Heart Attack Analysis

DSC-680 Project Proposal

Isabella Sturm

2021-12-04

Topic

Given patient information such as age, sex, and blood pressure, the goal of this project is to determine what factors mean has high risk of having a heart attack.

Business Problem

Heart Disease is the leading cause of death in the United States, killing nearly 700,000 people each year. Heart attacks are a form of heart disease, affecting about 805,000 people every year. Detecting a heart attack early is important in reducing damage to the heart, so beyond knowing what to look out for if a person is experiencing one, knowing whether a person is at high risk for heart disease and heart attacks is also important.

The goal of this project is to determine key features that determine who is at high risk for heart attacks.

Datasets

The dataset I am using for this project is a Heart Disease Dataset (Comprehensive) from IEEE.org. It is a combined dataset of five other heart disease datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog (Heart) Data Set:

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

The dataset has 1190 rows with 11 features detailed in the following table:

Feature	**Description**
Age	Age of the patient
Sex	Sex of the patient (1=M, 0=F)
Chest pain type	Chest pain types: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 3 = asymptomatic
Resting bp s	Resting blood pressure (mm Hg)
Cholesterol	Cholesterol (gm/dl) from BMI sensor
Fasting blood sugar	1 = true (fasting blood pressure > 120 mg/dl), 0 = false
Resting ecg	Resting electrocardiographic results: 0 = normal 1= having ST-T wave abnormality (T wave inversion and/or ST elevation or depression of > 0.05 mV) 2 = showing probably or definite left ventrical hypertrophy by Estes' criteria
Max heart rate	Maximum heart rate achieved
Exercise angina	Exercise induced angina (0 = false, 1 = true)
Oldpeak	ST depression
ST slope	The slope of the peak exercise ST segment: 1 = upsloping 2 = flat 3 = downsloping
Target	Class: 0 = normal 1 = heart disease

Methods

This is a classification problem with a binary target indicating whether or not someone has or is at risk for heart disease.

I will start with some exploratory analysis to see how each factor correlate to each other and how the values are distributed. After that, I will do some data modeling of the data.

There are several methods I plan to try for this project starting with logistic regression. Other methods include decision tree and random forest. I would like to investigate other classification options, but I will start with these.

Ethical Considerations

With medical related data, an obvious ethical concern is patient privacy. In this case, all identifying information is already removed. However, in a scenario where I would be collecting data or have all the patient data, it would be a major ethical concern to use that data improperly.

Challenges/Issues

While I do not suspect too many issues, especially considering the data seems relatively clean, one concern I have is about the data size. Being that there is only about 1200 rows, it is possible there could be limited data for predictions and modeling the data. Also, the data may be unbalanced, partially because of the limited data size.

References

The Centers for Disease Control and Prevention (CDC) has facts that I will reference to support my project topic and the results that come out of the analysis.

- https://www.cdc.gov/heartdisease/facts.htm
- https://www.cdc.gov/heartdisease/heart_attack.htm