Noah Scott

Machine Learning

Final Project

Heart Disease Data Mining

The project I decided to do was a data mining project on heart disease patients. The reason I decided on this particular data set was because the personal connection I have with heart disease. The first thing I did when I decided on the data set that I was going to use was pull out all the data and figure out what they meant if I had any questions. The data is as followed:

”**age** - age in years ----**sex** - (1 = male; 0 = female------**cp** - chest pain type-------**trestbps** - resting blood pressure (in mm Hg on admission to the hospital)-----**chol** - serum cholestoral in mg/dl-------**fbs** - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)------**restecg** - resting electrocardiographic results------**thalach** - maximum heart rate achieved-------**exang** - exercise induced angina (1 = yes; 0 = no)- ----**oldpeak** - ST depression induced by exercise relative to rest------**slope** - the slope of the peak exercise ST segment------**ca** - number of major vessels (0-3) colored by flourosopy------**thal** - 3 = normal; 6 = fixed defect; 7 = reversable defect-target - have disease or not (1=yes, 0=no)”

With all this data and attributes I wanted to experiment and dive deeper to see if there were any particular reasons that people may experience heart disease. Meaning that if someone’s age was older than a certain age or what sex they are how more or less likely they will be to get a heart disease. I wanted to gather information that included a various amount of data to see if I could see any particular trends or reasons why some groups of people have a heart disease and some do not.

As I began the experimentation portion of the project I needed to decide on what particular data sets I wanted to focus on and which ones I thought would be more closely related patients getting heart disease. That’s how I came to use Age, Sex, Chest Pain, Fasting Blood Sugar, Blood pressure, and Maximum Heart Rate. The main factors on why I choose these particular data attributes is because I know what is wrong with my mom and I kind of have an idea of what she experiences and some of the were out of curiosity like the age and sex categories was chosen because I know my mom was born with a heart disease but I have met quite a few people that got a heart disease later in their life. I choose chest pain because I felt like this could be directly related to heart disease but also I felt as if this could be maybe a lung problem or something else that isn’t a heart problem because not all chest pain can be a heart disease because I feel as if a lot of people would have them. I wanted to see how well the dietary aspects of someone’s life weighed in them getting a heart disease that is why I choose to use the fasting blood sugar. This is where they test the glucose levels after a night of fasting to see if they are diabetic or not and I wanted to see if this effected patients hearts. I chose blood pressure because this does have a direct correlation with your heart because it measures how hard the blood is pushing against the walls of arteries. I also choose to use sex is because the different strain women and men put on their body and like hormones and testosterone and I wanted to see if there was a great difference of those with and without heart disease that are different sex. Lastly setting up this project I chose to use mostly all bar graphs mainly because they are very easy to read and understand them but I also had one scatter plot to handle the data that came with maximum heart rate.

As we look at all the results in my project I was a little taken back on some of the data but also is some I was not surprised at all and my hypothesis was correct. To start off with I would like to point out in figure 1(below) that takes in the data of how many male and female patients were tested there were a lot more male so that could have thrown off some of the data because it was not an equal and fair data set. To be able to get to plot this I first had to index the data to only have the column of “sex” that there gave me all the rows within sex which consisted of females and males which were represented by 0’s (Females) and 1’s(males). To index I needed to make a name such as countFemales then go into df (“Data Frame” where all columns and rows of data are) and only use the column that was equal to the string ‘sex’ which allowed me to only have the column sex and the rows filled with 0’s and 1’s which are female and male. When I did this I also used a snippet of code that gave me the length of the row that contained 0’s which means that it gave me every single female in the column of ‘sex’. I then put the two data sets of females and males into a bar graph (shown below).

A screenshot of a cell phone

Description automatically generated(Figure 1)

I then got the percentage of the numbers by taking the countFemale and the countMale and dividing them by the complete row of the ‘sex’ column to get my answer. But as you can see male patients out numbered female patients by almost 37% I am not sure if this is because that is just the number of patients they had at the time or why this number is so different. So for the rest of the presented data we need to be aware that there are more males then there are females.

One of the first data points I examined was seeing if there were more or less patients with or without heart disease in their early life or late life. How I got this was by using the Pandas built in function “crosstab”. This was a very helpful function to use throughout my whole code because it allowed me to take all the rows in the age column and cross reference it with the target column. Which the target column is whether the patients had a heart disease or did not have a heart disease and that was also represented by 0’s and 1’s. then within the same line of code I could use the “.plot” and from there all I had to do was decide what type of graph I wanted to use the ‘figsize’ and the color. The ‘figsize’ is attribute allows us to specify the width and height of a figure and that makes the appearance of the graph presentable to the reader so all the data on the graph is spaced out and looks good. Moving forward from the basics of the graph I needed to add the x and y axis title with the tittle of the graph and the legend for the graph which was done by using the ’matplotlib.pyplot’ library. A picture containing object

Description automatically generated(Figure 2)

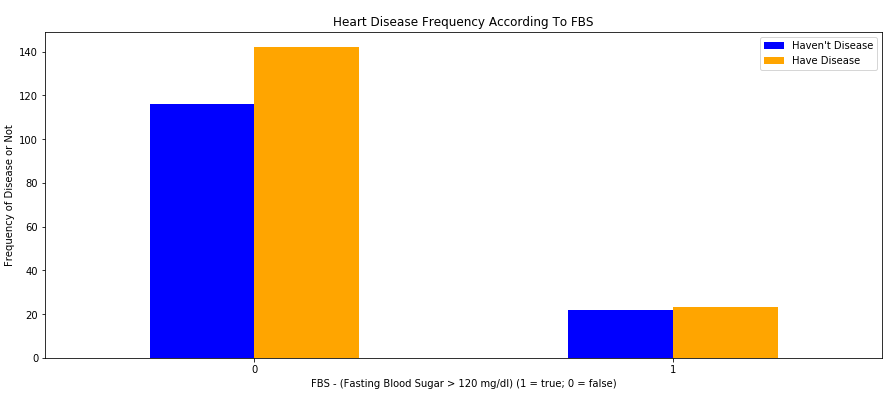
As you can see in figure 2 when getting all the data plotted there was a lot of questions that it left me. To start off with we can see that we had a lot more cases of people having heart disease from 41 to about 54 but the spread is quite over the place. The main questions I asked myself on this graph is whether or not these people have had the disease their whole life and were diagnosed at these ages or if bad habits and other health factors led them to acquire a heart disease at a later age and that data was not provided in this set.

As we move onto the next data that I wanted to examine was chest pain. I really wanted to look closer to this because of the many questions I had about what kind of chest pain this is and was this a result of a heart problem or from something else. The approach I took from this is actually quite simple the patients ranked their pain from zero to three. Assuming that zero is the least amount of pain and three being the most amount of pain. I used the same function as I did with the age one as I did with the chest pain and all the others. I used the ‘crosstabs’ function and did everything the exact same but I plugged in the column name for chest pain which was labeled as ‘cp’. Also one more distinct thing I did differently with this code was that numbers on the x-axis were facing the wrong way so with much research I found the ‘xticks(rotation)’ function that allowed me to turn the numbers on the x-axis so that they were facing the right way as you can see in figure 3.

A screenshot of a cell phone

Description automatically generated(Figure 3)

This data set was quite stunning to see mainly because how low the frequency of the disease was with the chest pain level three. As we look at the first level of chest pain which is zero the number of people that do not have a disease is very assuming and did not really surprise me I figured that number would be higher than the rest. Also while looking at this we see that people that had a chest pain level of two had a higher frequency of having a heart disease than in three which had me thinking that maybe the reason the number in chest pain level number two is so high is because they pass away before the can reach number three or because they get a heart transplant or cured of their disease.

Moving onto my next plotted data set we have fasting blood sugar. My hypothesis about this one before I plotted it was that the higher the blood sugar the higher the risk of a heart disease. I began this code as I did the others with the ‘crosstabs’ function because of the easy use of it. As we look at figure 4 (below) we can see that my hypothesis was extremely wrong and actually the opposite of what I thought was going to happen.(Figure 4)

As we can see here in figure 4 we see that the lower the fasting blood sugar is the more there is a frequency in heart disease. After doing more and more research on what exactly a fasting blood sugar is it made more sense on why my hypothesis was wrong. Fasting blood sugar mainly has to deal with diabetics and the higher a patients blood sugar is the more likely they are to be diabetic that explains why we see that in figure 4 on the bar graph that is labeled one which has lower frequency of heart disease is probably because those patients are diabetic. On the other hand if we look at the patients that have a low blood sugar which is normal we see that the frequency of heart disease is about even and there is nothing really exciting about this information or much that we should note.

Looking at the sex category of my data I wasn’t to surprised on what sex had more patients that had heart disease because we had a greater number of male patients to start with. But I plotted this bar graph by using the ‘crosstables’ functions again and only changing the labels title and axis labels. A screenshot of a cell phone

Description automatically generated(Figure 5)

As we can clearly see in figure 5 there are more males with heart disease than females and this is most likely because the sample size had more males than females to begin with.

As I approached the last two pieces of data that I had left to evaluate I had to change the bar graph in to a scatter plot for one set of data. I chose to do this was because I really wanted to see better if there was any correlation and it was better seen in the scatter plot than it was on a bar graph.to be gin with the first thing I did to make this work was have to figure out what my x and y’s were going to be which was age and the target (have disease or not have disease). After this I used the scatter plot function that was built in by ‘matplotlib.pyplot’ and I had to determine the x by indexing the age based on if the patient had or did not have a heart disease. Then almost exactly the same way I did the x I did the y I had to plot all the heart rates but I indexed them on if those patients had a heart disease or not as well.

A screenshot of a cell phone

Description automatically generated

(Figure 6)

As we look at figure 6 we can see that there was no distinct answer on whether or not heart rate affected the frequency to get a heart disease. Although there is no real clear answer of one thing affecting another we can definitely see that most of the patients that have a higher heart rate also have more of the heart disease meaning that on the graph we don’t really see any red dots in the bottom right corner but they are mostly higher up on the graph and right in the middle. This makes me conclude that people from around 40 to 60 with a high heart rate could possibly have a greater chance of having a heart disease.

As we look at the last data that we plotted I find this one to be quite shocking. Again like the first bar graphs we used the ‘crosstabs’ function to help build this almost exactly like the others just changing the labels, colors, and data that we used. Here in figure 7(below) I already have some knowledge on how high and low an average person’s blood pressure should be and that is why I was shocked to see the bar graph. Since I myself have pretty high blood pressure that needs to be checked often I hypothesized that the higher the patients’ blood pressure the more likely someone will have a heart disease because high blood pressure is not healthy.

A close up of a device

Description automatically generated

(Figure 7)

So knowing that an average person’s blood pressure should be around 120 or so I was shocked to see that the highest frequency of having a heart disease was in the 120’s and 140’s. this disproved my hypothesis.

All in all I have figured out a lot more unique things about heart disease but was not pleased about some of my data and I was sometimes disappointed with my failed hypothesis. I conclude that looking at how wide spread the data is and how there are no real single pieces of data that stands out I think at the end of this that anyone can get heart disease because in almost all of the data that I sampled there was someone that had a heart disease in each category. The main things I would definitely change to proceed onto do more testing would be to gather a lot bigger data set from everywhere across the world because I do not think there were enough patients examined. Another thing I could do moving on in the future is to have an even amount of men and women because this could also be throwing off my data and my solution. Lastly I would see if there were more test to view and do to the patients to see what more closely cause heart disease like if the patients parents or grandparents had any sort of heart disease then that would be something very useful to look at next time.