**Mid-term Proposal**

**Title:** Heart Disease Prediction

**Introduction**

World Health Organization estimated there were 12 million deaths were due to heart disease worldwide every year. Half of the deaths in the United States and other developed countries are due to cardiovascular diseases. Cardiovascular disease is the conditions that involve narrowed or blocked blood vessels, by definition. Cardiovascular disease could lead to a heart attack, chest pain, or stoke. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and reduce the complications. This dataset intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using machine learning.

**Data**

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information including over 4,000 records and 15 attributes variables. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors. The dataset includes binary variables like the patient’s gender, smoke or not, whether the patient had a stroke previously, whether the patient was hypertensive, etc. The dataset contains categorical data such as education level and separates education levels into 1 through 4 representing from “elementary school” to “college or above”. There are numerical variables like clinical measurement from hospitals such as BMI (continuous), systolic blood pressure(continuous), heart rate(discrete), glucose(discrete), etc. The data types are very comprehensive for prediction modeling, and I will be able to apply different classification techniques on this dataset.

**Proposed Methods**

The dataset is mainly used for classification to predict if the patient will potentially be risky to have coronary heart disease in 10 years. The prediction will be binary: 1 means “Risky”, and “0” means “not risky”. Firstly, I will run "t-SNE" and linear plots for basic visualization. I will apply statistical tests with linear models by calculating p-values to see which variable is the most statically significant for the prediction. Then I will use logistic regression as my base model for classification. I will also apply the KNN, decision tree classifier, gradient boost, support vector machine, and random forest to compare the accuracies. Since the downside of this dataset is the size is small (only 4,000+ data). I am a little worried that the accuracy may not be high enough. Therefore, some ensemble machine learning techniques like bagging and boosting could improve the model accuracy. Reduce overfitting if needed. Neural network models will be additional-extension modeling. I don’t try to over-achieve, so I will see my workload and decide if I will model the neural network later.

**Timeline**

Data cleaning: The data is very clean. I will spend 1 week (week 4) for next-step cleaning.

Modeling/interpretation: This is the big part. I will start modeling/Interpretation from week 5 - 7.

Presentation work: Presentation and rehearsal will be done in 1 and half weeks (Week 8-9).