Problem Statement:

* An introduction to the problem: What is the problem? Who is the Client? (Feel free to reuse points 1-2 from your proposal document)
* A deeper dive into the data set:
  + What important fields and information does the data set have?
  + What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + What kind of cleaning and wrangling did you need to do?
  + Are there other datasets you can find, use and combine with, to answer the questions that matter?
* Any preliminary exploration you’ve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. You first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.
* Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

**I am interested in predicting with high accuracy whether or not a person will develop heart disease as it is the leading cause of death in the US. Knowing whether someone is likely to develop heart disease can ideally improve preventive measures. The clients would be health care practitioners. The goal is that if we can reliably predict whether somebody will develop heart disease in ample time, then the health care practitioner can help put in place a preventative plan.**

**I began by exploring the data. In this heart disease dataset I eliminated many features and ended with 20 relevant features and 5 possible outcomes. This dataset has a number of relevant features, such as blood pressure, heart rate, exercise induced heart rate, age, sex, and gender among other features. It’s important to note that this is a multiclass problem, so subjects in this data set are classified as 0 (no heart disease) or a 1,2,3 or 4 depending on the severity of heart disease.**

**With this dataset, there are some limitations. For instance, there is no information about genetic factors of subject. Literature supports that genetic factors are related to the development of heart disease. I also imagine that one’s diet and exercise regiment may have bearing on their likelihood to develop heart disease. Additionally, pre-existing heart problems or other health problems might be relevant predictors of heart disease. Therefore, while this data does have useful features of predicting heart disease but this dataset also has some limitations as it may not give a ‘full picture’ of predictors of heart disease. With this data set I cannot answer whether genetic factors, diet, history of exercise, and preexisting conditions relate to the development of heart disease.**

**To begin, I had to clean and wrangle the data. First, the data set had to be combined from 3 different dataset. Each dataset initially had 76 columns, many of which were blank. I tried to read each dataset into a pandas dataframe, and found that the dataframe was not shaped correctly – it had 9 columns instead of 76. Some of the columns were being put in as rows. The first bit of wrangling I had to do involved figuring out how to get my data into the right shape. Using io methods and the replace function, I was able to replace the ‘newlines’ with empty strings so that new rows weren’t formed for attributes 10-76. Next I removed 56 features of the data of which 51 contained no data at all and one contained the name of each subject. Finally I combined the 3 dataframes into one dataframe and labeled each feature column correctly.**

**Then, upon exploring the data, I found that many features contained the value ‘-9’ in numerous places. I ascertained that in this data set, the value ‘-9’ is akin to a value that was not collected. For instance, one cannot have a heart rate or blood pressure or -9, and therefore I knew that the -9 values indicated data that was not reported for some reason. Nearly every feature contains -9 values. I have a few options with negative values. If I’m worried that the negative values or missing values will cause errors with model fitting, then I can completely rows containing negative values. However, this will drastically reduce my dataset, and I do not yet know the effect that negative values will have on model fitting. Another option is to impute values. For instance, I could impute missing heart rate values with the mean heart rate, for example. However, this also does not seem like a great idea in a problem such as this where the actual heart rate could potentially be a significant predictor of heart disease. Over- or under- reporting a true heart rate by imputing the mean could therefore negatively impact the accuracy of our model. For now, I have replaced the ‘-9’ values with the value ‘0.’ I chose to do this simply so that the data makes more logical sense to me. Negative values can be hard to assign meaning to in this context whereas 0s are a clear indicator that the respective value is missing. Having replaced the negative values with 0s, I began preliminary analysis and model training just to see how the models act with the 0s. I will revisit the problem of the 0 values after initial model fitting.**

**I did find another heart disease dataset[[1]](#footnote-1), and while it may have some useful features, there are two problems with combining this dataset with the current dataset I am using: 1) We have no indication that the features collected in the new dataset were collected in the same way as the original dataset, and 2) The outcome of the new dataset is binary rather than multiclass. Variability surrounding data collection is an important factor to consider when combining datasets. If a feature (such as heart rate, for example) is collected by two different methods in two different datasets, such as via a wireless sensor vs recording electrodes analyzed by medical specialist, then it is plausible that the two heart rates will not be in agreeance with each other for the same individual. Therefore, to reduce an effect of variability during data collecting, I have chosen not to combine the new dataset with the original dataset. Secondly, the new dataset has a binary outcome. This will be difficult to combine with the original dataset. I could use it to only combine the ‘0’ outcome. However, my original dataset has unbalanced classes with far more ‘0’ outcomes then 1,2,3 or 4. Therefore, I have chosen not to combine the two datasets.**

**In terms of preliminary analysis, I began with a multiclass logistic regression. I implemented a train/test split of the data and cross validation. To determine the accuracy of the basic logistic regression I calculated the accuracy score and also looked at a confusion matrix and ROC curve. The accuracy score was quite low (0.4), and I could see a lot of errors from the confusion matrix. The ROC curve told me that the model was a lot better at classifying outcome 0 then it was at classifying outcome 4.**

**This preliminary analysis has told me that perhaps the negative values are affecting the model as the accuracy was only 0.4 and also that the imbalanced class sizes may be a reason that the logistic model is far better at predicting class 0 compared to class 4. Moving forward, I need to address these two issues. Additionally, I will test a decision tree model as it may be more accurate in classifying a dataset with many features. With decision trees, I will have to be careful of overfitting the data and will use hyperparameter optimization methodology to find the best parameters for a decision tree.**

1. <http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data> [↑](#footnote-ref-1)