# Abby Willard

DASC 2113 Final Project

## Pre-Analysis

### Development of question/hypothesis

When deciding what our topic for this project would be, Bobby and I decided upon a heart disease dataset. We were between this dataset and a salary dataset of Data Scientists and other engineers, but we both agreed that the heart failure dataset had a good ratio of quantitative data and categorical data. The dataset also has a very natural interpreted question to answer. How does one acquire a heart disease (values 0 or 1 on HeartDisease variable) based on the variables provided in the dataset? From there, it was easy for the three of us to collaborate and create our project description and plan assignment. We all contributed for the collaboration of the assignment, but it was my main idea to do the decision tree and, with the mention of Dr. Cothren, the binary logistic regression.

### Data Research

Because I don’t have a great background in health terms, specifically heart terms, I did investigate research that has already been done to get an idea of how our variables would relate to the heart disease variable. I looked up all the terms I was somewhat comfortable with and terms I had never heard of and made my own data dictionary, looking up common values among healthy heart individuals and looking up what the variable actually means.

### Literature Review

Because our dataset is a conglomerate data set of 5 smaller datasets, I did find research done on the smaller, more specific datasets, and looked at those to again determine more about the variables, but didn’t want to spend too much time looking at those for worry that it would sway our approach.

## Analysis strategy and Analysis Code

### Exploratory Data Analysis

After collaborating, we all went about our own exploratory data analysis. I think that was important for us to do as I didn’t wany my own findings to persuade the findings of my partners. My initial EDA was in the form of creating pivot tables of different variables, descriptive statistics (count, mean, std, min, 25%, 50%, 75%, max) for all variables, histograms to determine the lay of the variables, and 3 pair plots to see how each variable has relations to the other variables. Of that initial analysis, the most descriptive EDA would be the pair plots, specifically the one that is colored by heart disease, as well as heat maps I created to show densities of the variables that are in positive heart disease and negative heart disease individuals. I was looking for similar clusters or notable relations from the pair plots and high densities in the heat maps to explore later. This EDA helped me

### Predictive Analysis

For my predictive measures, I tried the data through multiple regressions: decision tree, binary logistic regression, and gradient boosting regression. I also created multiple tests to evaluate how well the models were.

For the decision tree, I created the feature selection, the parts of the data that lead to the target (heart disease), split the data with train\_test\_split function from sklearn.model\_selection, and optimized the decision tree performance to create our final tree.

After creating the tree, I broke it down by evaluating it by confusion matrix, accuracy, precision, and recall, and then creating a classification report. All those evaluations came back to show high accuracy and thus, the tree classification came back successful.

For the binary logistic regression, I broke up (split the data), created the logistic regression, and fit it with the trained X and y. The binary logistic odd table came back with specific weights of the variables that equate to heart disease. Of those, old peak and fasting blood sugar had the most weight. By checking the accuracy of the prediction, it came back as 77% so a relatively good model. To evaluate its accuracy further, I again created a confusion matrix, classification report, and calculated the accuracy, precision, and recall scores, as well as the ROC and AUC curves for the regression.

Lastly, I created a gradient boosting regression, as that algorithm can be used to train models for both regression and classification problems and shows the weight of the variables. When I created it, it was an accuracy level of 30% so I did not proceed it further. I did try to increase the accuracy level by changing the values of the parameters of the gradient boosting regression, but I never obtained an accuracy above .445, so I still dropped it.

A breakdown of who did what: I did around 50% of the EDA published in our finalized group project, as well as all of the predictive measures in the group project.

## Code Review

As we shared code between each other, I was able to stay updated on what my partners were doing. If there was some of their code I was curious about, I would review it with them before or after class or text them in our group chat asking for greater analysis of what they found. Moreover, if they had a question, they would often find me after class for my input on their code.

## Work Planning and Organization

In our communication with one another, I stayed involved in letting my team members know what I am doing and helped administer setting up times to meet up to work on the project, finish our work, and rehearse our presentation. It was understood in the group that I would try to tackle the majority of the predictive analysis, so I allowed my group members to do more of the exploratory analysis. There were a few instances, were I had specific ideas for the project that I wanted to be accomplished that I knew I wouldn’t have enough time to do, so I forwarded those ideas to my partners, who both were interested in the ideas and were able to complete them for our project. For example, I gave them the idea of using hypothesis tests to determine the significance of our variables, and they executed it afterwards.

## Improving Teamwork and collaboration

## As I said above, I worked to keep the communication in the forefront of our team. However, communication was somewhat difficult for me as Bobby mentioned to me that he usually doesn’t check his phone messages, but tends to respond better on Microsoft Teams. So, usually when I needed to reach both of them, I had to use both text messages and Teams to get my idea across, which took some getting used to. As I worked primarily alone (and a lot over the thanksgiving break), I kept sending updates, so they knew where I was in my analysis. We met up a few times over the course of the project to go over who was doing what, checking in on progress, and asking for input from others in the group.

## Testing code and procedures

When testing our code, the majority of that remained as testing the accuracy of the predictive regressions. I accuracy tested all regressions to make sure they were a good fit for our data, and if they weren’t high in accuracy, I tried to increase its accuracy or dropped the test in general. For the decision tree I created a confusion matrix and classification report to get deeper analysis on the accuracy of the tree. As for the binary regression, I still had a confusion matrix, but I also used an ROC and AUC curve to get another view of the accuracy and it still remained with a high enough accuracy to be accounted for. Also, as mentioned earlier, we used hypothesis testing to check our procedures and ensure our findings were statistically significant.

## Writing Report

As I did my EDA and Predictive analysis, I formulated it heavily with markdown text to describe what each section was for and asked my teammates to do the same. I got my teammates work from their notebooks and took those to create our one cohesive report that was to be turned in. A challenge that arose when joining our work into one notebook, is that we all had different names for our dataframe when we read in our csv file. So, I went through all of their work and used the name of the file that I used for my EDA and Predictive applications (because I used it in more instances than they used theirs). I kept the markdown text in a cohesive manner throughout the report and tried to make sure all our code and analysis was thoroughly described and the introduction and concluding remarks summarized our findings. Lastly, I pushed our repository to git (once it was back running).