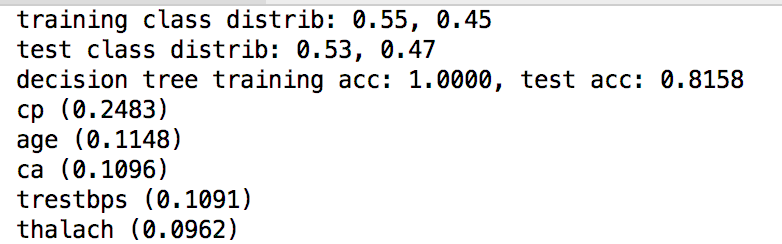
Bhogesh Maddirala

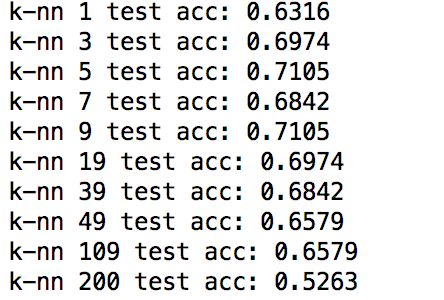
CS 397A

Predicting Coronary Heart Disease

I chose a dataset that collected data on heart disease. I wanted to explore what various variables are linked with coronary heart disease, as it’s a major issue in this country. I found the dataset on Kaggle, so processing it wasn’t hard. I conducted a few different tests, for example, I used sklearn to try and predict whether a particular individual would have heart disease or not. Additionally, I used sklearn to find what variables weighed the most, when trying to predict heart disease. After those tests, I decided to plot a few scatter plots of well-known variables that affect heart disease like cholesterol levels and heart rate. These were also used as indicators to check whether I was using the data correctly because if my graph was negatively correlated, I know something is wrong because it’s common knowledge that high cholesterol levels leads to an increased risk of heart disease. In terms of problems, one was that for some reason for my target variable, the knn test accuracy was reducing after it passed a certain scale. I tried dropping a variable from my data frame because I thought it was interfering with it, and it did improve the accuracy, but the issue still persisted. I have included scatter plots and images of what happens when I run my program.

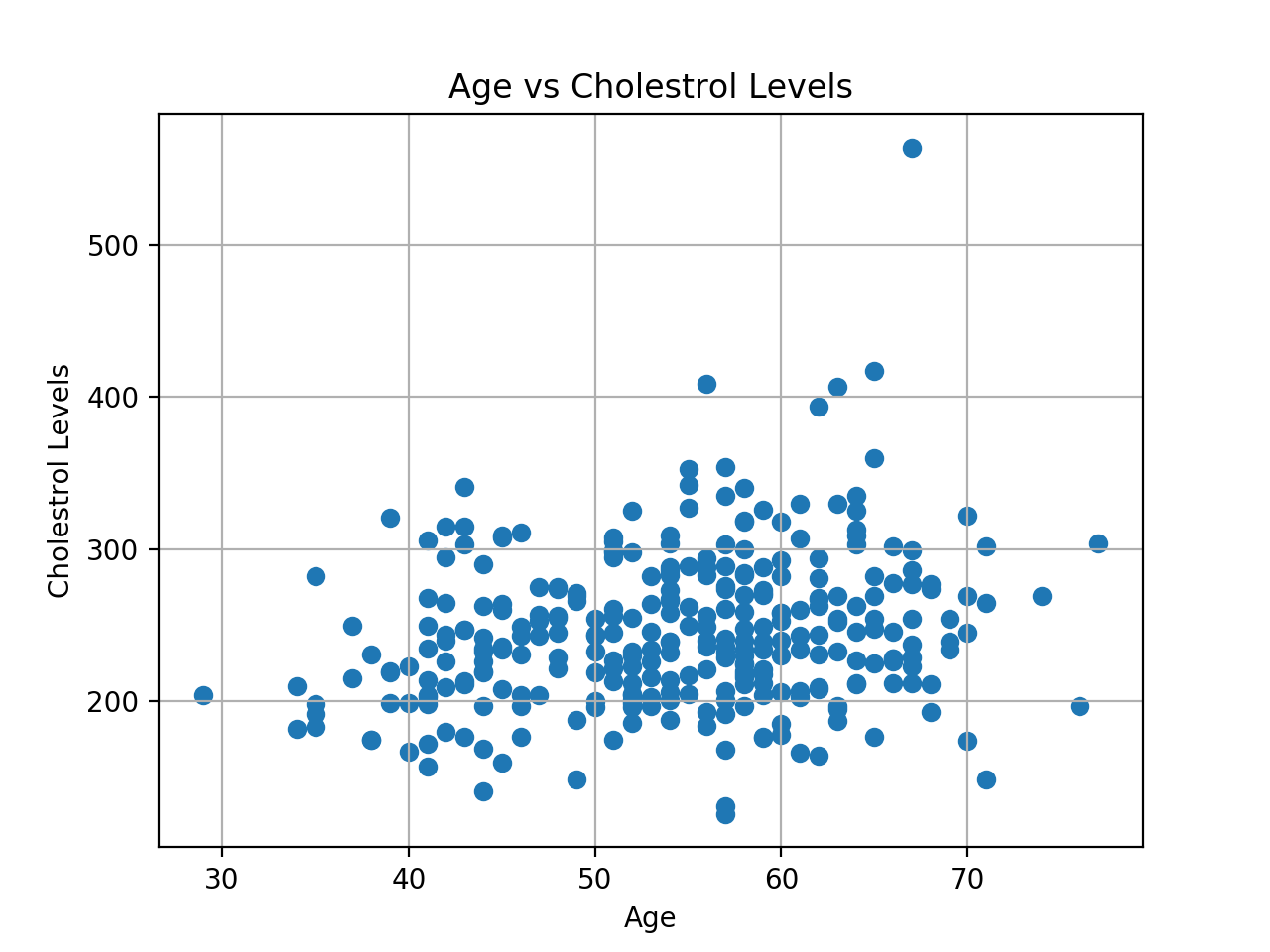
Heart disease has been the cause of a lot of death’s in this country, so I decided to find what variables were the most prevalent for the people affected. Thus, we can learn from it and improve the quality of life, so we don’t suffer the same fate. I conducted the sklearn tests below, in order to find that information.



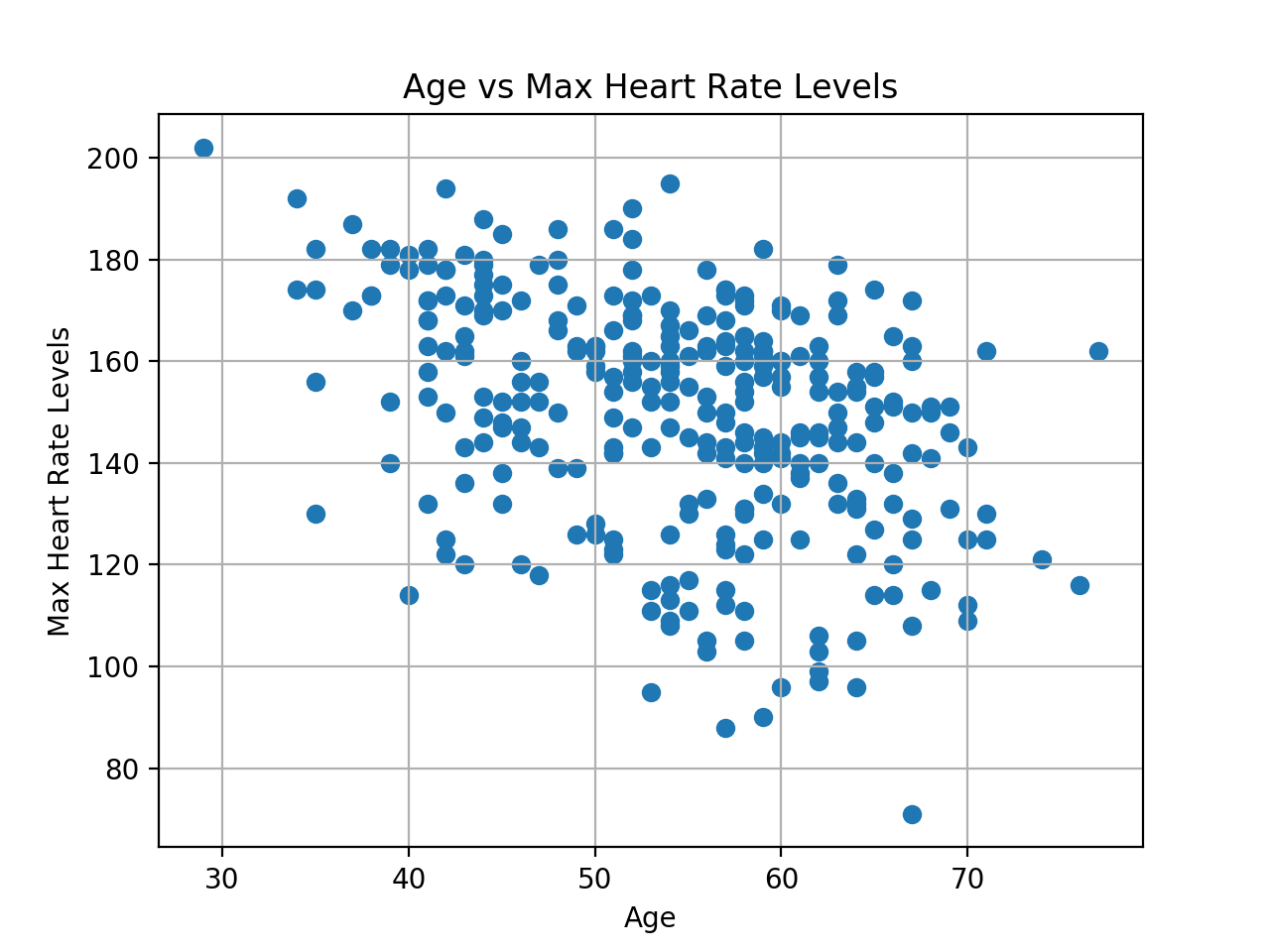


The tests showed that chest pain, age, number of major vessels, resting heart rate, and maximum heart rate reached were the variables most related with the target, who had heart disease. These variables’ association were found by using sklearn, to predict which variable is important to getting heart disease. Furthermore, this observation makes sense as the more severe the chest pain you have or the older you are, then the greater the chance you have of having coronary heart disease. This data analysis is telling us that we should seek immediate consultation, if are suffering from chest pain, as it could be a heart problem. Furthermore, this is telling us that the older you get, the more careful you have to be with your diet and health as you’re more susceptible to heart disease. The other three variables have also showed this same trend, but not to the degree of chest pain or age. The general accuracy could have been higher and the knn test accuracy started to reduce, after we added more neighbors didn’t make sense but as previously stated I tried my best to try and fix this error.

After conducting tests sklearn, I decided to use the matplotlib skills I learned from the first homework to create two scatter plots one is between age and cholesterol levels, and the other is between age and maximum heart rate. I chose these variables because there have been several studies done that showed that these variables were correlated with each other and heart disease. I plotted them because I wanted to confirm these previous studies and also to make sure my dataset is accurate, as these widely accepted theories. Thus, if there was no correlation or the wrong correlation, I knew the dataset or I was wrong.



The scatterplot between age and cholesterol levels shows a strong positive correlation between these two variables. This means that as a person gets older, more likely than not, his cholesterol levels will increase. We know that high amounts of cholesterol isn’t good for the heart, thus this plot is reminding us that we need to take better care of ourselves, as we age.



The scatterplot between age and max heart rate levels shows a strong negative correlation between the two variables. This shows that as a person gets older, your maximum heart rate reduces. This implies that an older person’s heart is weaker than a younger person’s, thus they probably can’t handle as much, and are more prone to heart related events.

I wanted to plot the relationship between every variable and the target variable, which showed whether you had the heart disease or not. But, I couldn’t manage to because the target variable is filled with either 0 or 1 thus it’s hard to find a correlation with a scatter plot or similar graphs. Perhaps, a frequency table to show how many people had heart disease compared to their age, but the data shown would be redundant to what I already found.

To conclude, I researched the relationship between having heart disease and different variables. I felt that it’s a very prominent topic, therefore educating myself, and using the skills I learned, to find things that I should be careful about, in order to avoid heart disease. Age, chest pain, maximum heart rate, resting heart rate, and cholesterol are all possible predictors for the outcome of having a coronary heart disease. Thus, if any symptom related to these variables arise, or you just get older then we should begin to take more precaution and seek more medical attention, than you would previously.

I worked on this by myself and the code used for my project was derived from my previous homework, therefore it should be easy to run and understand. However, I put comments in my code, in case someone else wants to be able to read, what I made. Almost everything worked, and I felt that I achieved the goal I wanted in this exploratory data analysis, which was to find relationships between variables affecting the heart.