**STAT 4255 Project Proposal**

**Predicting heart attack**

**Members: Urja Desai and Akriti Bhattarai**

**Description of the problem:**

According to the CDC, heart disease is the number one killer in the United States; one in four deaths can be attributed to heart disease. This is because the most common form of heart disease, coronary artery disease, often leads to a myocardial infarction (heart attack) by blocking arteries that supply the heart tissue. The lack of blood supply eventually leads to cell damage and death. The heart can no longer effectively pump blood to the rest of the body. The heart can heal, but there are often permanent affects due to scar tissue. Some of the risk factors for heart disease are very common, like smoking or hypertension. Genetics and age also contribute to risk of heart disease.

Prior studies have attempted to predict heart disease using machine learning. Attributes that have been associated with heart disease include cholesterol levels, blood pressure, and age. Since some of the risk factors for coronary artery disease can be treated or controlled, predicting the presence of heart disease would allow for early intervention. This can reduce the risk of death and improve the quality of life of patients.

**Description of the dataset**

This data is the Heart Disease Data Set from the UCI Machine Learning Repository.

The original description of the dataset can be found here: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Within the data folder, there is the file “processed.cleaveland.data”. The data were uploaded by David W. Aha, who indicates that this is the correct file.

1. Age
   1. Continuous (in years)
2. Sex
   1. 0=female
   2. 1=male
3. Chest pain type (cp)
   1. 1=typical angina
   2. 2=atypical angina
   3. 3=non-anginal pain
   4. 4=asymptomatic
4. Resting blood pressure (trestbps)
   1. Continuous (in mm Hg when admitted to the hospital)
5. Cholesterol (chol)
   1. Continuous (cholesterol levels in mg/dl)
6. Fasting blood sugar (fbs)
   1. 0=fbs< 120 mg/dl
   2. 1=fbs>120 mg/dl
7. Resting ecg
   1. 0=normal
   2. 1=having ST-T wave abnormality
   3. 2=showing probable or definite left ventricular hypertrophy
8. Maximum heart rate achieved (thalach)
   1. Continuous (bpm)
9. Exercise induced angina
   1. 0=no
   2. 1=yes
10. ST depression induced by exercise relative to rest (oldpeak)
11. Slope of the peak exercise ST segment (slope)
    1. 1=upsloping
    2. 2=flat
    3. 3=down sloping
12. Number of major vessels colored by fluoroscopy (ca)
    1. Discrete, 0-3
13. Results of thallium heart scan (thal)
    1. 3=normal
    2. 6=fixed defect
    3. 7=reversable defect
14. Heart disease diagnosis
    1. 0=no heart disease (less than 50% vessel diameter narrowing)
    2. 1=heart disease (more than 50% diameter narrowing)

**Type of problem/method:**

This is a supervised classification problem as we have 2 outcomes, heart disease and no heart disease.

**Comments and concerns:**

Some of the medical terms are unfamiliar to us and will require some background understanding to make interpretations. This dataset has been used before, and others have noted that there are some missing values that we will have to deal with.

In the data itself, the heart disease diagnosis column takes values from 0 to 4. However, every study with this data uses 0 and 1 to describe the presence and absence of heart disease; values 1 to 4 are replaced with 1. In prior studies, the continuous variables are also converted to discrete variables. Choosing whether and how we should assign discrete values to continuous variables may be a concern. As of now, those are the concerns we have.