**Goals**

Healthcare is a field that requires precisions in the judgment about a patient’s health; a wrong estimation can cost both life and money. One predictive challenge in the healthcare field is understanding if an individual is prone to have a heart attack or not based on prior health problems and demographics. This particular project can be useful for health officials to potentially identify and take action for an individual who is more prone to having a heart attack. A potential predictive model for this can save lives. As you identify people who are more prone to heart attacks through the model, they can possibly receive earlier treatment to reduce the chances of having a heart attack.

**Forecasting Variable(s) of Interest**

The variable of interest here is the *target* which indicates whether an individual has had a heart attack or not. This is represented as a binary variable where 0 = Heart Attack hasn't happened, 1 = Heart Attack has happened.

The prediction model will help in classifying individuals in the target categories, based on their past medical history and other related variables. Helping to direct the correct treatment plan for more heart attack prone patients.

**Observation**

Row in the Dataset: The data set has 303 observations capturing the demographics and past medical history of the patient.



Unit of Analysis: Target (1 = Heart Attack Occurred, 0 = Heart Attack didn’t occur)

**Correlates**

We plan to use variables associated with the demographics and health records of the individual, which include:

1) Age

2) Sex

3) Chest pain type (4 values)

4) Resting blood pressure

5) Serum cholesterol in mg/dl

6) Fasting blood sugar > 120 mg/dl

7) Resting electrocardiographic results (values 0,1,2)

8) Maximum heart rate achieved

9) Exercise induced angina

10) Oldpeak = ST depression induced by exercise relative to rest

11) Slope of the peak exercise ST segment

12) Number of major vessels (0-3) colored by fluoroscopy

13) Thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

We believe that these variables allow us to better understand what the patient’s prior health history is like as well as how their environment, i.e. their demographics, play a role in their risk of a heart attack. These variables are comprehensive in considering both the patient's health as well external factors which can play a role in the occurrence of a heart attack and hence we think they will be helpful in predicting whether a heart attack will happen or not.

Some of the health readings like blood pressure, cholesterol, fasting blood sugar, etc are medically known to be highly correlated to the occurrence of heart attack and will be good variables to help classify patients better.

**Data**

The dataset is from Kaggle: (<https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility>), and is originally taken from UCI Machine Learning Repository. There are 303 rows and 14 columns in the dataset. In terms of data preparation, there is not a lot of cleaning needed, as there are no missing rows in the dataset. This dataset is only for one of the locations (Cleveland) from the UCI repository, so we are attempting to extract the data for the other 3 locations as well to get more data.

**Plan**

We will first start with data preprocessing which we hope to accomplish by Week 5 of the quarter.

The models we considered so far include:

* KNN
* Logistic regression
* Decision tree
* SVM
* Neural Networks