Ali Husain
MATH 4323, Fall 2024, Part II: Data Selection
11/14/2024

Link to dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

The data set linked above is from the website Kaggle, one of the many resources that provides high quality datasets. The dataset itself revolves around heart failure and predicting it based on certain attributes like cholesterol or beats per minute of the patient's heart. Cardiovascular diseases affect a wide range of people so, trying to predict the factors that could cause and subsequently try to prevent them by knowing what causes it is a huge motivation of mine. CVD is also present in my family and most definitely in many others so researching something like this is very meaningful.
Regarding the dataset, we have a healthy mix of both categorical and numerical variables to ensure our dataset works with multiple testing methods. We will also be doing supervised learning as we get provided the result of each patient.
To surmise, we are preforming a supervised learning project with a classification task. Our data set has well over 300 observations (around 917) with both categorical and numerical variables (12 variables to be exact).
Below is each column with a brief description partly provided by the author of the dataset:

- Age: age of the patient measured in years
- Sex: sex of the patient M for Male, F for Female
- Chest Pain Type: chest pain type that a patient is experiencing, TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic
- Resting blood pressure (RestingBP): resting blood pressure measured in mm Hg (millimeters of mercury)
- Cholesterol: Cholesterol levels measured in milligrams per deciliter
- Fasting blood sugar (FastingBS): fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- Resting electrocardiogram results (RestingECG): Measured in a categorical [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- Maximum heart rate (MaxHR): maximum heart rate that the patient had, numeric value between 60 and 202
- Exercise Angina (ExerciseAngina): exercise-induced angina Y for Yes, N for No
- Oldpeak: oldpeak represented by ST a Numeric value measured in depression
- Slope of ST (ST_Slope): the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: down sloping]
- HeartDisease: if the patient has heart disease 1 for yes heart disease 0 for No heart disease

Looking at the variables, we clearly have both categorical and numerical variables for this dataset. Cholesterol and Resting blood pressure are both examples of numerical variables with the former being measured in mm/dl and the former measured in mm Hg with Sex and HeartDisease being examples of categorical variables. Yellow highlights numerical with gray highlighting categorical.