Heart Failure Prediction

I have chosen, for my capstone a dataset from Kaggle that contains 11 features (Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Pressure, Resting ECG, Max Heart Rate, Exercise Angina, Oldpeak, St_Slope, and, ultimately, Heart Failure). I will use these features to help predict heart failure.

1. **Problem Identification**

Almost 18 million people per year are killed by cardiovascular diseases. This accounts for 31% of worldwide deaths. Heart Failure itself is a common event of those who have CVD's. The problem to be addressed in this project is, how can we use given health metrics (listed above) to create a model to help with early detection and management.

2. **Data Wrangling**

Since this is a dataset that comes from Kaggle, obtaining the csv file will not be difficult. From there I plan on going through and seeing how the data is laid out. I plan to look at the various columns and rows, check for how many data entries I have altogether and in each column, and look for what values are missing. Before I try to impute any missing values I should look at the rest of the included data to determine whether it should be imputed, or dropped.

3. **Exploratory Data Analysis**

During this step, I will still be looking at the quality of the data, but I will be looking for obscurities in the data. For instance, are there any instances where information is submitted in different units of measure (cm vs inches, etc.)? I will look for various trends/complex relationships within the data by utilizing different graphing and visualization tools. I will also look for trends/patterns in any missing data. Ultimately, I will be playing and 'getting a feel' for the data.

4. **Pre-Processing and Training Data Development**

During this step, I will be actually looking at how to impute any missing data. If any of the metrics contain different units of measure I will want to transform them all to match at this point. I will also be looking at how I impute any missing data (finding a median, mean, or other statistical value). If there are any categorical data columns containing strings, I will need to create a new column assigning the value of a number to a given category. I will also use cross-validation tools and model selection to make sure I accurately check the model performance.

5. **Modeling**

_____Here I will be working through the different models to determine what set has the highest level of accuracy, or whatever it is that I may be testing, in this case, the highest levels of correlation among various metrics and whether or not that individual experienced Heart Failure.

6. **Documentation**

_____During this final stage I will be taking my models and findings and summarizing them into one report. I will take my suggestions for early detection and management of Heart Disease and present them to the audience, while then going back and explaining in everyday terms what I was doing to get to that ultimate result. I will not highlight every single step in the project, moreso the thought processes behind a specific step.