Daniel Theng

Dr. Banuelos

MATH 191T

6 November 2020

Final Project Proposal

For my project, I would like to perform exploratory data analysis and create a classification model this [Cardiovascular Disease dataset](https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). The dataset has 70,000 entries with 11 features plus the target variable for the presence of cardiovascular disease. The data has 3 types of features; objective (age, height, weight, gender), examination (information from medical examination), and subjective (information from the patient). In this project, I hope to determine:

1. General trends between the features and the target variable.
2. Which features contribute the most towards cardiovascular disease.
3. Whether a classification model can accurately predict the presence of cardiovascular disease using the features given.

For the exploratory data analysis, I will use plotting packages to create pairplots to compare the features and also find any general trends. To determine which features contribute the most, I will use feature importance or feature coefficients for each model to measure the importance of each feature and plot the importance for each feature to determine which feature comes out on top. As for the models, I plan to use a logistic regression, support vector, and random forest classifiers. All of these models are from the sklearn package. Since the dataset only has 1 file for training, I will have to use the train\_test\_split() function to create a training and test dataset. I can imagine that the exploratory data analysis will take a long time, and given the amount of features it might be difficult to make any generalizations. In addition, some of the features in the data are categorical variables. From what I have read in order for a model to understand categorical variables they need to be encoded. While I have never had to encode categorical variables, there are possible solutions in functions like LabelEncoder and One Hot Encoder that are available in the sklearn package that make encoding easier.