Explore the relationship among various factors: Analysis of logistic regression

Jiatong Zhang

No.1 High School of Liaohe Oil Field, Liaoning, 124010, China

zjtyxdyx@outlook.com

Abstract. Heart disease is a common condition among living beings, especially humans. It can lead to severe consequences, including death. Therefore, it is essential to take preventive measures. Early detection can significantly increase the survival rate and reduce suffering. This study aims to demonstrate the application of logistic regression using heart disease as a case study. First, the paper introduces heart disease, including its causes, risk groups, contributing factors, current status, and available treatments. Next, it presents the relevant variables and pre-diction methods used in the logistic regression model. The results are then discussed, showing clear relationships between the predictors and the presence of heart disease. In conclusion, the study uses visual charts and statistical analysis to illustrate the value of logistic regression in medical diagnostics. According to the findings, logistic regression can effectively support the identification of disease risks and contribute to early diagnosis efforts. Future research should focus on optimizing model performance, integrating additional variables, and combining logistic regression with other advanced machine learning techniques to enhance diagnostic accuracy and reliability further.

Keywords: Logistic regression, heart disease, morbidity rate.

1. Introduction

Heart disease is currently one of the leading threats to human health. Many individuals suffer from the pain and consequences associated with cardiovascular conditions. Although heart disease may seem daunting, proactive measures can be taken to reduce the risks and protect one's health. In this study, logistic regression is employed as a predictive tool to estimate the likelihood of heart disease in a specific population.

Medically, heart disease encompasses a variety of conditions affecting the heart's structure and function. As the core organ of the human body, the heart is responsible for pumping blood throughout the entire system, supplying oxygen and nutrients to sustain life. Any dysfunction in its operation can lead to serious health problems. Among the many forms of heart disease, coronary atherosclerotic heart disease (CHD) is particularly prevalent. CHD can result in myocardial ischemia, oxygen deficiency, and even tissue necrosis due to narrowing or blockage of the coronary arteries. It is a major contributor to the increasing mortality rate, emphasizing the need for regular exercise and healthy living to prevent its onset (Wang and Liang, 2023).

Elderly individuals are at higher risk of developing acquired heart disease (AHD), though young people are not immune. For adolescents, factors such as poor lifestyle habits, irregular sleep schedules, prolonged sedentary behavior, and excessive psychological stress may con-tribute to cardiovascular issues. These behaviors disrupt biological rhythms, slow metabolism, and interfere with hormonal balance. For older adults, reduced physical activity, unbalanced diets, and irregular routines further increase vulnerability to heart disease (Zhou et al., 2024).

Heart disease is rarely caused by a single factor. Instead, it results from a complex interplay of physical, psychological, social, and environmental influences. While some mechanisms-such as the inflammatory processes associated with atherosclerosis-remain unclear, mental stress has been identified as a significant contributor (Ke et al., 2015). Drug therapies play an essential role in managing heart disease by regulating indicators like blood sugar, cholesterol levels, body mass, and blood pressure. These health metrics are closely tied to a person’s lifestyle and environment (Li et al., 2017).

Cardiovascular disease remains a leading cause of death globally. For example, coronary disease alone is responsible for approximately one-third of total deaths, with similar prevalence in men and women (Lekha et al., 2017). Although advances in medical technology have delayed disease onset, the mortality rate continues to rise. In recent years, the number of AHD patients has grown substantially (Liu et al., 2019). According to the China Cardiovascular Disease Report (2018), over 11 million individuals in China are affected, with approximately 1.1 million deaths occurring each year (Chen et al., 2021).

To address this crisis, timely intervention and treatment are essential. Hospitals often rely on two main categories of drug prescriptions: those aimed at relieving symptoms and those that improve long-term outcomes (Sackner-Bernstein, 2005). For example, treatment for unstable angina may include nitrates, beta-blockers, calcium antagonists, and medications such as antiplatelet agents, ACE inhibitors, ARBs, and Statins. Each type of drug plays a different role in reducing cardiovascular risk, managing symptoms, and slowing disease progression. However, treatment must be tailored to individual patient needs (Chen et al., 2018).

In addition to medication, lifestyle modifications are critical. Medical professionals emphasize the importance of a balanced diet rich in vegetables, fruits, nuts, and lean proteins. Regular physical activity, incorporating both aerobic and resistance exercises, is also strongly recommended. Reducing sedentary behavior is particularly beneficial for both prevention and recovery (Wu et al., 2024).

Technological advances have also improved rehabilitation practices. Effective cardiac recovery and secondary prevention now depend on interdisciplinary cooperation, convenient medical devices, and innovations such as remote monitoring via the Internet of Things. Although China’s cardiac rehabilitation efforts have expanded rapidly in recent years, there re-mains significant room for development (Li et al., 2017).

2. Methods

2.1. Data source and description

The dataset used in this study was obtained from Kaggle and originates from a compilation conducted in 1988. It includes records from four sources: Cleveland, Hungary, Switzerland, and Long Beach V. While the original dataset comprises 76 attributes, most published studies have focused on a subset of 14 key attributes. Among these, one attribute labeled "target" is used to indicate whether a patient has been diagnosed with heart disease. A value of 0 represents the absence of disease, while a value of 1 signifies the presence of disease.

In total, the dataset includes data from 1190 individuals. Each record contains various physiological and demographic indicators relevant to heart disease, such as age, sex, chest pain type, cholesterol levels, fasting blood sugar, and maximum heart rate.

2.2. Variable description

Table 1 provides an overview of 10 selected variables used in this analysis. These variables were selected due to their documented relevance in the occurrence and diagnosis of heart dis-ease. Each one may influence or be associated with other variables, making it essential to analyze their interactions.

Table 1. Data selection table

Variable Name	Description	Value Range
Age	Age of the individual	25 to 80 years
Sex	Biological sex (0 = female, 1 = male)	0 or 1
Chest Pain Type	Type of chest pain experienced	1 to 4 (severity increases)
Cholesterol	Cholesterol level in mg/dL	0 to 603
Fasting Blood Sugar	Blood sugar after fasting (1 = >120 mg/dL)	0 or 1
Max Heart Rate	Maximum heart rate measured during exercise	60 to 202 bpm
Exercise-Induced Angina	Angina caused by exercise (1 = yes, 0 = no)	0 or 1
Target	Diagnosis result (1 = disease, 0 = no disease)	0 or 1

2.3. Model introduction

Given the study's objective-to predict the likelihood of heart disease-logistic regression was chosen as the primary analytical tool. This model is particularly suited for binary classification problems where the outcome variable has two categories, such as disease presence or absence.

In statistical terms, logistic regression estimates the probability of a specific event (in this case, the diagnosis of heart disease) based on a set of independent variables. Unlike linear regression, which predicts a continuous outcome, logistic regression outputs a value between 0 and 1, representing the probability of the target event. This method allows people to determine which variables have the most significant influence on the likelihood of developing heart dis-ease.

Correlation analysis was also conducted to assess the strength and direction of the relation-ships between variables. This helped identify which factors may serve as potential predictors in the logistic regression model.

3. Results and discussion

3.1. Descriptive statistics and distribution

Figure 1 presents the distribution of heart disease across different age groups. The analysis shows that individuals aged over 45 are significantly more likely to suffer from heart disease compared to younger groups. The incidence rate increases with age, highlighting age as a critical risk factor.

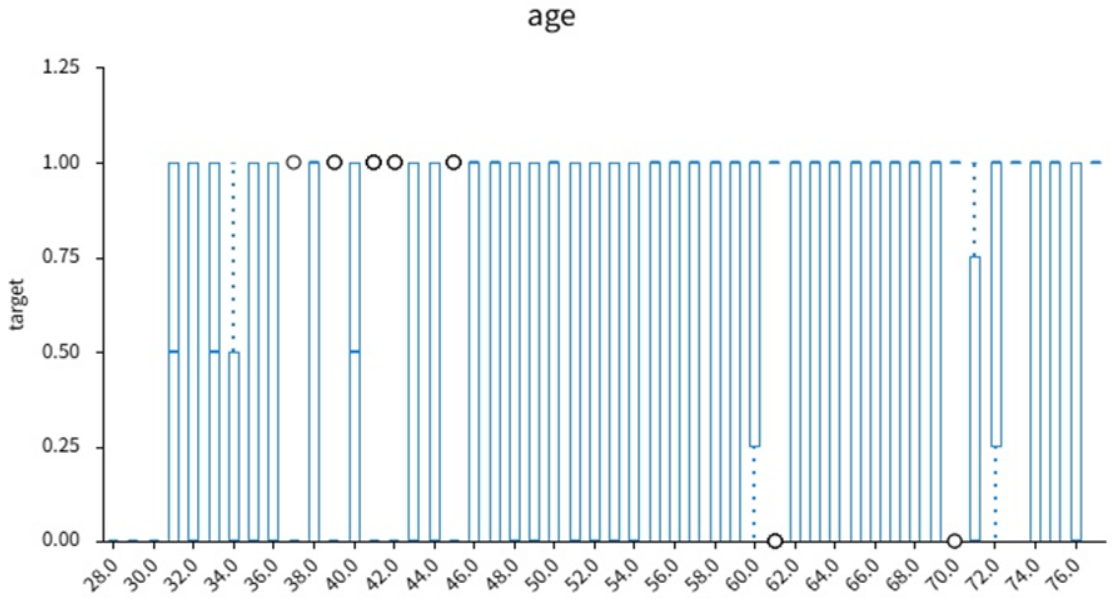

Figure 1. Target changes with age.

Alt Text for Figure 1: A bar plot showing the distribution of heart disease across different age groups.

Figure 2 shows the relationship between chest pain type and the presence of heart disease. Patients exhibiting type 1 (typical angina) and type 3 (non-anginal pain) chest pain are more likely to be diagnosed with heart disease. Their associated "target" values are closer to 1.0, indicating a higher probability of diagnosis. In contrast, chest pain types 2 and 4 are associated with much lower target values, suggesting a lower risk.

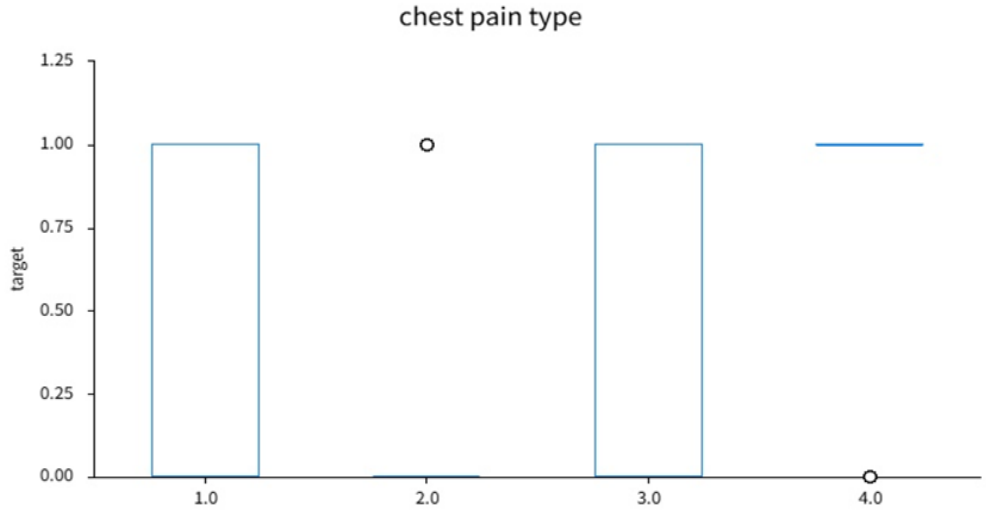

Figure 2. Target changes with chest pain type.

Alt Text for Figure 2: A bar plot showing the relationship between chest pain type and the presence of heart disease.

Figure 3 explores the association between maximum heart rate and heart disease. Individuals with a maximum heart rate of 110 bpm or higher are at greater risk, especially those reaching or exceeding 130 bpm. This suggests that elevated heart rates during physical exertion may serve as a warning signal for underlying cardiovascular conditions.

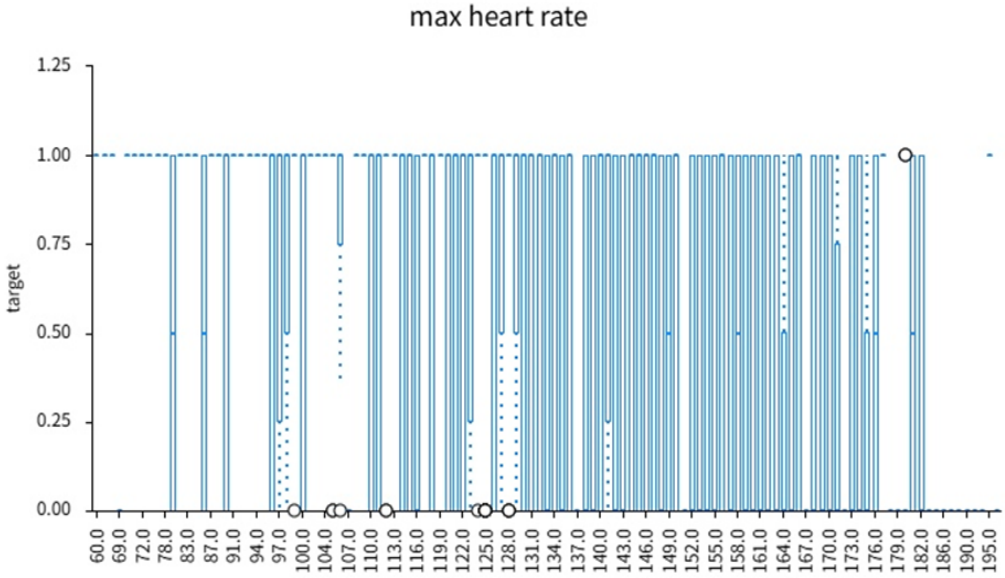

Figure 3. Target changes with max heart rate.

Alt Text for Figure 3: A bar plot showing the association between maximum heart rate and heart disease.

3.2. Correlation analysis

Figure 4 provides the correlation coefficients between age and other key variables. Several statistically significant relationships were found: Age and chest pain type: 0.16 (positive correlation).

Age and fasting blood sugar: 0.18 (positive correlation). Age and maximum heart rate: -0.35 (negative correlation). Age and exercise-induced angina: 0.19 (positive correlation). Age and heart disease diagnosis (target): 0.27 (positive correlation).

These findings suggest that as individuals age, they are more likely to develop higher blood sugar levels, experience more severe chest pain, and suffer from exercise-induced angina, all of which are associated with an increased likelihood of heart disease. Interestingly, maximum heart rate tends to decrease with age, which may also play a role in risk prediction.

Other variables, such as age and sex, or cholesterol and sex, did not demonstrate statistically significant correlations ($p > 0.05$), indicating weaker or non-existent relationships in this dataset.

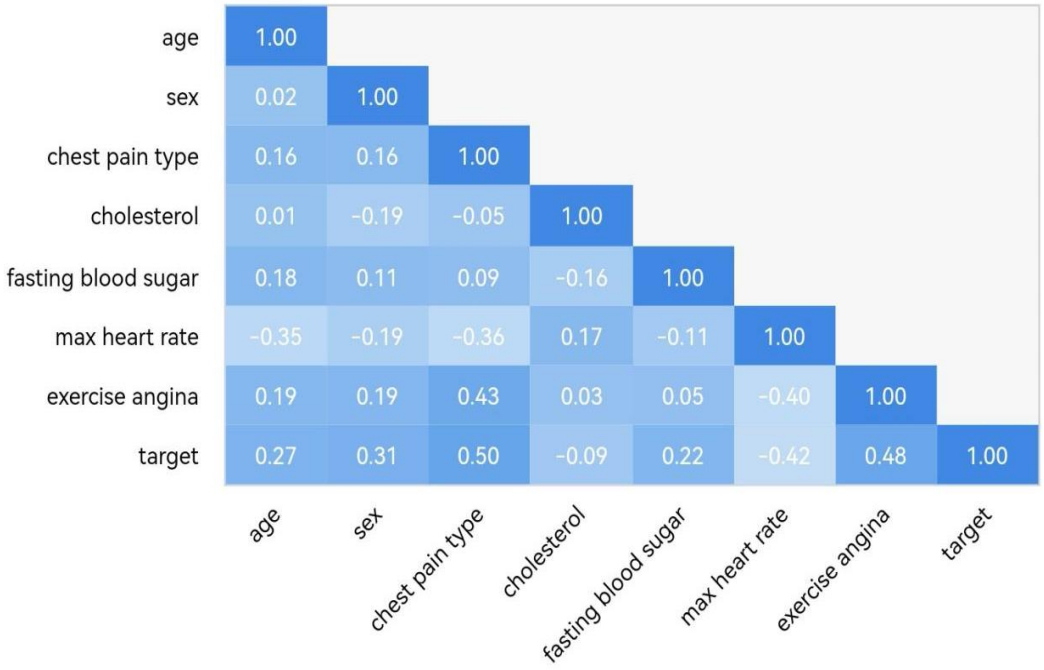

Figure 4. The relationship between 8 variables.

Alt Text for Figure 4: A heatmap showing the correlation coefficients between age and other key variables.

3.3. Model results

To evaluate the predictive capability of the logistic regression model, table 2 summarizes the prediction accuracy. The overall accuracy of 78.66% indicates that the logistic regression model performs reasonably well in distinguishing between individuals with and without heart disease. While not perfect, it offers a practical and interpretable approach to medical prediction and diagnosis, with relatively balanced performance across both categories.

Table 2. Logistic regression model results.

true value	predicted as 0	predicted as 1	Accuracy	Error Rate
0	438	123	78.07%	21.93%
1	131	498	79.17%	20.83%
Overall	-	-	78.66%	21.34%

4. Conclusion

This study analyzed the relationship between various factors and the likelihood of developing heart disease using logistic regression. Based on the results in Table 2, the prediction accuracy of the logistic regression model was 78.66%, with an error rate of 21.34%. In the field of medical diagnosis, higher accuracy contributes significantly to early detection and effective treatment of diseases.

Therefore, improving prediction accuracy is of great importance for public health and clinical decision-making.

Although logistic regression is a useful tool for identifying potential risk factors and estimating the probability of disease, it has limitations. Specifically, it can only reveal statistical associations between variables, rather than establishing direct causal relationships. Moreover, prediction errors are inevitable due to the complexity of medical data and the limitations of the model itself.

Despite these challenges, the findings demonstrate the practical value of logistic regression in disease prediction and prevention. It enables researchers and healthcare professionals to identify high-risk groups and prioritize preventive interventions. Future research should focus on optimizing model performance, integrating additional variables, and combining logistic regression with other advanced machine learning techniques to enhance diagnostic accuracy and reliability further.

References

[1] Chen, R., Zhang, C.K., Wang, Y.Q., Yan, H.X. & Guo, R. 2021. Study on the correlation between the characteristics of recursive quantitative analysis of pulse graph and blood coagulation function in patients with coronary atherosclerotic heart disease. *Chinese Journal of Traditional Chinese Medicine Information* 28(1): 107-112.

[2] Ke, W., Ramon, et al. 2015. Developmental origin of age-related coronary artery disease. *Cardiovascular Research.*

[3] Lekha, A.P., Salil, S. & Ronak, R. 2017. Jaideep Rajebahadur Coronary artery disease in women. *Indian Heart Journal* 9: 532-538.

[4] Li, Z.R., Liu, D.G., Liu, J.J., Yang, X.Y. & Xiao, C.J. 2017. Research progress on cardiac rehabilitation and prevention of coronary heart disease. *Hunan Journal of Traditional Chinese Medicine* 33(2): 152-154.

[5] Liu, Y.B., Yang, S.H., Chen, H.W. & Xing Y.S. 2019. The correlation between adipocytokines and the severity of coronary artery disease in elderly patients with coronary heart disease. *Journal of Practical Medicine* 35(11): 1799-1804.

[6] Sackner-Bernstein, J.D. 2005. Risk of worsening renal function with nesiritide in patients with acutely decompensated heart failure. *Circulation* 111(12), 1487-1491.

[7] Song, X.D., Zhao, D.D., Du, T.H. & Wang, Y.P. 2018. Research progress on cardiac rehabilitation therapy and prevention of cardiovascular diseases. *Modern Distance Education of Traditional Chinese Medicine in China* 16(4): 151-153.

[8] Wang, Y.G. & Liang, F. 2023. Guidelines for the Management of Chronic Diseases in Elderly Coronary Heart Disease. *Research on Integrated Traditional Chinese and Western Medicine* 15 (01): 30-42.

[9] Wu, Y., Li, B.C., Ding, Y.K. & Sun, X. 2024. Interpretation of the 2023 AHA/ ACC/ ACCP/ ASPC/NLA/ PCNA Guidelines for the Management of Chronic Coronary Heart Disease. *Chinese Journal of Evidence Based Medicine* 24(9): 1094-109

[10] Zhou, K.Y., Chen, Q. & Liao, X. 2024. Progress in diagnosis and clinical treatment of senile atherosclerotic heart disease. *Hebei Medical Journal* 30 (1): 168-171.