Final Project Proposal - IST 707

Jacqueline Parks, Laura Larregui, Courtney Rowland, Jacob Arluck

**PREDICTING HEART DISEASE**

**BACKGROUND / DATA MINING PROBLEM**

According to the Centers for Disease Control and Prevention, heart disease is the term given to a range of heart condition types. These diseases are often referred to as being silent because they sometimes are not diagnosed until the person experiences an event like a heart attack or arrhythmia. Since heart disease is the leading cause of death for both men and women of most racial and ethnic groups within the U.S., heart disease is a major cause for concern. Due to the large number of risk factors such as high blood pressure, cholesterol levels, smoking, diabetes, obesity, diet, physical inactivity, and excessive alcohol usage, it can be difficult to identify heart disease in individuals. The data mining problem is to create a model to predict whether target patients have heart disease based on the attribute subset.

**DATA SET USED**

The heart disease data set is from Kaggle but originated from the University of California, Irvine’s (UCI) machine learning repository. It is compiled from 4 databases, based on data collected in Cleveland, Hungary, Switzerland, and VA Long Beach. It contains 76 attributes but the project will focus on a subset of 14 attributes. There are a total of 303 instances formatted in a csv file.

Data Source: Naresh (2020). Health care: Data set on Heart Attack Possibility (Version 1) [CSV file]. Retrieved from https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility.

**INITIAL STRATEGIES**

*Exploratory Data Analysis*: (histograms, barplots, treemaps)

*Correlation Matrix*: Which predictor variables go together? Can inform colinearity assumptions for other strategies.

*Association Rule Mining*: This method is used to observe frequently occurring patterns.

*Decision Tree*: To predict potential outcomes.and create an easy-to-interpret model.

*Random Forest*: More advanced decision trees, reduce risk of overfitting.

*K-means Clustering*: Identify groups of patients with similar vitals. Are there clusters (diabetics, etc) who might need special treatment?

*Naive Bayes*: To predict the probability of future high risk patients for heart disease based on the data provided.

*Support Vector Machine*: SVM will be utilized to determine which variable combination generates the most parsimonious model for predicting which patients are at a higher risk for heart disease.