Building a Model to Predict Heart Disease

Fergus Xu and Rohan Pandey

**Research Questions**

1. Which demographics are more important to consider (have a higher correlation) when diagnosing cardiovascular disease?
2. Which model, when passed in some features of a patient (such as demographic, age, symptoms, etc.), provides the highest accuracy prediction of contracting the disease for that patient?

Feedback: Try to use/build a very basic baseline model, using the dataset, and then keep filtering and continue

1. How do our results compare to traditional models for evaluating risk? Is there a statistically significant improvement over non-machine learning models?
2. What biases and limitations are present within our model and the data that we have available? How can we possibly improve our model or methodology in order to reduce bias?

**Motivation**

Cardiovascular diseases are a group of disorders that affect the heart and blood vessels. Together, they are the leading cause of death globally. However, most can be prevented by addressing behavioral risk factors, and early detection is important so that management with behavior and medicine can begin as early as possible. By analyzing past data, we hope to shed light on which factors are the most important to consider and develop a model that can be used to help predict which people are most at risk.

**Data Setting**

Datasets:

* https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
* https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction
* https://archive-beta.ics.uci.edu/dataset/45/heart+disease
* https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive
* https://archive.ics.uci.edu/ml/datasets/heart+Disease

These datasets contain information about different patients containing different health benchmarks and the presence of certain predictors that can be used to train a model. Before using our data, we may need to clean the dataset in order to remove outliers of certain physical features in case they were incorrectly recorded (such as impossible body measurements), or that they impact our dataset in a way that would not accurately represent the rest of our data. Additionally, it may also be helpful to combine certain columns into more useful metrics, such as considering bmi over height and weight independently. With regards to the context of our data, the context is not provided for all datasets. However, for the ones that are provided, they are sourced from developed countries. Additionally, these populations tend to skew towards white, so the data likely does too, which ultimately may lead to bias within our model since we are not able to account for race based factors such as genetic risk factors or economic based factors.

**Challenge Goals**

1. Machine learning
   * One of the goals of this project is to establish a machine-learning model that can be used to predict whether the patient has or is at risk of contracting the disease. In order to do so, we will be training several different models and optimizing them using cross validation and regularization and comparing them using measurements such as the F1 score and AUC to determine which one provides the most accurate assessment. Additionally, through establishing a model, we hope to shed some light on which factors are the most important to consider.
2. Multiple datasets

* This project will aim at collecting multiple datasets and using the data to build a more accurate model. And with multiple sources of data, we are able to collect new features and potential causes of heart disease that our model wouldn’t have access to without more datasets. This also means we will have to filter out data that isn’t as useful or as necessary for our model.

**Method**

Create a method to read the data and create a Pandas DataFrame with relevant columns for the project.

Create methods to plot the data targeting the research questions that have been created, and through inspection of the generated graphs, we can begin our analysis process and answer the questions.

Then create a method that filters the data even more to just the factors that correspond to the disease and using the new dataset, build several machine learning models that use different algorithms using ScikitLearn to determine the best working prediction model.

If time permits we will provide an interface to the client who will be able to answer certain questions/input their personal data into our model and receive an accuracy rating.

Lastly, we’ll create statistical tests to compare and verify the validity of our predictions.

**Plan**

* Find and filter through the initial dataset, and create data frames for analysis. ~ 2 hours
* The first one will have the challenge of finding a good data set, so it’s important to work together if needed.
* Use Seaborn, or Matplotlib to create plots to answer some of the research questions above. ~ 1 hour
* Then continue filtering and find the most directly correlated causes/symptoms for the disease. ~ 1 hour
* Lastly, to build the machine learning model, we’ll sort through the symptoms and build models by feeding it lots of training data sets (we’re not exactly sure how many but it’ll definitely be more than 100 helpful data values). ~ 3 hours
* Build a patient class and instantiate it as an object and use the machine learning model to predict the accuracy scores of the prediction values of the sample patient! ~ 2 hours
* Using our model’s predicted values, and the values from our dataset, use statistical tests to verify the validity of our predictions. ~ 1 hour

**Development Environment**

VSCode