MATH4570 - Project Proposal - Outline:

Group Members: Chia Hsu, Alec Bulkin, Angela Wu, Emily Gu, Brendan Clarke, Bridget Foote

Data set Link:
https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

- What is the application field of research?
  - The most broad field is healthcare, more specifically cardiovascular health.
- Why is the problem important?
  - Heart disease is the number one leading cause of death in the United States and the most common type is coronary artery disease. Coronary disease or CAD which is a major cause for heart attacks. Hence, identifying factors that directly lead to heart attacks is fundamental to creating preventative measures.
- What questions will you try to answer?
  - Identify highest risk factors for having a heart attack
  - Identify where preventative care may be necessary
- What will be the main work for your project?
  -
- What the published literature suggests about your problem/plan
  -
- Your proposed timeline and division of labor
  - Process data & build model (Chia, Alec)- analyze error, predictions, etc. (1 week)
  - Write paper (Modeling, Analysis (Bridget), Simulations (Brendan), Conclusions (Angela)) (2 weeks)
  - Presentation (Alec) (1 week to make slides + 1 week to practice)
- Talk about what data sources you plan to use, including the number of samples and number of features.
  - Data source is “Heart Attack Analysis & Prediction Dataset” from Rashik Rahman on Kaggle. There are 303 data points with 10 features.
- Explain what methods and approach you are planning to use and why.
  - Logistic regression

Identifying Heart Attack Risk Factors

By: Chia Hsu, Alec Bulkin, Angela Wu, Emily Gu, Brendan Clarke, Bridget Foote

Heart disease is the number one leading cause of death in the United States and the most common type is coronary artery disease (CAD), which is a major cause for heart attacks. Hence, identifying factors that directly lead to heart attacks is fundamental to creating preventative measures. The goal of our project is to identify the risk factors that are most highly correlated to having a heart attack so to identify where preventative care may be necessary. We are planning to use the “Heart Attack Analysis & Prediction Dataset” from Rashik Rahman on Kaggle which has 303 data points and 10 features. We are planning on using a logistic regression model because this is a classification problem with two distinct classes, the subject had a heart attack or the subject did not have a heart attack.

We have divided the work for this project into three phases. First is to build the model, then analyze the results and write the paper, and finally to create and practice the presentation. Chia and Alec will work on cleaning the data and building the model first which we estimate will take approximately one week. Then Bridget, Brendan, and Angela will work on analyzing the outcomes of our model and write the paper which we estimate will take approximately two weeks. Finally Alec and Emily will create the presentation slides based on the contents of our model and paper and then we will all meet to practice, which we estimate will take approximately 2 weeks in total.