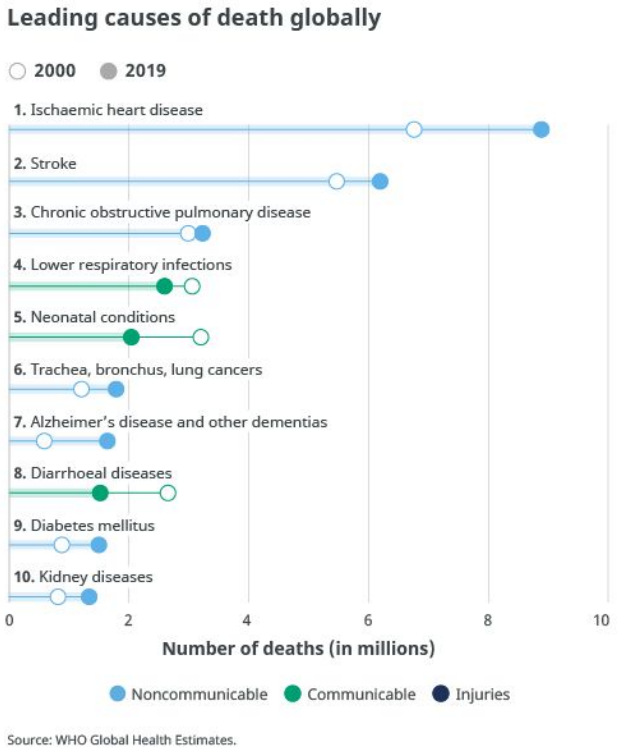
Heart Failure Prediction Analysis

By Bala Kandikonda

The Problem

Deaths caused by Heart Failure have been rising over the past 20 years and account for the biggest share of world wide deaths

Can we predict if an individual is more susceptible susceptible to Heart Disease?



The Data

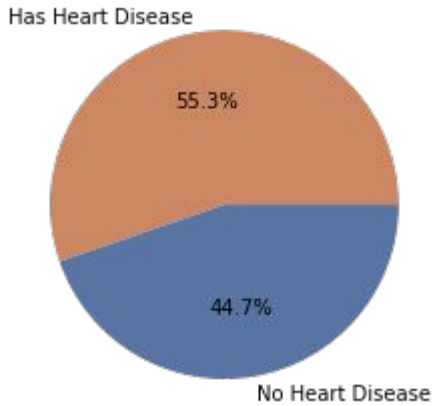
Dataset: Compiled from numerous studies done around the world, initial 918 observations and 12 features

Features in the dataset:

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [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]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: output class [1: heart disease, 0: Normal]

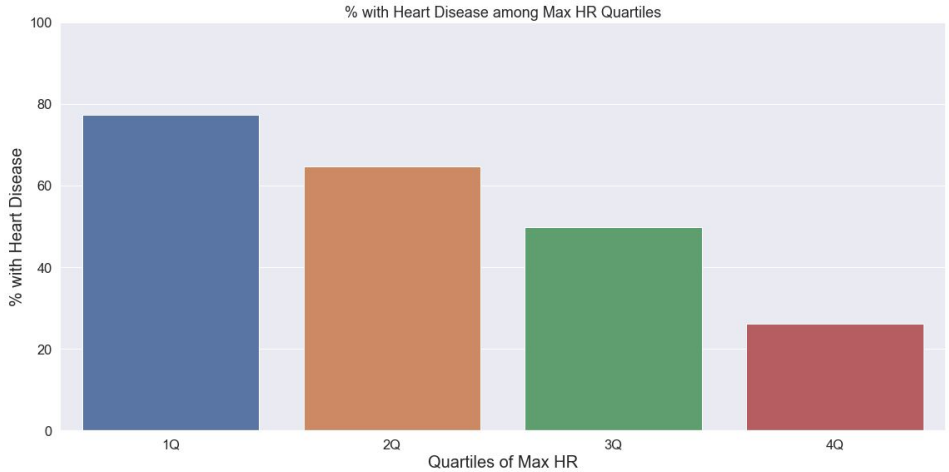
Exploratory Data Analysis

Classification Breakdown of Data



The dataset was almost evenly split with participants having and not having heart disease

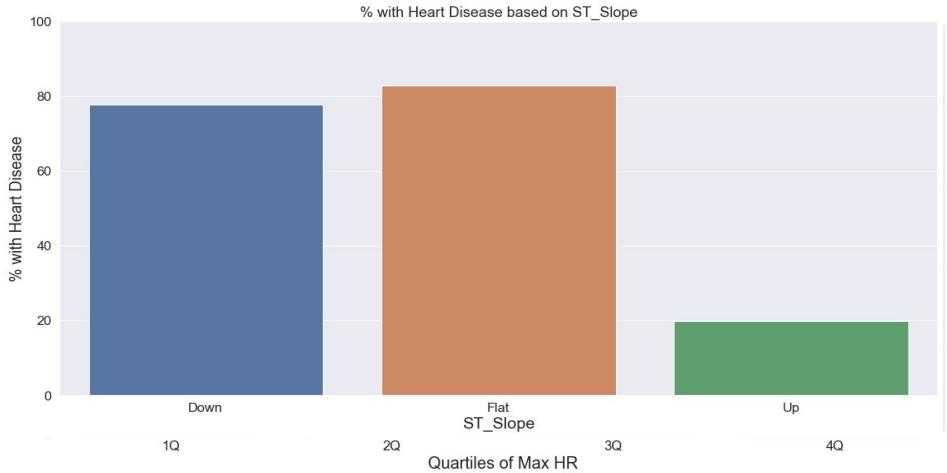
Max HR Breakdown



Max HR of an individual has a direct indication of whether or not they have Heart Disease

This was true later in the Feature Importance section of the analysis, interesting finding to research further beyond this analysis

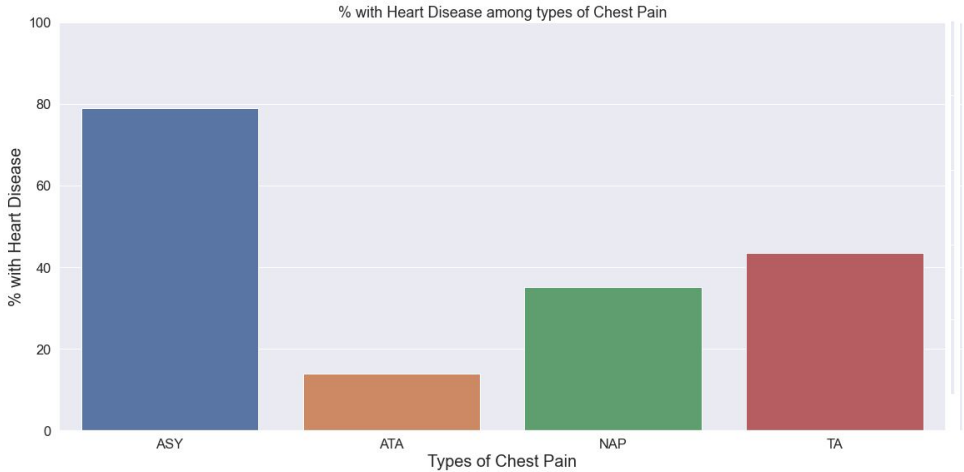
ST Slope Breakdown



The ST Slope is one of the most important features for predicting Heart Disease.

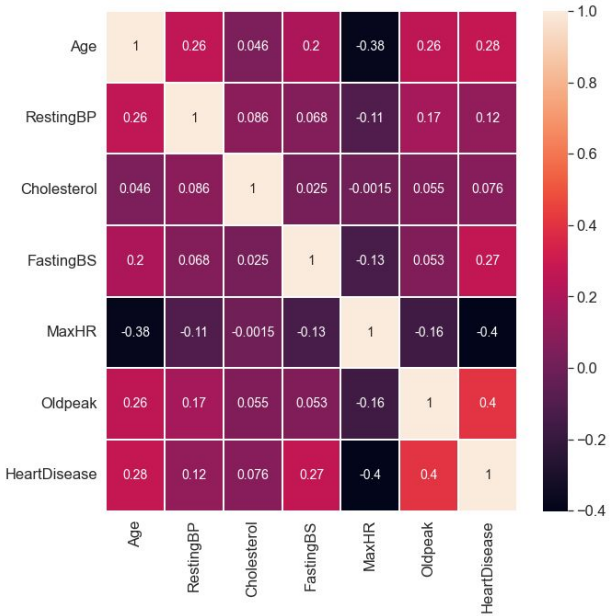
Participants with Flat and Down ST Slopes are much more likely to have Heart Disease vs participants with an Up ST Slope

Chest Pain Breakdown



Participants who have or are experiencing Asymptotic Chest Pain are significantly more likely to have Heart Failure than participants who experience any other type of Chest Pain

Correlation Matrix

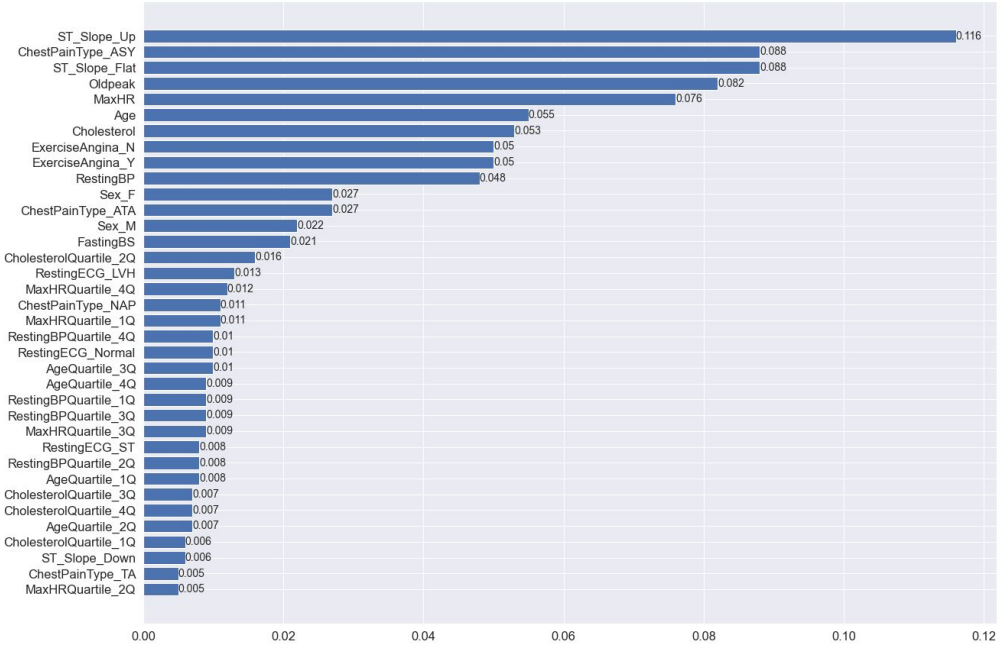


The Correlation Matrix for the quantitative features in the data set didn't show many strong relationships positively or negatively among the features

2 of the most important features in the dataset:

- Max Heart Rate and Heart Disease have the strongest negative correlation at -0.4
- Oldpeak and Heart Disease have the strongest positive correlation at 0.4

Feature Importance



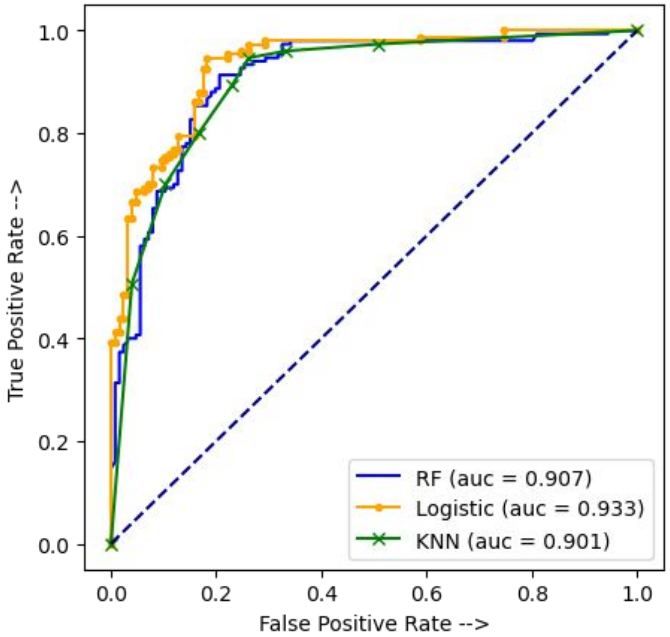
ST Slope, Chest Pain Type, Oldpeak, Max Heart Rate, and Age were the most important features for accurately predicting Heart Disease/Failure

Modeling

Models Used

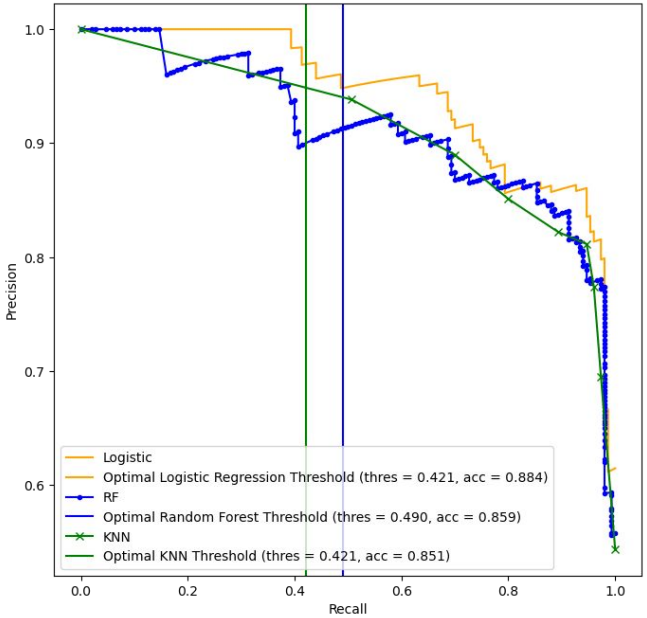
- Logistic Regression Classification
- Random Forest Classifier
- K-Nearest Neighbors

ROC-AUC Curve



The Logistic Regression model performed the best out of the 3 using the ROC-AUC curve, with an AUC of 0.933.

Precision-Recall Curve



Logistic Regression wins out again with the Precision-Recall curve with the highest accuracy

Logistic Regression and KNN have the same optimal threshold

Takeaways

- The most useful features at predicting Heart Failure are:
  - Chest Pain Type
  - ST Slope
  - Max Heart Rate
  - Oldpeak
- Logistic Regression model is the best ML model for this dataset and use-case for predicting Heart Failure

Further Research

- The relationship between Max Heart Rate/Resting Heart Rate and Heart Disease should be further studied
    - Lower Resting Heart Rate levels are a strong indicator of not having Heart Disease
    - While lower Max Heart Rate is a strong indicator of having Heart Disease