

Coronary Heart Disease Prediction from Common Risk Factors

Jianing Song *

College of Chang Wang, Nanjing University of Information and Science Technology, Nanjing,
Jiangsu Province, 210000, P. R. China

* Corresponding Author Email: shenmiao@ldy.edu.rs

Abstract. Coronary heart disease (CHD) is the most prevalent type of heart diseases, and is one of the most common chronic diseases leading to death. This study analyzed the data collected from coronary heart disease patients and developed a predictive model for coronary heart disease diagnosis. The dataset from Kaggle website contains 462 observations, one target variable and 9 common risk factors, such as systolic blood pressure (sbl), the use of tobacco each year, low density lipoprotein (ldl), body adiposity index (BAI), family history, the score of type A personality, body mass index (BMI), daily use of alcohol and age. At first, the data preprocessing indicates that there are no missing value and extreme distribution, and the correlation and collinearity between the variables are weak. Then the study develops two prediction models using the methods such as k-nearest neighbor (KNN) and logistic regression. To choose the most suitable one, the indicators like kappa value and accuracy are the main factors to evaluate the model's performance. The results shows that the logistic regression model have the higher accuracy, which is equal to 0.7174. After dropping some variables which have little impact on final prediction, the outputs present that yearly use of tobacco, low density lipoprotein, family history and BMI are the main factors which may influence the diagnosis of CHD. In the end, more clinical patients' data still needs to be collected for improving the model's accuracy.

Keywords: Logistic regression; K-nearest neighbor; prediction model; coronary heart disease.

1. Introduction

Heart disease is a kind of chronic diseases which is the most prevalent, and is the leading reason of death no matter sexes or racial groups in the world. In United States, millions of people are impacted by heart disease, about 695,000 people died from heart disease in 2021 according to the American Heart Association [1]. Also, the disease exerts great pressure on the economy, the cost including care services, medicines and productivity lost caused by death was up to \$239.9 billion from 2018 to 2019 [2]. Coronary Heart Disease (CHD), also named coronary artery disease (CAD) is a type of heart disease which is the most prevalent, and caused the death of 375,476 people in American, 2021[1]. The coronary disease caused by atherosclerosis, which means the arteries are narrowed due to the fatty deposits when carrying blood to the heart [3]. The patients may have following symptoms such as chest pain, heart attack which represents a part of heart muscle damaged or died and the arteries are blocked when the blood flowing to the heart, heart failure which means the heart does not have enough energy to offer blood and oxygen to other organs in body, and arrhythmia [4]. The Centers for Disease Control and Prevention (CDC) introduced that the high pressure of blood, high blood cholesterol level and tobacco are main factors for CHD, almost half of Americans have one of three mentioned factors. Moreover, the family history, age, sex, inflammation and lipoproteins can also be the biomarkers for CHD [5, 6]. These factors can be used in clinical diagnosis and therapies for coronary heart disease, which people can take an initial test of common risk factors and bloodwork or other tests followed. As the human lifespan increased, the older adult population keep rising. The Behavioral Risk Factor Surveillance System (BRFSS) is a phone-called survey which ask questions about risk behaviors, some chronic diseases, and so on, collecting data from American residents. According to the survey in 2015-2017, almost 7% adults aged between 45 and 64 years who are reported having CHD and nearly 20% of people aged over 65 years, and there are still numerous people who are not aware of or being told they have CHD. The preventative

measures and prediction before negative outcomes taking place are beneficial not only to increase the quality of life but to relieve the pressure for public health professionals and economy.

The clinical predictions according to some main risk factors for CHD diagnosis may still need to improve nowadays, caused by a great deal of cases are not detected by these scores [7]. Therefore, this article aims to create a coronary heart disease indicators prediction model based on machine learning algorithms, choosing the possible risk factors as variables. To choose the most suitable model, the author may compare different models such as random forest model, KNN model, logistic regression model, the kappa value and accuracy are the main factors to show the model performance.

2. Methodology

2.1. Data Source

The coronary heart disease patients' condition data used in this paper comes from the Kaggle website. The original data was saved in .CSV format.

2.2. Data Introduction

The CHD dataset in the article includes 462 observations and 10 variables, which have none missing values. Table 1 presents the meaning of 10 variables, among which have 1 character variable and 9 numeric variables.

Table 1. The meaning of 10 variables

Variable	Type	Meaning
sbp	numeric	The pressure of systolic blood
tobacco	numeric	Yearly tobacco use (in kg)
ldl	numeric	Low density lipoprotein
adiposity	numeric	Body adiposity index (BAI)
famhist	character	Family history: 'Present' means family members do have disease history, 'Absent' means not.
typea	numeric	The score of type A personality
obesity	numeric	Body mass index (BMI)
alcohol	numeric	Alcohol use daily (in g)
age	numeric	age
chd	numeric	The diagnosis of CHD: 0= no, 1=yes

Before creating prediction models, the correlation and distribution of each variable should be checked, since the collinearity may influences the performance of the models. To make a prediction of the risk of CHD, the target variable is 'chd'. The famhist character variable has two factors, which is absent and present. For the following study, the 'absent' is equal to 0, and the 'present' is equal to 1.

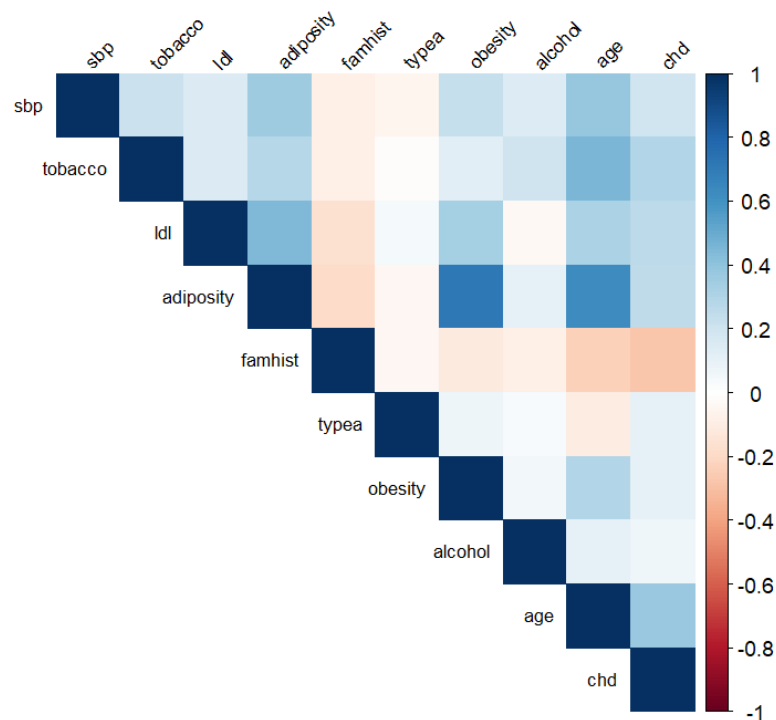


Fig 1. Collinearity between the target variable and variables

Correlation with variables

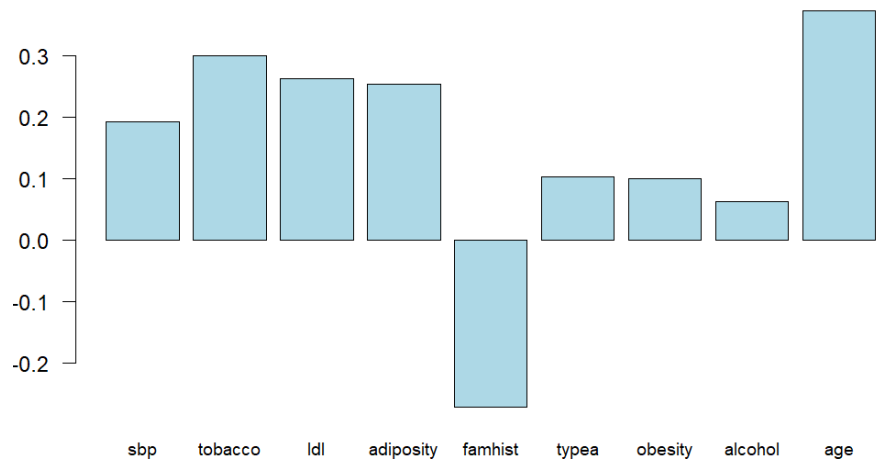


Fig 2. Correlation of variables

Figure 1 and Figure 2 presents a corr-plot and a bar-plot, which show the collinearity between the target variable and left variables, and correlation of all variables. Shown as the two figures, both the collinearity and correlation of variables are weak, which means the variables are all independent.

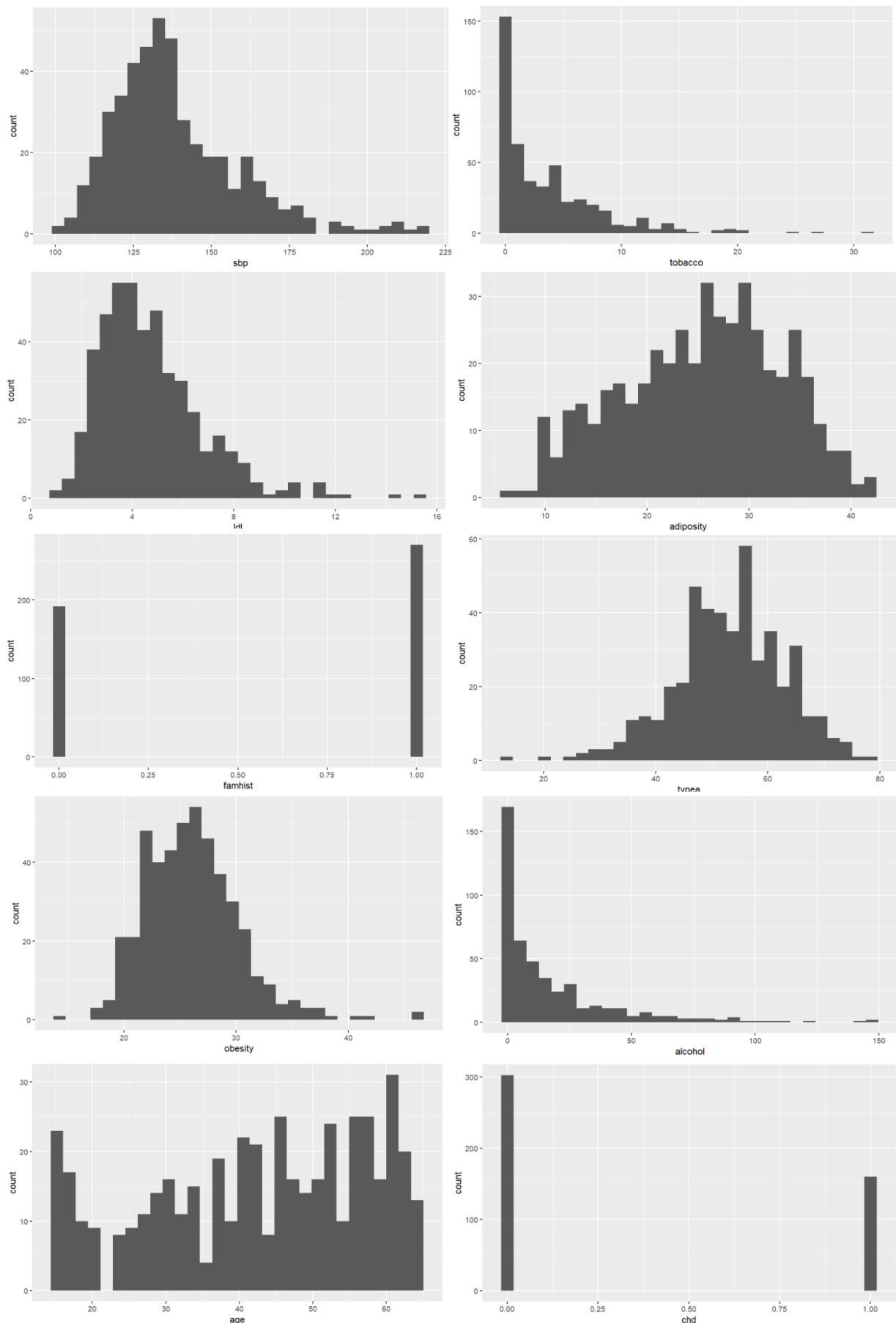


Fig 3. The histograms of 10 variables distributions

Figure 3. indicates the histograms of 10 variables distributions, which shows that there is no extreme distribution in the dataset.

2.3. Method Introduction

The methods in the study are k-nearest neighbor (KNN) and logistic regression for prediction. The kappa value and the accuracy are the main factors to evaluate the performance. If Kappa is equal to 1, the prediction of model and the actual observation are exactly same. If kappa is equal to 0, the performance of model is the same as the random speculation. If kappa is less than 0, the performance of model may be terrible. Accuracy represents the percentage of correct classifications in total number of classifications. The performance of prediction may be better if the accuracy is near to 1. The first step is to build these two models. And then compare their kappa value and accuracy, choose the most suitable model. Finally, analyze the results and discussion.

3. Results and Discussion

3.1. KNN Model

K-nearest neighbor model is used for classification and regression. Firstly, separate the dataset into a training data and a test data. The basic principle of KNN algorithm is classifying a new observation from given data with group labels, and looking at the objects that are closest to the new one. Since the KNN model focuses on the points which is near to the new observation point, the labels around the points and the number of the points which are observed are significant. However, the problem, the anti-noise, exists because of the classification of the new observations by focusing on the other observations. A few outliers in the dataset may impact the final performance of the model.

In order to solve this problem, cross-validation method can be used to reduce the disadvantages by dividing different training datasets and test datasets. Assign 10 to the cross-validations, which means output the best prediction results among ten times under the current k value. This method can ensure the accuracy and become more efficient without too many crossing times. Then determine the most suitable k value can evaluate the number of observations which are appropriate to observe. Figure 4. indicates the accuracy under different k values, and the highest value occurs when k is equal to 14.

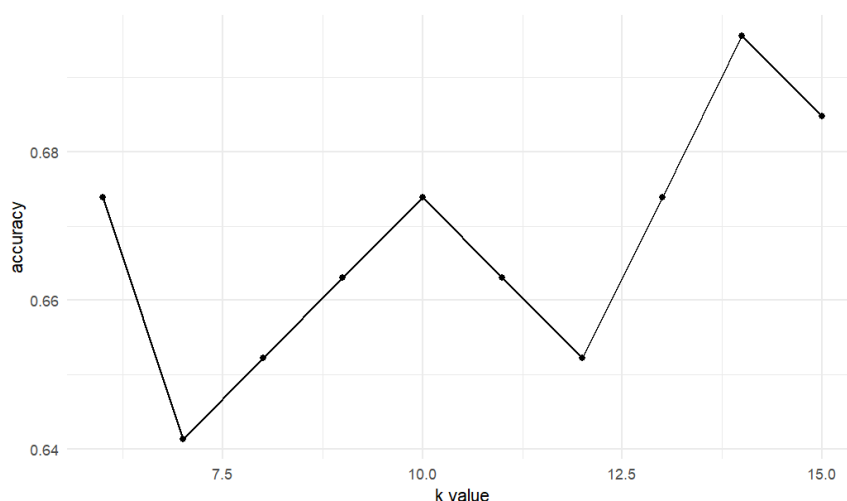


Fig 4. The accuracy of KNN model under different k values

Table 2. The indicators of KNN model

Accuracy	Kappa	Precision	Recall	F1 Score
0.6413	0.1404	0.6901	0.8167	0.7481

According to Table 2, the accuracy of KNN model is equal to 0.6413, and the kappa value is equal to 0.1404.

3.2. Logistic Regression Model

Logistic regression model allows to assign different coefficients to different variables and get an intercept. The formula using these values can predict the probability of the diagnosis of coronary heart disease. Table 3. represents the results of the model.

Table 3. The coefficients of variables in the logistic regression model

Variables	Estimate	Std. Error
(Intercept)	-4.5810	
sbp	0.0074	0.0066
tobacco	0.0987	0.0305
ldl	0.1925	0.0703
adiposity	0.0403	0.0342
famhist	-1.0019	0.2610
typea	0.0491	0.0144
obesity	-0.1262	0.0556
alcohol	0.0007	0.0053
age	0.0378	0.0134

From the previous table, the P-values of sbp, adiposity, typea, alcohol and age are less than 0.05, which means these variables have little influence for the accuracy and may only make the model complex. So that the regression model should be fitted again after dropping these variables. The coefficients of the fitted model are shown in Table 4.

Table 4. The coefficients of variables after dropping

Variables	Estimate	Std. Error
(Intercept)	-0.7445	
tobacco	0.1385	0.0282
ldl	0.2415	0.0672
famhist	-1.1266	0.2462
obesity	-0.0391	0.0339

The formula can be presented:

$$\log\left(\frac{P}{1-P}\right) = -0.7445 + 0.1385\text{tobacco} + 0.2415\text{ldl} - 1.1266\text{famhist} - 0.0391\text{obesity} \quad (1)$$

Overall, the famhist and obesity increase each unit, logodds, $P/(1-P)$, may decrease. And the sbp, adiposity, typea, alcohol, tobacco, ldl and age increase each unit, the logodds may increase.

Table 5. The indicators of logistic regression model

Accuracy	Kappa	Precision	Recall	F1 Score
0.7174	0.3385	0.75	0.85	0.7969

According to the Table 5, the accuracy of the logistic regression model is equal to 0.7174, and the kappa value is equal to 0.3385.

3.3. Discussion

To sum up, the logistic regression model is the most suitable prediction model, which has higher kappa value and accuracy, listed in Table 5, compared to the KNN model listed in Table 2. According to the study, the yearly use of tobacco, low density lipoprotein, family history and BMI are the main factors which may influence the diagnosis of coronary heart disease. According to the Amine's research, the obesity was more prevalent among the female compared to the male [8]. While the prevalence of smoking was significantly higher, in Jordon University of Science and Technology students' research [9]. Though the high systolic blood pressure, abnormal body adiposity index, high score of type A personality, numerous alcohol use daily and high age can add the probability of CHD, the influence is too small to be considered compared with the previous main factors.

Referring to the study conducted in Karachi-Pakistan, only about 20% people were aware of the coronary heart disease [10]. Early diagnosis can lead to not only lifestyle changes but also more effective treatment.

4. Conclusion

With the development of the society, people have been paying more attention to healthy lifestyle. Reducing the yearly tobacco use and keeping the low density lipoprotein at the normal concentration can decrease the possibility of CHD diagnosis. And the patients have family disease history or abnormal BMI may face more risks, so that they should pay more attention to the body check in daily life, and keep the BMI value around 25 Kg/m². The left factors, mentioned before can also be the references for patients' clinical check. Moreover, the research still needs to collect more clinical data for improving and fitting the model. In this study, the dataset only contains 462 observations without differentiating sexes. If more data, such as from different ages, regions or sex etc. can be included, the final model and formula may be more accurate.

References

- [1] Tsao C W, Aday A W, Almarzooq Z I, Beaton A Z, Bittencourt M S, Boehme A K, et al. Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. *Circulation*. 2023, 147.
- [2] National Center for Health Statistics. Percentage of coronary heart disease for adults aged 18 and over, United States, 2019—2021. National Health Interview Survey, 2023.
- [3] Visseren F L J, et al. 2021 ESC guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*, 2021, 42(34): 3227-3337.
- [4] Nasa Sinnott-Armstrong, Y. Tanigawa, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*, 2021, 53(2): 185-194.
- [5] Johnston N, Jernberg T, Lagerqvist B, Siegbahn A, Wallentin L. Improved identification of patients with coronary artery disease by the use of new lipid and lipoprotein biomarkers. *Am J Cardiol*, 2006, 97(5): 640-645.
- [6] Zakynthinos Z, Pappa N. Inflammatory biomarkers in coronary artery disease. *J Cardiol*, 2009, 53(3): 317-333.
- [7] Wilson P, Castelli W P, Kannel W B. Coronary risk prediction in adults. *Am J Cardiol*. 1987, 59(14): 91-94.
- [8] Amine E K, Samy M. Obesity among Female University Students in United Arab Emirates. *The Journal of the Royal Society for the Prevention of Health*, 1996, 116: 91-96.
- [9] Khader Y S, Alsadi A A. Smoking Habits among University Students in Jordon: Prevalence and Associated Factors. *Eastern Mediterranean Health Journal*, 2008, 14: 897-904.
- [10] Dodani S, Mistry R, Farooqi M, Khwaja A, Qureshi R. Prevalence and Awareness of Risk Factors and Behaviors of Coronay Heart Disease in an Urban Population of Karachi, the Largest City of Pakistan: A Community Survey. *Journal of Public Health*, 2004, 26: 245-249.