# The Research about Heart Attack Prediction Model

Tianyu Shi *

New York University, New York, USA

* Corresponding Author Email: ts4036@nyu.edu

**Abstract.** Nowadays, coronary heart disease is becoming the most important cause of death all over the world. The use of science technology makes it much more easier for people to analyze the causes for the different kinds of diseases. This article uses the Behavioral Risk Factor Surveillance System to assemble data on health-related risk behaviors from more than 400,000 Americans and analyze these data to provide a precise prediction model for coronary heart disease. The article uses methods like Logistic regression, support vector machine (SVM) and random forest (RF) to explore a better prediction model. After all the calculation, the accuracy of logistic regression is not very high because the accuracy is 85%. SVM and RF proves to make a better prediction since the accuracy of these two models are both above 90% and random forest is the best predictor. To discover how the variables contributes to construct the model, the weight of these variables are crucial. Of all these variables, the proportion of BMI is 44.34%, which has the highest weight and plays a key role in model construction. Moreover, further predictions and better models should be set up to analyze the data. In conclusion, the study shows that SVM and RF make better predictions of heart disease than simple logistic regression. More comprehensive data from more respondents and more suitable analytic models need to use to improve prediction.

**Keywords:** Heart disease; logistic regression; prediction model.

## 1. Introduction

Nowadays, heart disease is the primary cause of death for the majority of individuals globally. About 285 million people worldwide suffered from heart and other related ailments in 1990; that figure rose to 350 million in 2000; by 2010, that figure had surpassed 430 million. Globally, the number of people with cardiovascular disease has risen since 1997. Worldwide, there were over 290 million women (53%) and 260 million men living with cardiovascular disease in 2019. This means that more women than males had the disease. The most prevalent cardiovascular conditions are atrial fibrillation (60 million), peripheral artery disease (110 million), stroke (100 million), and coronary heart disease (200 million) globally [1-3].

It used to be exceedingly rare for someone under 40 to have a heart attack, but today one in five heart attack patients is under 40. Between 2002 and 2014, the annual rate of heart attacks among those under 40 increased by 2%. The incidence of coronary heart disease among youths has gradually increased as a result of alcohol consumption, obesity (high BMI), inactivity, and other factors that may negatively impact the cardiovascular system. Hypertension and high blood pressure are common risk factors for cardiovascular disease. It is regrettable that young individuals' hypertension is increasing more quickly than that of older ones, following the trend in cardiac disorders. One of the most vital organs in the human body, the heart is essential to the existence of the human [4, 5]. In China, the second most common cause of cardiovascular death is coronary heart disease, or CHD. Cardiovascular mortality account for roughly 22% of deaths in urban regions and 13% in rural ones [6, 7].

There are a lot of people in the United States who have heart disease, despite the fact that there are far fewer cardiac patients there than in China. Millions of Americans suffer from heart disease every year, which has a substantial financial impact on the nation's economy. About 647,000 lives are lost to heart disease in the United States alone each year, making it the leading cause of death. Heart disease can be caused by plaque buildup in major coronary arteries, age-related molecular alterations, chronic inflammation, excessive blood pressure, and diabetes [8, 9].

Despite the fact that coronary artery disease comes in a variety of forms, most people only become aware of it when they experience symptoms like chest pain or heart attacks. This information highlights the value of screenings and preventative interventions that can reliably identify heart disease in the general population. A more precise model must be developed in order to forecast heart attacks. Rapid breakthroughs in science have made it simpler to observe data and make predictions in the modern day. One example is the Behavioral Risk Factor Monitoring System. The CDC conducts an annual telephone survey on health-related topics called the Behavioral Risk Factor Surveillance System (BRFSS). The annual survey, which questions about health-risk behaviors, chronic conditions, and use of preventative services, is completed by over 400,000 Americans [10]. This 2015 Kaggle original dataset has 330 features and responses from 441,455 users. These traits, which include high blood pressure, diabetes, and alcohol usage, are either questions posed to participants directly or computed variables based on their answers. Only a few traits are included, based on research on heart disease and the factors that contribute to heart disease and other chronic health issues. 23,893 respondents had heart illness, while 229,787 respondents do not or did not have heart disease. There are 253,680 rows and 22 columns in the dataset. Predictions are made using techniques like random forest, support vector machines, and logistic regression.

## 2. Methods

### 2.1. Data Source

The heart attack patients' responses are collected by the CDC. It is found on Kaggle. The original data was saved in .CSV format.

### 2.2. Data Introduction

The original data set has 22 columns and 253,680 rows. There are 22 different variables in the original dataset and we will study on the 9 different variables listed below. We have 5 factor variables, which are binary, and 4 numeric variables. Of these binary variables, one means yes and zero means no (Table 1).

**Table 1.** Different types of variables

| Term | Type | Meaning |
|---|---|---|
| Heart Diseaseor Attack | factor | Respondents that reported having coronary heart disease |
| High BP | factor | Adults that have been told by a doctor in a authoritative hospital that they have high blood pressure |
| BMI | numeric | Body Mass Index |
| age | numeric | Fourteen-level age category (five years) |
| High Chol | factor | Adults that have been told that they have high blood cholesterol in the hospital |
| Smoker | factor | Individuals stated that they have smoked 100 or more cigarettes during their lifetime [Note: 5 packs = 100 cigarettes |
| Diabetes | numeric | People who have diabetes |
| PhysActivity | factor | Adults who said they had exercised or engaged in physical activity within the previous 30 days |
| MentHlth | numeric | How many days in the last 30 days did you feel that your mental health wasn't doing well? |

The first variable heart disease attack is our main target to show that whether respondents have ever reported having coronary heart disease. Any respondent who has ever experienced coronary heart disease will be included in our dataset as one. It's difficult for us to study on each variable

because we have 9 different variables. It is easier to analyze these data if we can analyze all the variables and their correlation. So the first thing to do is that we need to find if there is any relationship between the variables themselves. we construct a correlation matrix to find if there is any relationship between these variables

The relationship of each of the variables is stronger if the color is darker, which means the value is closer to one or minus one. On this graph, we can not find any dark color in each of these small squares. So there is almost no correlation between each variables.

However, based on our common sense, Individuals with diabetes, high blood pressure, and excessive cholesterol are more likely to develop coronary heart disease. We can discover the relationship between these variables and the number of people having heart disease. Figure 1, 2, 3 are three scatter plots generated through R studio. They can show the linear relationship between high blood pressure and heart disease attack, the linear relationship between high blood cholesterol and heart disease attack and the linear relationship between diabetes and heart disease attack. These three plots show that all of these variables have a strong positive linear relationship with the number of people having heart disease. It is clear that those who respond who do not have diabetes, high blood pressure, or high cholesterol have a lower risk of heart attack, which means that people having these three diseases are more likely to have coronary heart disease.

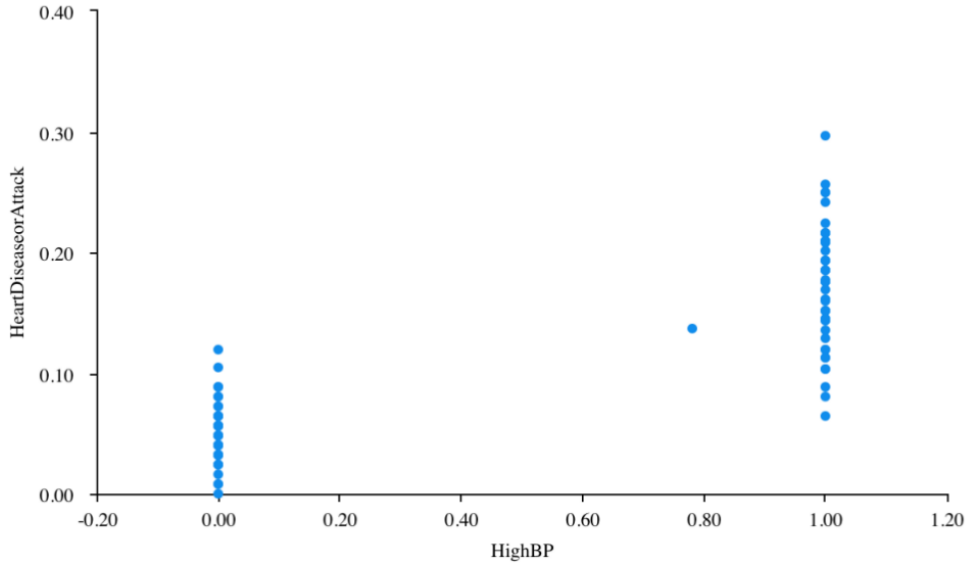


**Fig 1.** Scatter plots of high blood pressure

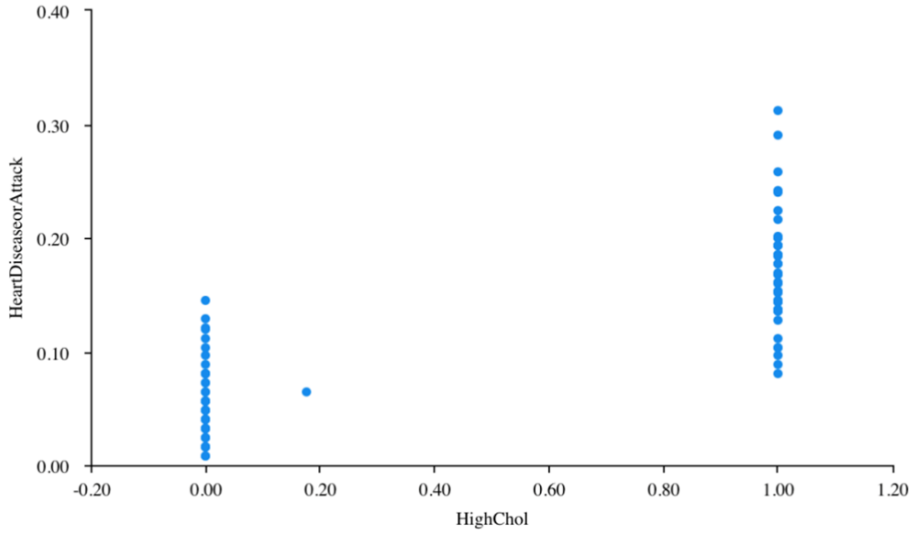


**Fig 2.** Scatter plots of high blood cholesterol

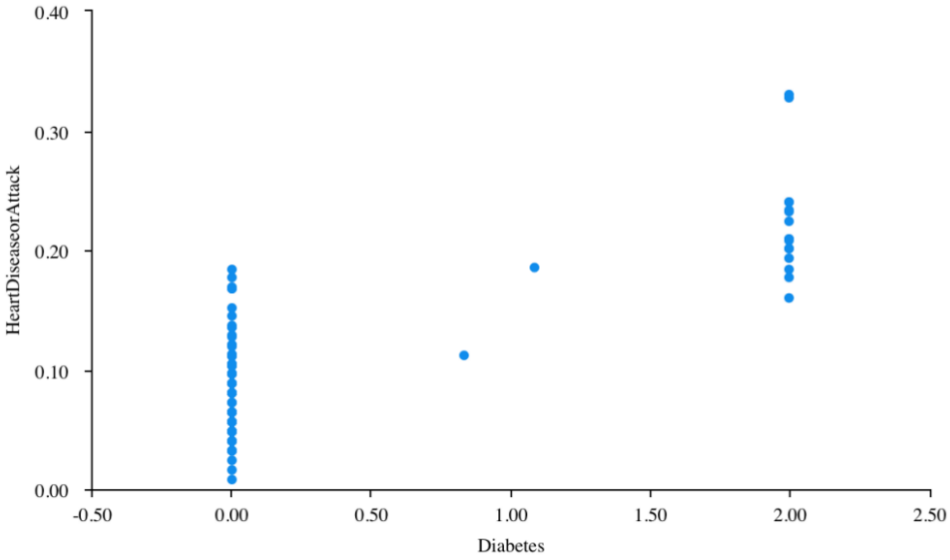


**Fig 3.** Scatter plots of diabetes

### 2.3. Method Introduction

The method used in this study are logistic regression, support vector machine (SVM) and random forest (RF) for each variable. The first step is to fits a logistic regression model and make predictions using all the variables. Then makes predictions using SVM and RF. Finally, calculates number and percentage of correct predictions.

## 3. Results and Discussion

Aside from diabetes, high blood pressure, and high blood cholesterol, all the factors are displayed in Figure 4. Age and body mass index seems to have a positive linear relationship with heart disease attack. For smoker, physical activity and mental health, the plot shows no relationship between them.

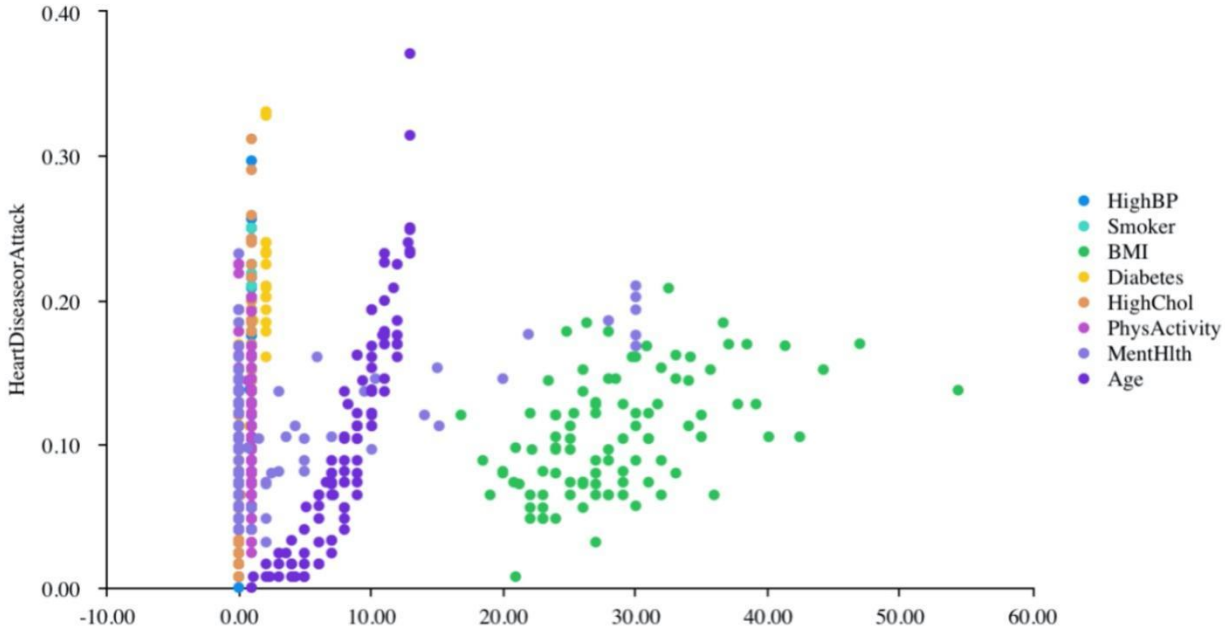


**Fig 4.** Scatter plots of potentially relative variables

Logistic regression modeling uses each variable individually to predict the number of people having coronary heart disease. The results are presented in Table 2 and table 3. As shown, the

accuracy is about 0.85. The percentage of samples with accurate prediction results throughout the entire sample is known as accuracy. A higher accuracy level indicates a better model.

**Table 2.** Result from Logistic regression

| Variable | coefficients |
|---|---|
| High BP | 0.0571625 |
| BMI | -0.0001598 |
| age | 0.0134899 |
| High Chol | 0.0480250 |
| Smoker | 0.0393831 |
| Diabetes | 0.0417967 |
| PhysActivity | -0.0217837 |
| MentHlth | 0.0019649 |
| intercept | -0.0741636 |

$$p = \frac{\exp(-0.074+0.057HB-0.0002BMI+0.013Age+\ldots+0.002MH)}{1+\exp(-0.074+0.057HB-0.0002BMI+0.013Age+\ldots+0.002MH} \quad (1)$$

So this is the logistical regression model. It's called logistic because the left-hand side involves the logit function. Here, p stands for the probability that the outcome is equal to 1. The coefficients of each of the variable correspond with those calculated in the Rstudio.

**Table 3.** Result from logistic regression

| | Train set | Test set |
|---|---|---|
| accuracy | 0.85 | 0.86 |
| F-1 score | 0.85 | 0.85 |

The accuracy of logistic regression is about 85% after all the calculation. Since the accuracy of logistic regression is not very high, it is necessary to use other models to make a better prediction. Here, support vector machine (SVM) and random forest (RF) are used and the results are shown (Table 4).

**Table 4.** Result from SVM and RF

| | Train set accuracy | Test set accuracy | Train set F-1 score | Test set F-1 score |
|---|---|---|---|---|
| support vector machine | 0.9 | 0.8 | 0.85 | 0.84 |
| random forest | 0.96 | 0.83 | 0.96 | 0.85 |

**Table 5.** Result from weights of each variable

| | Weights |
|---|---|
| BMI | 0.4434 |
| HighBP | 0.2178 |
| Diabetes | 0.1533 |
| HighChol | 0.0435 |
| Age | 0.0401 |
| MentHealth | 0.0397 |
| Physactivity | 0.0317 |
| Smoker | 0.0306 |

It is obvious that the accuracy of these two models is definitely higher than that of logistic regression since the accuracy of these two models are both above 90%. Random forest, which has an

accuracy of 96%, is the best predictor. Of all these variables, BMI has the largest weight of 0.44 and play the most important role in the model construction (Table 5).

## 4. Conclusion

This study indicates that the accuracy of logistic regression is about 85% and the accuracy of SVM and RF are both above 90%. To choose the best indicators, random forest has the highest accuracy and will absolutely make better predictions compared to the other two models. For all the weights of each of the variables, BMI, which has the largest weight, has the largest influence on the model construction. However, from the results of the models, it is important that further predictions and better models should be set up to analyze the data. Discovering the relations between each variable and the relation between the variables with the target is also crucial.

## References

[1] Ancel Keys. Coronary heart disease-The global picture. Atherosclerosis. 1975, 22(2): 149-192.

[2] Jie Hoffman, Kaplan S, Liberthson R R. Prevalence of congenital heart disease. American heart journal, 2004, 147(3): 425-439.

[3] James E. Dalen M D, Joseph S. Alpert M D , Robert J. Goldberg M D , Ronald S. Weinstein M D. The Epidemic of the 20th Century: Coronary Heart Disease. The American Journal of Medical, 2014, 127(9): 807-812.

[4] Rubin J B, Borden W B. Coronary Heart Disease in Young Adults. Current atherosclerosis reports, 2012, 14(1): 140-149.

[5] Gertler M M, Garn S M, White P D. Young Candidates For Coronary Heart Disease. Journal of the American Medical, 1951, 147(7): 621-625.

[6] Zhang X H, Lu Z L, Liu L. Coronary heart disease in China. Heart, 2008, 94(9): 1126-1131.

[7] Zhao D, Liu J, Wang Z, Zhang X, Zhou M. Epidemiology of cardiovascular disease in China: current features and implications. Nature Reviews Cardiology, 2019, 22(16): 203-212.

[8] Benjamin E J, Virani S S, Callaway C W. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation, 2018, 137(12): 137.

[9] Capewell S, Ford E S, Croft J B. Cardiovascular risk factor trends and potential for reducing coronary heart disease mortality in the United States of America. Bulletin of the World, 2010, 88(1): 120-130.

[10] Centers for Disease Control and Prevention (CDC). Prevalence of coronary heart disease-United States, 2006--2010. MMWR: Morbidity & Mortality Weekly Report, 2011, 60(40): 1377.