Heart Failure Prediction Dataset

Problem statement

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management.

Source Data

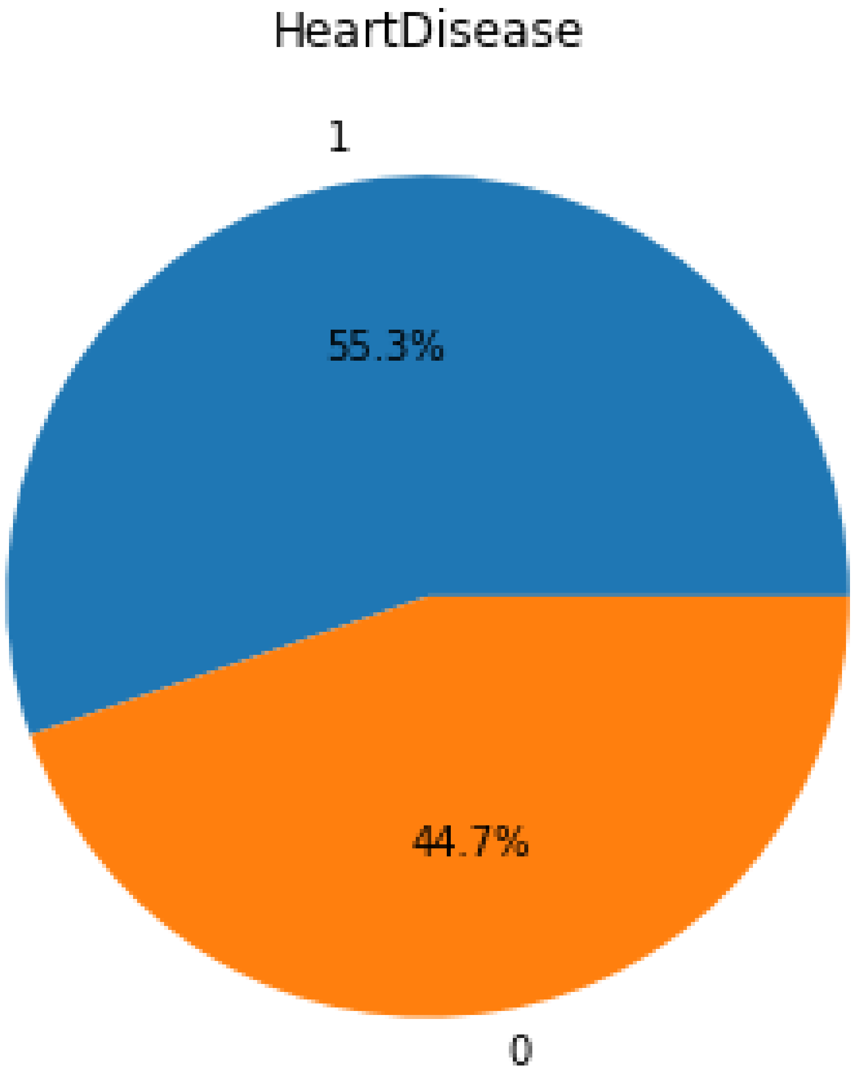
This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

Cleveland: 303 observations
Hungarian: 294 observations
Switzerland: 123 observations
Long Beach VA: 200 observations
Stalog (Heart) Data Set: 270 observations
Total: 1190 observations
Duplicated: 272 observations

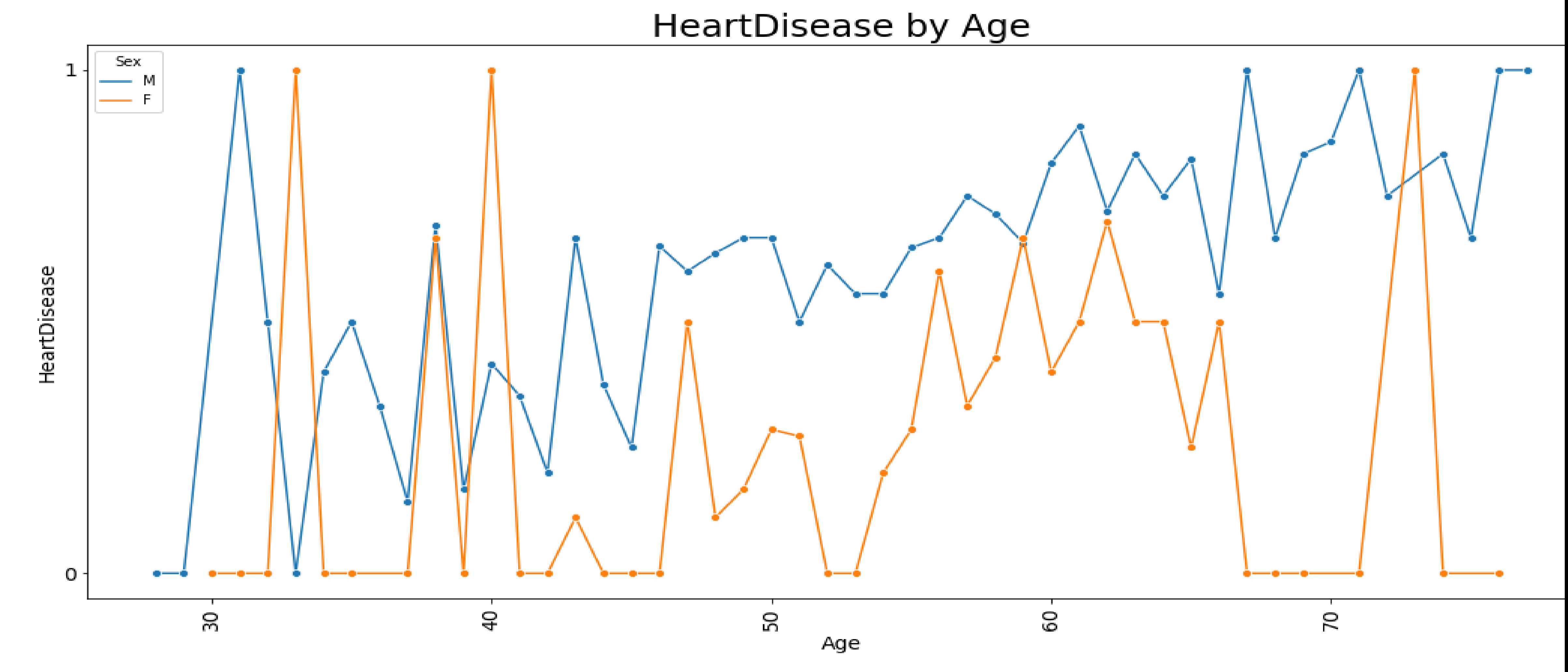
Final dataset: 918 observations

Attribute Information

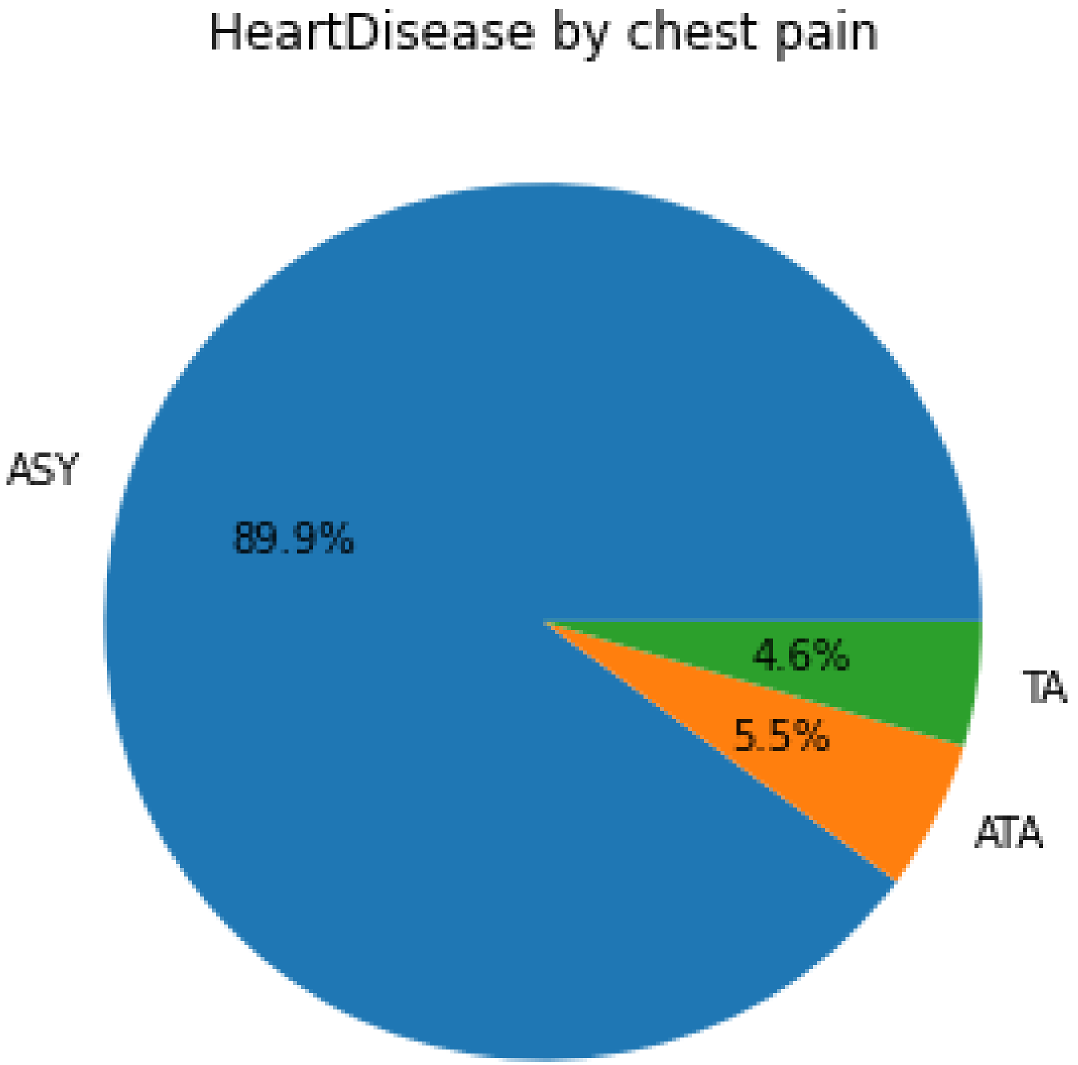
1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. 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]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]



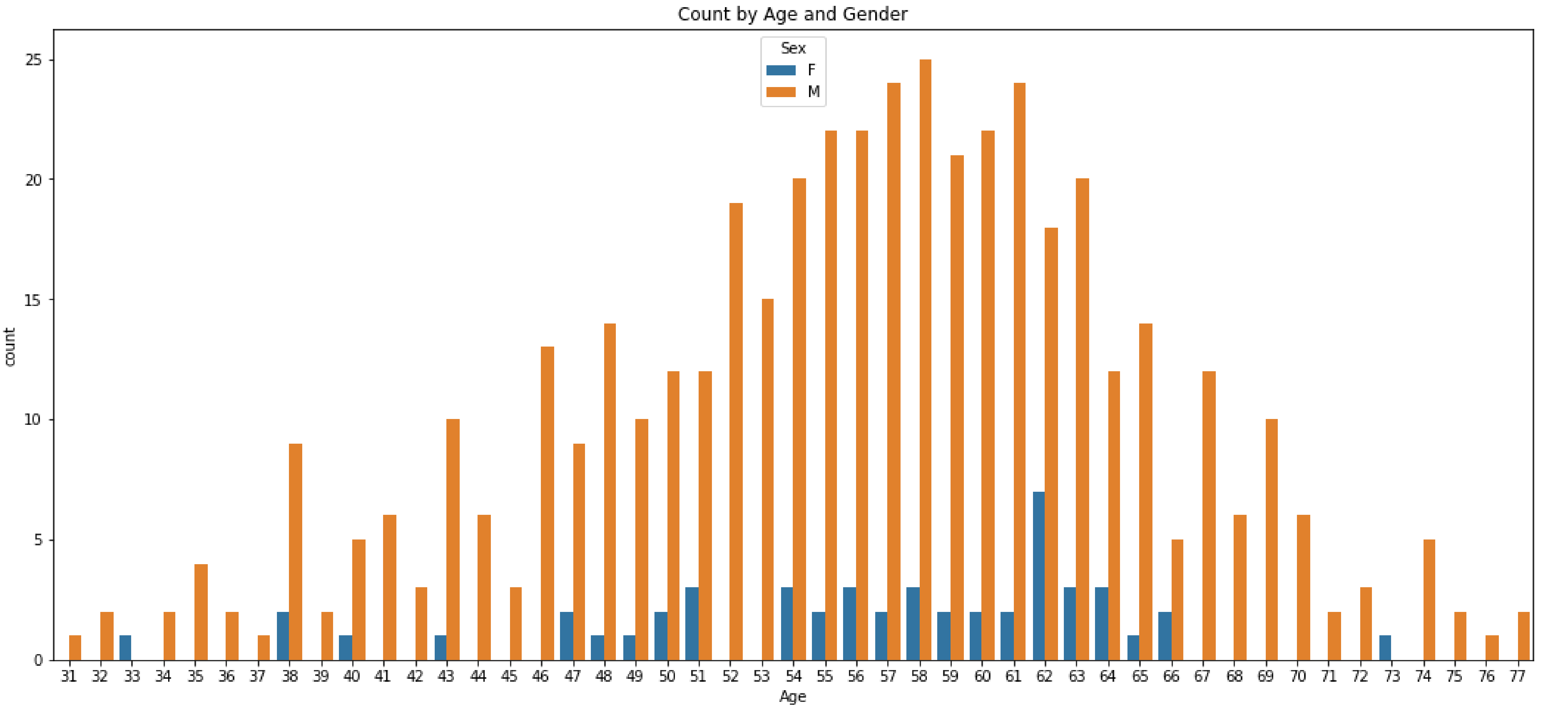
55.3% people are at high risk
44.7% people are normal



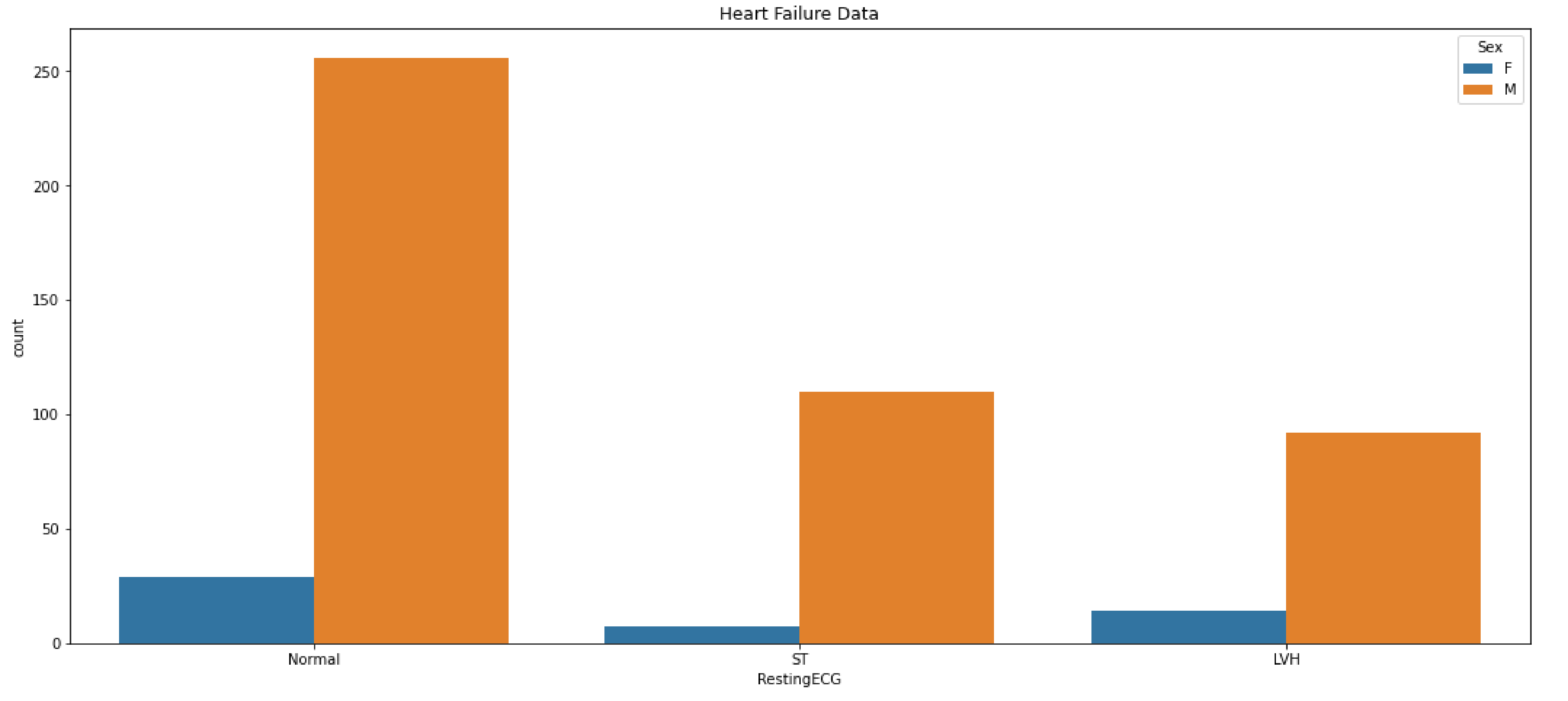
The above line graph shows the heart disease by age for men and women. It seems like men are at higher risk for heart disease than women. Women are more in danger between the age of 30 and 70.



Out of all people at risk 89.9% are having Asymptomatic chest pain. 5.5% are having Atypical Angina and 4.6% are having Typical Angina



Above graph shows the number of people by age and gender who are at high cardiovascular risk.



Resting electrocardiogram results by gender. This graph shows the number of people having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) and LVH (showing probable or definite left ventricular hypertrophy by Estes' criteria)

Models:

Logistic Regression model

False negative rate – 0.14% (16) with an accuracy of 87%
Tuned the model with a decision threshold of 0.3 and the results as below,
False negative rate – 0.083% (11) with an accuracy of 89%
L2 tuning – 0.14% (19) with an accuracy of 87%
Note: Results are better with the change in decision threshold when compared to L1 & L2 tuning

KNN Model

False negative rate – 0.12% (16) with an accuracy of 88%
At best K value – 0.091% (12) with an accuracy of 90%
Hyperparameter Tuning – 0.11% (14)
Tuned the model with a decision threshold of 0.3 and the results as below,
False negative rate – 0.038% (5) with an accuracy of 89% ← Best score

AdaBoost

False negative rate – 0.18% (24) with an accuracy of 85%
Hyperparameter Tuning – 0.18% (24) with an accuracy of 84%

XGBoost

False negative rate – 0.15% (20) with an accuracy of 87%
Hyperparameter Tuning – 0.14% (18) with an accuracy of 87%

Results & Conclusion:

After evaluating various model performances, the best performing model I would recommend is KNN model. In this project my target is to decrease False Negatives and make it equals to zero. KNN model worked best at a decision threshold of 0.3 with accuracy of 89%. False negative count went down to 0.038% which is better than all other models trained.

In the final model, false negative rate decreased as we anticipated but this increased the false positive rate. This trade off is acceptable because it is better to have false positive cases and proactively provide the precautions to the patients than ignoring the false negative cases. Here we don’t want to let go potential patients undetected.