Capstone Project 1: Project Proposal

Heart disease is the deadliest condition amongst adult humans in the world; it was estimated to be responsible for more than 30% of deaths worldwide in 2015.[1] Despite this, heart disease is largely preventable through a healthy diet and exercise, and can be treated by lowering high blood pressures and cholesterol levels. In both cases, identifying the presence of heart disease early or identifying whether the patient is at risk of developing heart disease is crucial to maximizing their chances of survival. To this extent, I am aiming to develop a model to predict the presence of heart disease in patients from readily available medical information.

The data set I am using is the Heart Disease Data Set from the UCI Machine Learning Repository, which contains information such as patients' age, sex, blood pressure, cholesterol level, and heart rate, in addition to an indicator of whether the patient in question has heart disease on a sliding scale from 0 (no presence) to 4.[2] Because the data set is relatively small with 303 examples, I plan on using logistic regression at first to perform a binary classification between absence of heart disease (0) and presence of heart disease (1-4). The data set is also split by locale, with patients originating from Cleveland, Long Beach, Hungary, and Switzerland, so I plan on comparing heart disease manifestation between those regions by creating smaller models for each locale and also adding a feature to the complete data set that indicates where the patient is from. Further down the line, I aim to attempt a softmax regression on the data to predict the exact numerical indicator of heart disease in the patient, but this may be difficult with the small size of the data set.

This project would be of interest to doctors and other medical professionals who interact with heart disease patients regularly, as it would allow them to make a quick prediction on whether a patient has or is at risk of heart disease from simple medical information. Based on the results of the prediction, the doctor can determine whether additional tests such as CT scans or electrocardiograms are necessary if the patient seems to be at risk. These additional tests would normally not be performed at a normal check-up, but this model could help identify whether such tests would be necessary. Even if the patient turns out to not exhibit heart disease, the model prediction could indicate that active prevention of heart disease onset may be important.

The deliverables of this project will include the code used to prepare the data and create the model, as well as a paper outlining the results and some analyses of the data from the data set.

References:

1. GBD 2015 Mortality and Causes of Death Collaborators (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)*, *388*(10053), 1459-1544.

2. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Heart+Disease