Deliverable 1
Marwan Jabbour
Due: October 5, 2020

Data Set: https://www.kaggle.com/ronitf/heart-disease-uci

Justification: The data set contains enough parameters (14) to allow me to me predict heart disease. These include: sex, age, chest pain type, resting blood pressure, fasting blood sugar, etc. It also contains enough instances of patients (303).

Data Preprocessing: The dataset I chose is feasible and reliable. It is based off data from the Hungarian Institute of Cardiology. There seems to be no duplicate examples and no omitted values. The feature labels are good and there is no inconsistency. The 14 variables provided are useful, as they are relevant to the prediction of heart disease (blood sugar, cholesterol, etc.) If I found a row with missing values (I didn't find any so far), I'll simply delete the row, as long as there is no bias introduced. The columns are all numerical, so I don't need to encode the data. Also, I am given a relatively equal number of patients with and without heart disease, so my predictions would not be biased.

Machine Learning Model: I want to be able to predict the presence of heart disease from the variables inputted by the user. An alternative model would be one similar to that provided by the (http://www.cvriskcalculator.com)l; however, my data does not include whether or not the user has diabetes or is a regular smoker, as that would overcomplicate the training process.

Evaluation Metric: My project is a classification problem: **Does the patient have heart disease?** The 14th column of my data set tells whether or not the person has heart disease (1) or does not have heart disease (0). So, I will evaluate my problem with metrics from category 2 (confusion matrix and accuracy/precision-recall/logistic loss).

Final Conceptualizations: My project will be relatively easy to showcase as a website similar to the one provided above. Users will be asked to input simple data, such as age, sex, blood sugar, etc. and, after clicking a button, they will receive an assessment on the possibility of heart disease.