Sheena Sharma – Capstone Proposal

1. What is the problem you want to solve

**I am interested in predicting with high accuracy whether or not a person will develop heart disease as it is the leading cause of death in the US. Knowing whether someone is likely to develop heart disease can ideally improve preventive measures.**

1. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

**The clients would be health care practitioners. The goal is that if we can reliably predict whether somebody will develop heart disease in ample time, then the health care practitioner can help put in place a preventative plan.**

1. What data are you going to use for this? How will you acquire this data?

**(https://archive.ics.uci.edu/ml/datasets/Heart+Disease)**

**I will be using data from the UCI repository. There are two options: 1) 4 ‘preprocessed’ CSV tables of 14 attributes or 2) 3 unprocessed data sets of 76 attributes. I am going to use Option 2. I’d like to peruse a large set of attributes to decide which ones are relevant or not.**

1. In brief, outline your approach to solving this problem (knowing that this might change later).

**First I need to figure out how to import the 3 datasets. The format is a little strange so I will need to first figure out how to get the data into tables of 76 columns. Next I need to merge the 3 data sets. Following this I will look for missing values or values that don’t make sense and will make a decision on how to deal with these values. Next I will explore the data to look for the relevant features. I will then explore different machine learning algorithms – maybe logistic or linear regressions would be appropriate here. I will also explore clustering as I think it is possible there will be ‘clusters’ of people more likely to develop one form of heart disease over another.**

1. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

**My deliverables are code along with slides outlining the mission statement, methods, and results. Additionally, I’d like to get out a probability of disease development for each person in the dataset.**