UC Berkeley MIDS
W200 Fall 2021 Project 2
November 16, 2021
Submitted By: Jinsoo Chung, Sunitha Haraginadoni, Haibi Lu
Team Github Repo: https://github.com/UC-Berkeley-I-School/Project2_Chung_Haraginadoni_Lu

# Project 2 Proposal
# What Causes Heart Disease?

## Introduction:

Heart disease was the number one leading cause of death in the United States in 2019 and accounted for over 23% of all deaths [1]. Coming in many forms, such as coronary artery disease, the most common form of heart disease in the US, heart arrhythmias, and heart valve disease, heart disease disrupts the quality of life and creates major health complications in the affected population [2]. In order to effectively prevent and manage the disease, we need to understand common factors that contribute to the disease. In this project, we will be examining multiple variables that may be related to heart disease and see if we could establish a prediction algorithm based on the available sample data set.

## Dataset:

**https://www.kaggle.com/fedesoriano/heart-failure-prediction**

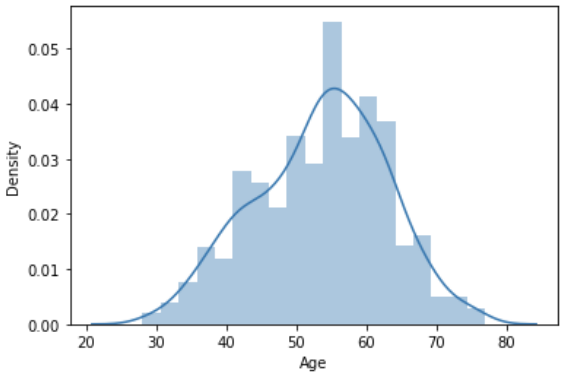
## Initial Data Exploration:

This dataset has 918 rows and 14 columns. The initial data exploration found 0 missing data.

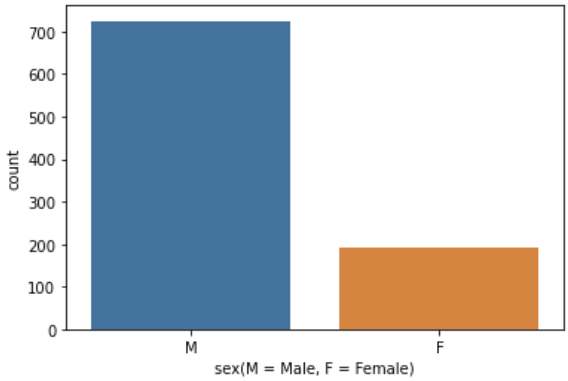
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
```

**Key Variables:**

1. Age:
   Use Seaborn package to plot the histograms of age distribution.
   a. Average age: 53.5
   b. Min: 28
   c. Max: 77
   d. Median: 54




2. Sex:
   The gender ratio between male and female is roughly 3.5: 1




3. Chest Paint Type
   a. TA: Typical Angina
   b. ATA: Atypical Angina
   c. ASY: No Symptoms
   d. NAP: Non-Angina
4. Resting BP
   a. Resting Blood Pressure
5. Cholesterol
6. Fasting BS
   a. Fasting blood glucose: (> 120 mg/dl=1; 0=false)
7. Resting ECG
   a. electrocardiogram (0=normal, 1=abnormal ST-T wave, 2=according to Estes standard showing possible or definite left ventricular hypertrophy, severe condition)
8. Max HR
   a. Maximum Heart Rate
9. Exercise Angina
   a. angina pectoris caused by exercise (Y=yes; N=no)
10. Old Peak
    a. ST-segment inhibition caused by exercise relative to rest.
11. ST Slope
    a. ECG of peak exercise
12. Heart Disease
    a. Heart Disease (0=no, 1=yes)

## Approach:

- Conduct initial distribution exploration on the demographical variables, such as age or sex, to see if we have any biases in our sample
- Determine correlation between heart disease and the independent variables listed above. This will help determine the factors that may contribute to heart disease
- Attempt to create a prediction model for heart disease based on the data set using Machine Learning
    - We will split up the data into 60% training, 20% validation, and 20% test set.
    - Potentially use the random forest classifier (decision tree) based Machine Learning model through scikitlearn to create a prediction model
    - Optimize hyperparameters & evaluate results on validation data set and test data set

## Summary:

Heart disease is the major cause of disability and premature death that is rapidly increasing in both economically developed and underdeveloped countries. This damage can be reduced considerably if the patient is diagnosed and treated in the early stages. It is difficult to manually predict the odds of getting heart disease based on the risk factors.In this project heart diseases are predicted by considering major risk factors like Age, sex, RestingBP, Cholesterol , ExerciseAngina with four types of chest pains. Machine Language prediction modeling is done using Scikit-learn which is simple and efficient for predictive data analysis. Healthcare providers can use this data to predict heart diseases and help to prevent and manage it effectively.

Reference:

1. https://www.cdc.gov/nchs/data/nvsr/nvsr70/nvsr70-09-508.pdf
2. https://www.cdc.gov/heartdisease/about.htm