Harris, C. C., Duell, E. J., Andrew, A. S., McLaughlin, J. R., Schwartz, A. G., Wenzlaff, A. S., Stucker, I. Hormone Use and Risk for Lung Cancer: A Pooled Analysis from the International Lung Cancer Consortium (ILCCO). British Journal of Cancer, 2013, 109(7), 1954-1964. https://doi.org/10.1038/bjc.2013.506
19. Venkatesan, G., Sasikala, V. A Statistical Analysis of Migration Using Logistic Regression Model. International Journal Of Scientific & Technology Research, 2019, 8(10), 1331-1336. http://www.ijstr.org/final-print/oct2019/A-Statistical-Analysis-Of-Migration-Using-Logistic-Regression-Model.pdf
20. Zriqat, I. M., Altamimi, A. M., Azzeh, M. A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. International Journal of Computer Science and Information Security (IJCSIS), 2016, 14(12), 868-879. https://arxiv.org/ftp/arxiv/papers/1704/1704.02799.pdf