

Data analysis to extract what affect heart disease

Hamza Saad

Abstract

Heart disease is all health problems that affect the heart and blood vessels. It is the main cause of death around the world. In 2016, the death rate due to heart attacks was about 17.9 million cases, according to the World Health Organization statistics, equivalent to 31% of the total deaths in that year, and the number of deaths due to a heart attack was approximately 15.2% of the total deaths 17.9 million. With the advancement of treatment, people are unable to stop the huge number of deaths, especially the huge death number came from the developing country. In this study, we analyzed data for thirteen variables that were used to predict heart disease for people between the ages of 29 and 77 years. Heart disease in this sample includes coronary heart disease, cardiomyopathy, and cardiovascular disease. We analyzed data for 303 patients to understand what affects heart disease by considering important variables. Data included two types: discrete (0, 1) and numerical attributes. Statistical analysis and data mining are used to predict heart disease and to extract the main variables that have directly affect the disease. The highest accuracy got from a neural network with an accuracy of 83% and area under the curve 90%. This accuracy was obtained from strong variables that were applied to predict diseases such as blood disorder, chest pain type, the number of major vessels, old peak, and exercise-induced angina. In Pearson correlation we got a significant relationship but also this was for limited variables. We did not get significant importance for age or gender because heart disease attacks everyone at any age.

Cardiovascular diseases (CVDs) or heart disease are the number one cause of death globally with 17.9 million death cases each year. CVDs are concertedly contributed by hypertension, diabetes, overweight and unhealthy lifestyles. You can read more on the heart disease statistics and causes for self-understanding. This project covers manual exploratory data analysis and using pandas profiling in Jupyter Notebook, on Google Colab. The dataset used in this project is UCI Heart Disease dataset, and both data and code for this project are available on my GitHub repository.

Initially, the dataset contains 76 features or attributes from 303 patients; however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, here we will be using the dataset consisting of 303 patients with 14 features set.

World Health Organization statistics, equivalent to 31% of the total deaths in that year, and the number of deaths due to a heart attack was approximately 15.2% of the total deaths 17.9 million. With the advancement of treatment, people are unable to stop the huge number of deaths, especially the huge death number came from the developing country. In this study, we analyzed data for thirteen variables that were used to predict heart disease for people between the ages of 29 and 77 years. Heart disease in this sample includes coronary heart disease, cardiomyopathy, and cardiovascular disease. We analyzed data for 303 patients to understand what affects heart disease by considering important variables. Data included two types: discrete (0, 1) and numerical attributes. Statistical analysis and data mining are used to predict heart disease and to extract the main variables that have directly affect the disease.

However, looking at the bar graph above, its raised a question of higher number of healthy subject having typical_angina. Or in other word, most of the healthy subject having chest pain, which is also discussed here. Chest pain can be subjective due to stress, physical activities and many more and varies between gender. Women and elderly patients usually have atypical symptoms with a history of disease. This article provide analysis comparing typical anginal vs nontypical angina patients in a clinical trial.

Fasting blood sugar or fbs is a diabetes indicator with fbs >120 mg/d is considered diabetic (True class). Here, we observe that the number for class true, is lower compared to class false. However, if we look closely, there are higher number of heart disease patient without diabetes. This provide an indication that fbs might not be a strong feature differentiating between heart disease an non-disease patient.

Bottom Note: This work is partly presented at 9th International Conference on Interventional Cardiology at August 09-10, 2021 | Webinar

Hamza Saad

Binghamton University School of Engineering, NY 13902, USA, E-mail: hamzasaadgh@gmail.com