Research on the heart attack prediction based on logistic regression

Siyi Wang*

School of Mathematics and Systems science, Xinjiang University, Urumqi, 830046, China

*Corresponding author: 20210803213@stu.xju.edu.cn

Abstract. Heart disease greatly threats human life and health. Through machine learning algorithms and models, computers can autonomously classify and predict data, thereby achieving analysis and prediction of unknown data. This research used Logistic Regression algorithm and 14 physical indicators from 302 patients to investigate heart attack. It found that because the correlation coefficient is greater than 0.4, the correlation that have to do with heart attack and what kind of chest pain it is, the maximum value of heart rate, whether have exercise induced angina and the ST depression was strong. The correlation have to do with heart attack and age, sex, number of main blood vessels, thalassemia is weaker. Correlation have to do with heart attack and cholestoral, the fasting blood sugar was the weakest. People who is older and men have more possibility to develop heart disease. The accuracy of the prediction is about 85.95%. The findings of this paper suggests that Logistic Regression algorithm will play a crucial role in preventing heart disease thus bringing better treatment outcomes to patients

Keywords: Logistic Regression algorithm, heart attack, dataset, decision tree

1. Introduction

Heart disease greatly treats human life and health. According to data, among cases of heart disease death, heart attack is the most common reason that lead to death. The occurrence of coronary heart disease, heart disease, cardiomyopathy and other cardiovascular diseases makes the overburdened heart worse, seriously threatening the health and life of human beings. According to epidemiological statistics, in 2018, cardiovascular death accounted for the first cause of death of urban and rural residents in China. The number of patients with cardiovascular disease is 330 million in China, including 11 million of stroke, nearly 11.4 million of coronary heart disease, pulmonary heart disease 5 millionn, almost 4.9 million of heart failure and atrial fibrillation, more than 2 million rheumatic heart disease, 2 million congenital heart disease and 245 million hypertensions.

For cardiovascular diseases, China's policy is to "focus on the grass-roots level, focus on prevention first", advocate the health management of the whole cardiovascular life cycle, and emphasize cardiovascular health in early life. Therefore, the prevention of heart disease is particularly important.

By reviewing previous studies about heart attack and using machine learning for heart attack prediction, there are studies that use K-Nearest Neighbour and Genetic Algorithm[1,2], to predict heart disease by using identification and data mining techniques to identify important features of heart disease[3], using decision tree (an efficient prediction system)to predict heart attack[4,5], and so on.

Researchers also found that Logistic Regression for prediction can get a good result [6-10]. Thus, this project focused on supporting the research and proving that Logistic Regression can play an important part in the prevention of heart disease. Also Logistic Regression can improve the prediction of clinical outcomes in the oncology field and explain the complex interactions between unassessed risk factors, helping to predict cardiac adverse events in high-risk populations [11].

2. Dataset

2.1. Introduction of the dataset

The database is called Heart Disease, it was published on UCI Machine Learning Repository in 1988. The data in this database comes from 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach. This database contains 76 attributes, a subset of 14 of them was used in this experiment.

This dataset included 14 physical indicators (Table1) for 302 patients aged between 29 and 77 years.

Age: People who is older have more possibility to develop heart disease.

Sex: Men have more possibility to develop heart disease than women.

Trtbps: Resting blood pressure, if people have higher blood pressure, the greater risk of having cardiovascular disease there will be.

Chol: Cholestoral obtained via BMI sensor, an important factor that induce heart disease.

Fbs: Fasting blood sugar, hyperglycemia can damage the blood vessels.

Thalachh: Maximum heart rate achieved, the higher the heart rate, the more likely heart disease.

Table 1. Indicators and definitions

physical indicators	Definitions
Age	Patients'age
Sex	Patients' sex
cp	What kind of chest pain it is Value1: angina that is typical Value 2: angina that is atypical Value 3: non-anginal pain Value 4: asymptomatic
trtbps	Tested blood pressure when at rest(mm/Hg)
chol	The amount of cholestoral (mg/dl)
fbs	The amount of fasting blood sugar(mg/dl) Value 1 :> 120 Value 0 :<=120
restecg	Electrocardiographic at rest Value 0: normal Value 1: abnormal ST-T wave abnormal Value 2: according to Estes' criteria, probably or definitely have left ventricular hypertrophy.
thalachh	Maximum value of the heart rate
exng	Exercise angina pectoris Value 1: have Value 0: not
oldpeak	ST depression
slp	value1:ST upsloping value2:ST flat value3: ST downsloping
caa	number of major vessels
thall	thalassemia
target	0=less chance of heart attack 1=more chance of heart attack

2.2. Descriptive statistics

Through the descriptive statistical analysis of the samples (Table 2), the total number of samples is 302. For the mean value, the age is 54.42, resting blood pressure is 131.60, cholestoral is 246.5, maximum heart rate achieved is 149.569536. For the maximum value, the age is 77, resting blood pressure is 200, cholestoralis 564, maximum heart rate achieved is 202. For the minimum value, the age is 29, resting blood pressure is 94, cholestoral is 126, maximum heart rate achieved is 71.

2.3. Distribution analysis

Most patients' age among 50 years old to 70 years old (Figure 1), more than 50% patients' resting blood pressure among 120~140 mm/hg(Figure 2),the distribution of the maximum heart rate achieved(Figure 3). the distribution of the st depression (Figure 4).

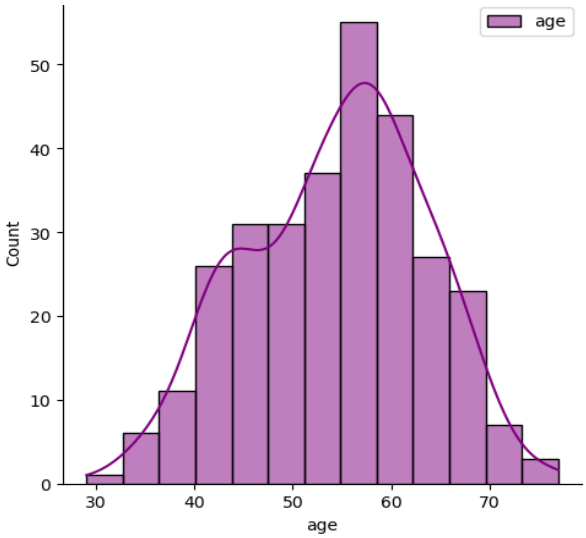

Fig. 1 Distribution of the age

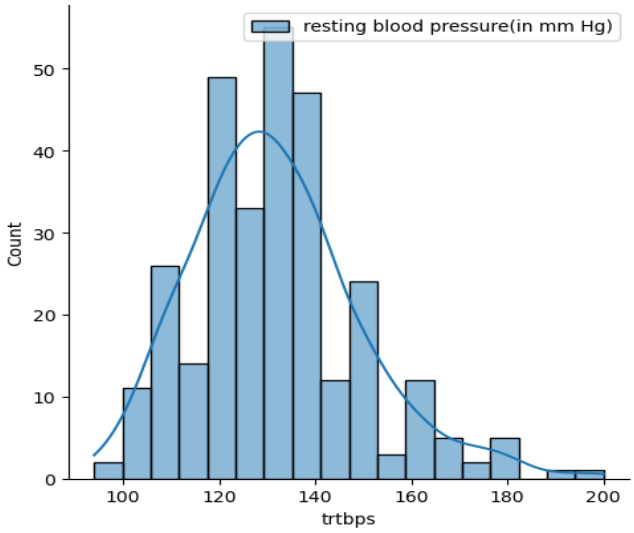

Fig. 2 Distribution of the resting blood pressure

Table 2. Descriptive statistical analysis

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
mean	54.42	0.68	0.96	131.60	246.50	0.15	0.53	149.57	0.33	1.04	1.40	0.72	2.31	0.54
std	9.05	0.47	1.03	17.56	51.75	0.36	0.53	22.90	0.47	1.16	0.62	1.01	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

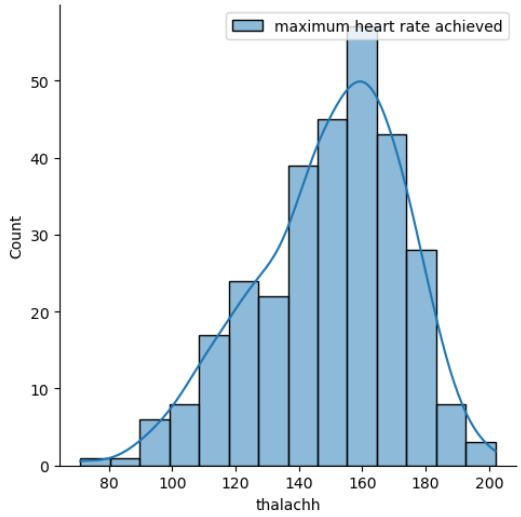

Fig. 3 Distribution of the maximum heart rate achieved

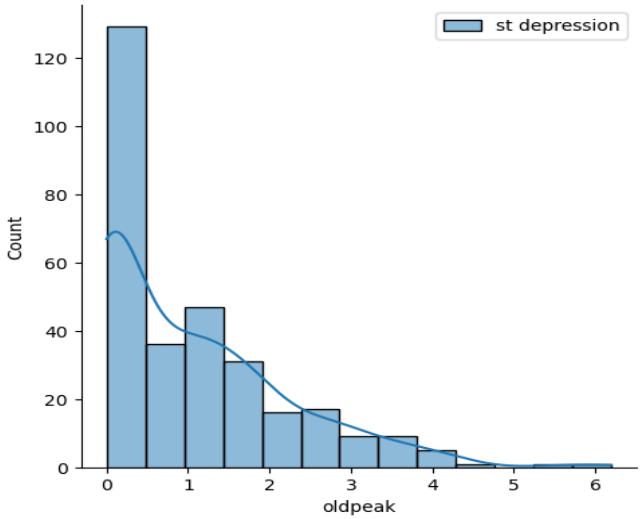

Fig. 4 Distribution of the st depression

2.4. Correlation analysis

Analyze the correlation and then draw the figures to make it more intuitive. From the Table 3, it's clear that age, sex, cp, resting blood pressure, restecg, thalachh, exng, oldpeak, slp, caa and thall have strong correlation with the heart attack, chol and fbs have weak correlation with the heart attack.

From the analysis, it shows that age, sex, trtbps, chol, fbs, exng, oldpeak, caa and thall were negatively associated with heart disease, others were positively associated with heart disease. It also indicates that because the correlation coefficient is greater than 0.4, the correlation that have to do with heart attack and what kind of chest pain it is, the maximum value of heart rate, whether have exercise induced angina and the ST depression was strong. The correlation has to do with heart attack and age, sex, number of main blood vessels,

thalassemia is weaker. Correlation have to do with heart attack and cholestoral, the fasting blood sugar was the weakest.

3. Method: logistic regression (LR)

Logistic Regression, also known as logistic Regression analysis, was proposed to solve the binary classification problem. It assumes the data obey the Bernoulli distribution and solves the optimal parameter with maximum likelihood estimation.

Table 3. Correlation with the heart attack

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.00	-0.09	-0.06	0.28	0.21	0.12	-0.11	-0.40	0.09	0.21	-0.16	0.30	0.07	-0.22
sex	-0.09	1.00	-0.05	-0.06	-0.20	0.05	-0.06	-0.05	0.14	0.10	-0.03	0.11	0.21	-0.28
cp	-0.06	-0.05	1.00	0.05	-0.07	0.10	0.04	0.29	-0.39	-0.15	0.12	-0.20	-0.16	0.43
trtbps	0.28	-0.06	0.05	1.00	0.13	0.18	-0.12	-0.05	0.07	0.19	-0.12	0.10	0.06	-0.15
chol	0.21	-0.20	-0.07	0.13	1.00	0.01	-0.15	-0.01	0.06	0.05	0.00	0.09	0.10	-0.08
fbs	0.12	0.05	0.10	0.18	0.01	1.00	-0.08	-0.01	0.02	0.00	-0.06	0.14	-0.03	-0.03
restecg	-0.11	-0.06	0.04	-0.12	-0.15	-0.08	1.00	0.04	-0.07	-0.06	0.09	-0.08	-0.01	0.13
thalachh	-0.40	-0.05	0.29	-0.05	-0.01	-0.01	0.04	1.00	-0.38	-0.34	0.38	-0.23	-0.09	0.42
exng	0.09	0.14	-0.39	0.07	0.06	0.02	-0.07	-0.38	1.00	0.29	-0.26	0.13	0.21	-0.44
oldpeak	0.21	0.10	-0.15	0.19	0.05	0.00	-0.06	-0.34	0.29	1.00	-0.58	0.24	0.21	-0.43
slp	-0.16	-0.03	0.12	-0.12	0.00	-0.06	0.09	0.38	-0.26	-0.58	1.00	-0.09	-0.10	0.34
caa	0.30	0.11	-0.20	0.10	0.09	0.14	-0.08	-0.23	0.13	0.24	-0.09	1.00	0.16	-0.41
thall	0.07	0.21	-0.16	0.06	0.10	-0.03	-0.01	-0.09	0.21	0.21	-0.10	0.16	1.00	-0.34
output	-0.22	-0.28	0.43	-0.15	-0.08	-0.03	0.13	0.42	-0.44	-0.43	0.34	-0.41	-0.34	1.00

The conditional probability function of the binomial logistic regression is defined as follows:

$P\ （Y=1|x）=\frac{1}{1+e^{-(\omega^T \bullet x+b)}}$ here $x \in R^n$ is the input, $Y \in \{0,1\}$, ω (For the column vector) and b is the parameters to be learned, $\omega^T \bullet x$ deputy the internal product of ω and x.

Launch by definition:

$$P(Y=0|x)=1-P(Y=1|x)=1-\frac{1}{1+e^{-(\omega^T \bullet x+b)}}=\frac{1}{1+e^{(\omega^T \bullet x+b)}} \quad (1)$$

The odds of an event is the probability of the event equals to the probability of the event not occurring, the log odds refers to the logarithm of the odds.

So there are:

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\omega^T \bullet x+b \quad (2)$$

Suitable conditions of the Logistic Regression model:

1.Because the variable is a digital variable, a classification variable, or an event occurrence rate. However, the repeated count phenomenon index is not applied in logistic regression.

2.Both residual and causal variables must follow two kinds of distributions; two kinds of distributions solve the estimation and verification of equations, which correspond to the classification variables and are neither static nor least squares.

3.The factor is linearly correlated with the logical probability.

4.The relationship between the observation objects is independent of each other.

The essence: Make a linear regression of the log odds of positive and negative cases.

4. Prediction

Process the data and predict heart attack by Logistic Regression, then analyze the accurcy. The roportion of the test data to the sample data is 40%, set the random number to seed is 6. As shown in Table 4, the first 10 predicted values are: 1, 1, 0, 0, 0, 1, 0, 1, 1, 1. The first 10 actual values are 1, 1, 1, 0, 0, 1, 0, 1, 1, 1. The training data's accuracy is about 83.98%, and the test data's accuracy is about 85.95%. The roc-curve was showed in Figure 5.

Table 4. Predicted value comparison

number	predicted value	actual values
1	1	1
2	1	1
3	0	1
4	0	0
5	0	0
6	1	1
7	0	0
8	1	1
9	1	1
10	1	1

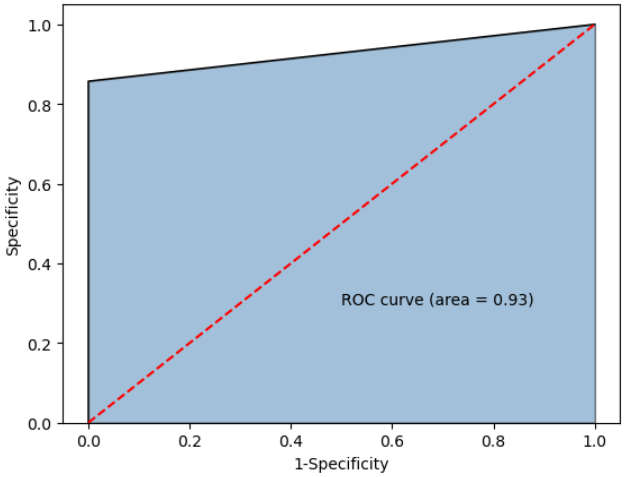

Fig. 5 Roc-curve

5. Conclusion

This study used Logistic Regression as a method to analyze the datasets on physical indicators of different populations to determine whether they have had a heart attack and predict the probability of one. The research shows that the accuracy of the prediction is about 85.95%. From the results of this research, it's clear that Logistic Regression has a high accuracy in the prediction of heart disease and can better integrate the data and accurately predict individual risks.

The study also shows that because the correlation coefficient is greater than 0.4, the correlation that have to do with heart attack and what kind of chest pain it is, the maximum value of heart rate, whether have exercise induced angina and the ST depression was strong. The correlation has to do with heart attack and age, sex, number of main blood vessels, thalassemia is weaker. Correlation have to do with heart attack and cholestoral, the fasting blood sugar was the weakest. This will open up new directions for personalized treatment, unlimitedly bringing better treatment outcomes to patients.

The weaknesses of this study are that as too many factors affecting heart disease, the existing data indicators are not fully leading to the accuracy of the prediction. What's more the effect of each index on the incidence of heart disease is not well reflected.

Reference

[1] M. Akhil jabbar, B.L. Deekshatulu, P. Chandra, J. Procedia Tech, **10**, (2013)

[2] Jeen-Shing Wang, Che-Wei Lin, Ya-Ting C. Yang, Neurocomputing, **116**, (2013)

[3] M. S. Amin, Y. K. Chiam, K. D. Varathan, J. Tele and Info, **36**, (2019)

[4] Purushottam, K. Saxena and R. Sharma, *Efficient heart disease prediction system using decision tree*, International Conference on Computing, Communication & Automation, Greater Noida, India, (2015)

[5] B. Liu, P. Li, M. Y. Zhu, S. P. Huan. J. CN Health Stat, **35**, (2018)

[6] X. Zhang, J. Mod Info Tech, **7**, (2023)

[7] A. Payam, A. Hasan, P. Jalal, M. A. Mohammad. Iranian J. pub health, **45**, (2016)

[8] J. Chen. Chinese Practical Medicine, **17**, (2022).

[9] M. Carina, J. Euro Soc Rev, **26**, (2010)

[10] B. Melinda, A. Raul, V. J. S. Hans, L. J. I. Zimmermann, G. Diego, G. W. D. Antonio, J. Ped card, **42**, (2020)

[11] Heilbroner, S. Peter, et al. J. immu ca, **9**,10 (2021)