Arti Mahaldar

DSC 530 Final Term Project

Date: 2/26/2022

**Introduction**

Cardiovascular disease (CVD) is a group of conditions that involve the heart and blood

vessels. Common complications include heart attack, chest pain (angina), or stroke.

The purpose of this project is to predict the effects of different parameters recorded in the data to predict mortality of the patient. Looking at the heart disease dataset, my statistical question is can heart disease be predicted? I decided to focus on the following five variables: age, sex, type of chest pain, cholesterol, and resting blood pressure.

**Outcome of EDA**

Going through the data, I found that the age data had a normal distribution. I also found that males comprised the majority of the study as well as asymptomatic chest pain was the major type of chest pain. The outcome of my exploratory data analysis was that resting blood pressure alone was not good enough to predict heart disease. I felt that I should have started to look at the different variables together instead of separate towards the end of the data analysis process. It might be case that it is multiple variables that affect heart disease instead of just a single variable.

**What do you feel was missed during the analysis?**

How over the years the heart decease changed and how lifestyle and eating habits can influence the heart disease prediction

**Were there any variables you felt could have helped in the analysis?**

I feel focusing on variables like Resting ECG , Oldpeak , and ST\_Slope may have been better because those were part of a test on the heart. I would assume that those could potentially predict heart disease better.

**Were there any assumptions you felt were incorrect?**

I think the major assumption I made was in choosing my variables as I continued down my data analysis process. Another assumption will be the sample size of the data which could have impacted the data analysis. I don’t think that is an issue because it is in the exploratory phase. I can always revisit the dataset with different variables.

**What challenges did you face, what did you not fully understand?**

The process of testing the hypothesis proved to be challenging for me. While I know that the process is four steps: choose test statistic, define null hypothesis, compute p-value, and interpret. The coding portion of it and choosing the test statistic was confusing.