Predicting Heart Disease

Paul Lipska
Coding Dojo
Project 2 part 5

Introduction

Heart Disease is the leading cause of death second only to Cancer in the US.

According to CDC.Gov One person dies every 34 seconds in the US from cardiovascular disease. In 2020 that roughly equated to one out of every five deaths.

From 2017 – 2018 the financial cost of heart disease was estimated to be $229 billion annually.

I will be reviewing my investigation of the kaggle.com dataset on this topic as we discuss ways to identify early signs of heart disease

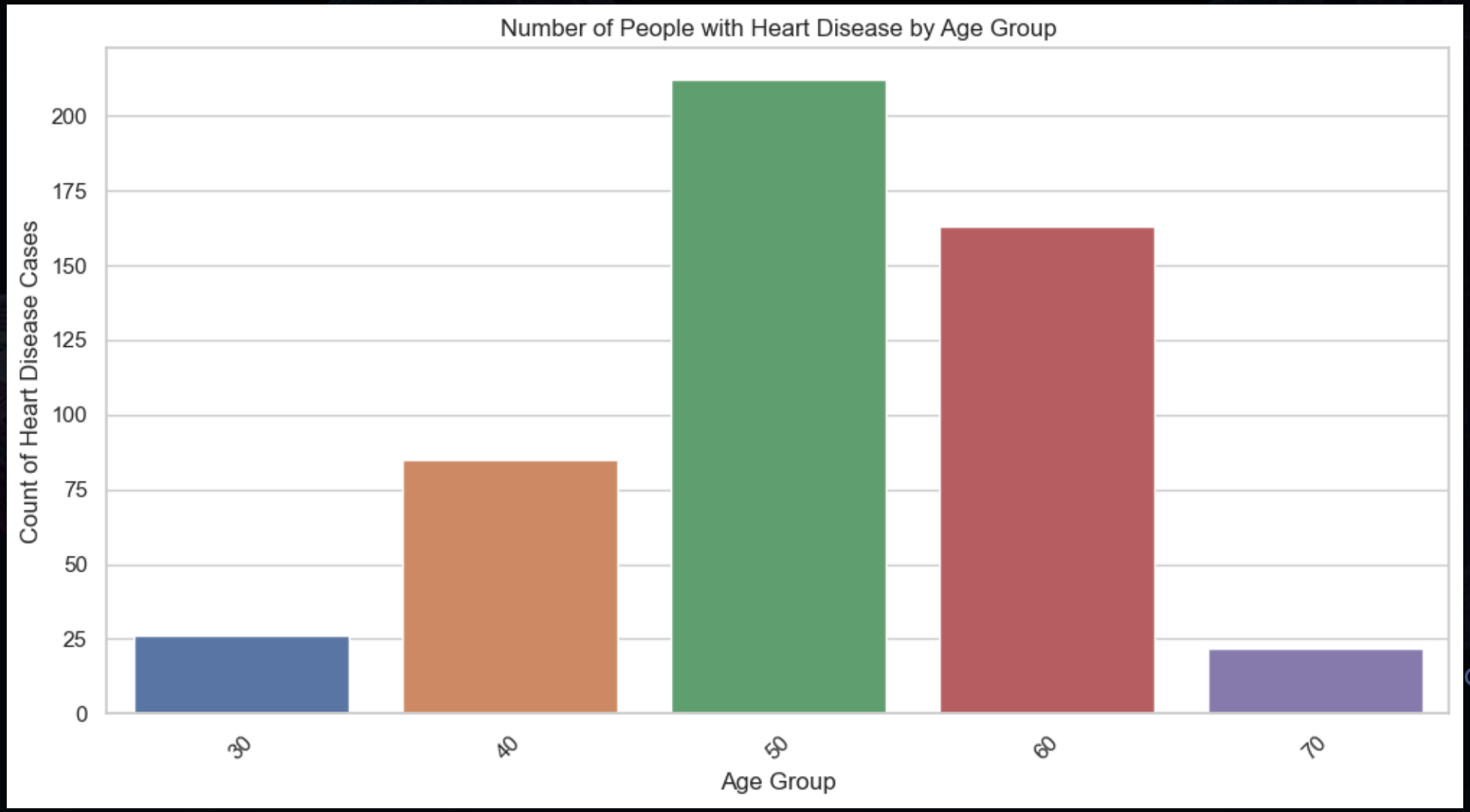
Stakeholder for this investigation would be Caregivers and Patients. The objective is to better identify the onset of Heart Disease

https://www.cdc.gov/heartdisease/facts.htm

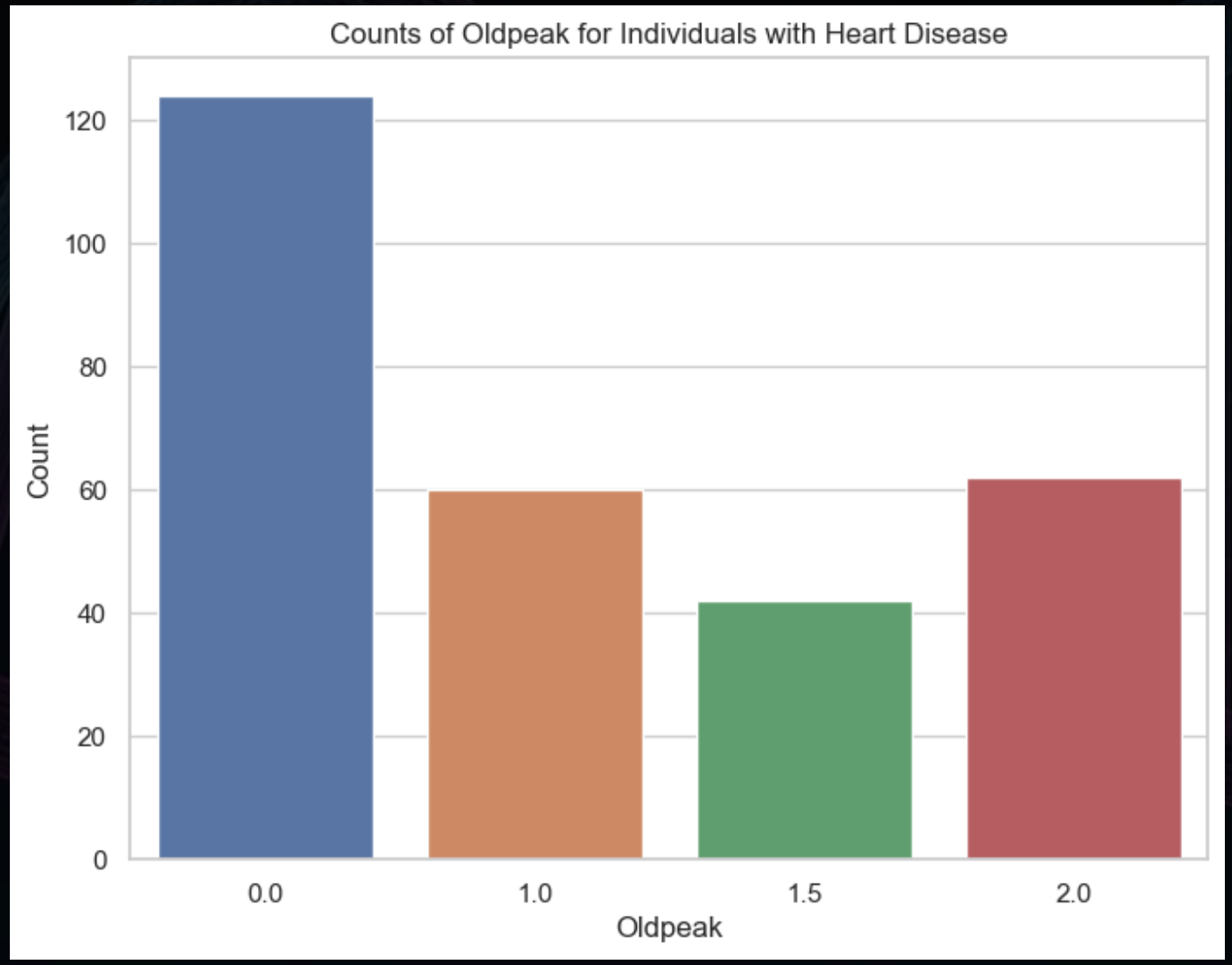
Dataset Description

- 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)

Heart Disease By Age Group



Old Peak Values with Heart Disease

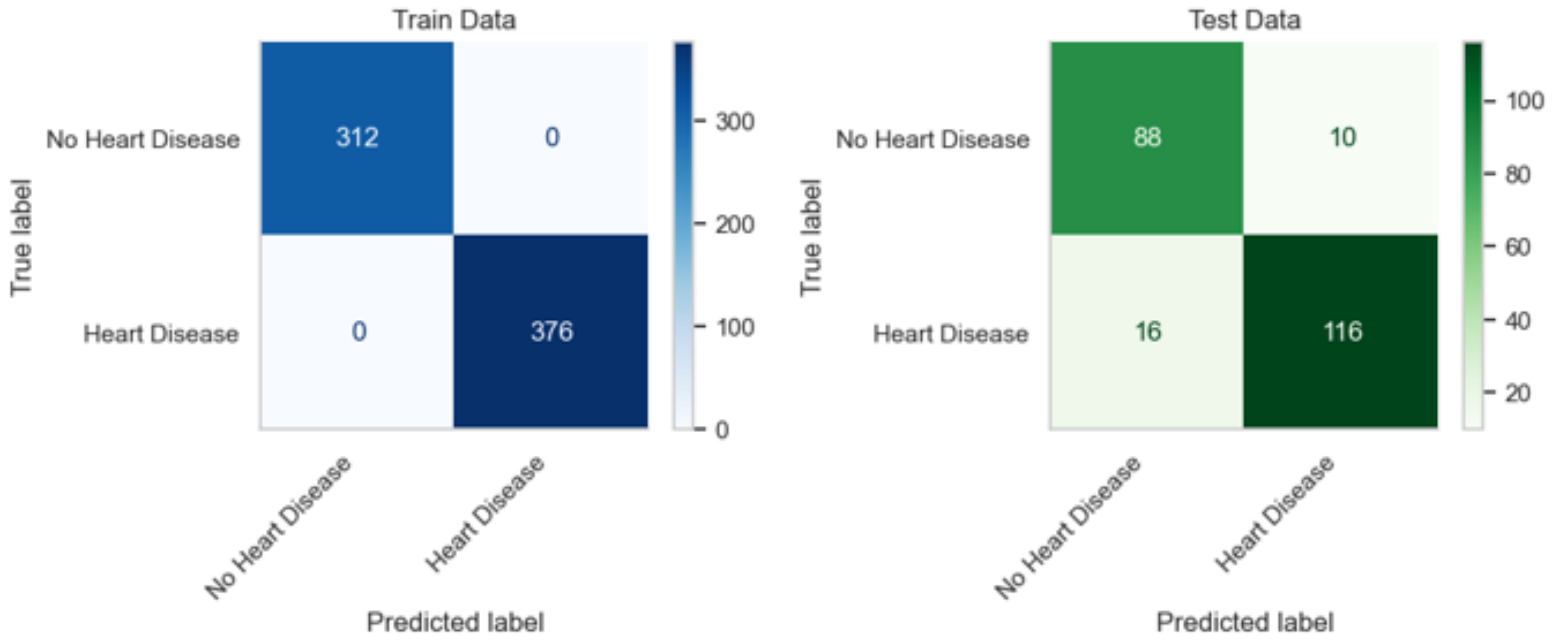


Methods for Training Data

Random Forest

Random Forest before training:

```
Accuracy score on the Random Forest train set:  1.0
Accuracy score on the Random Forest test set:  0.8869565217391304
```

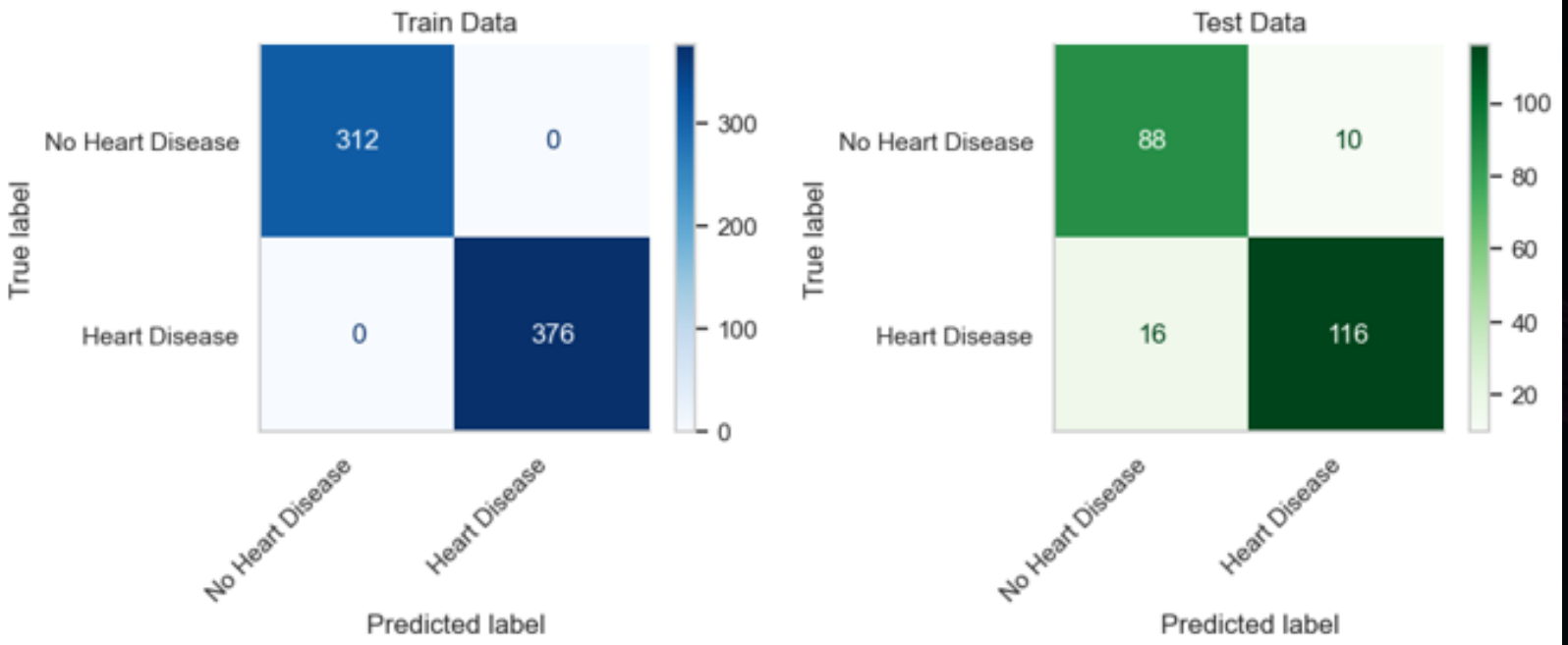


Methods for Training Data

XGBoost

XGBoost before training:

```
Accuracy score on XGBoost train set:  1.0
Accuracy score on XGBoost test set:  0.8869565217391304
```

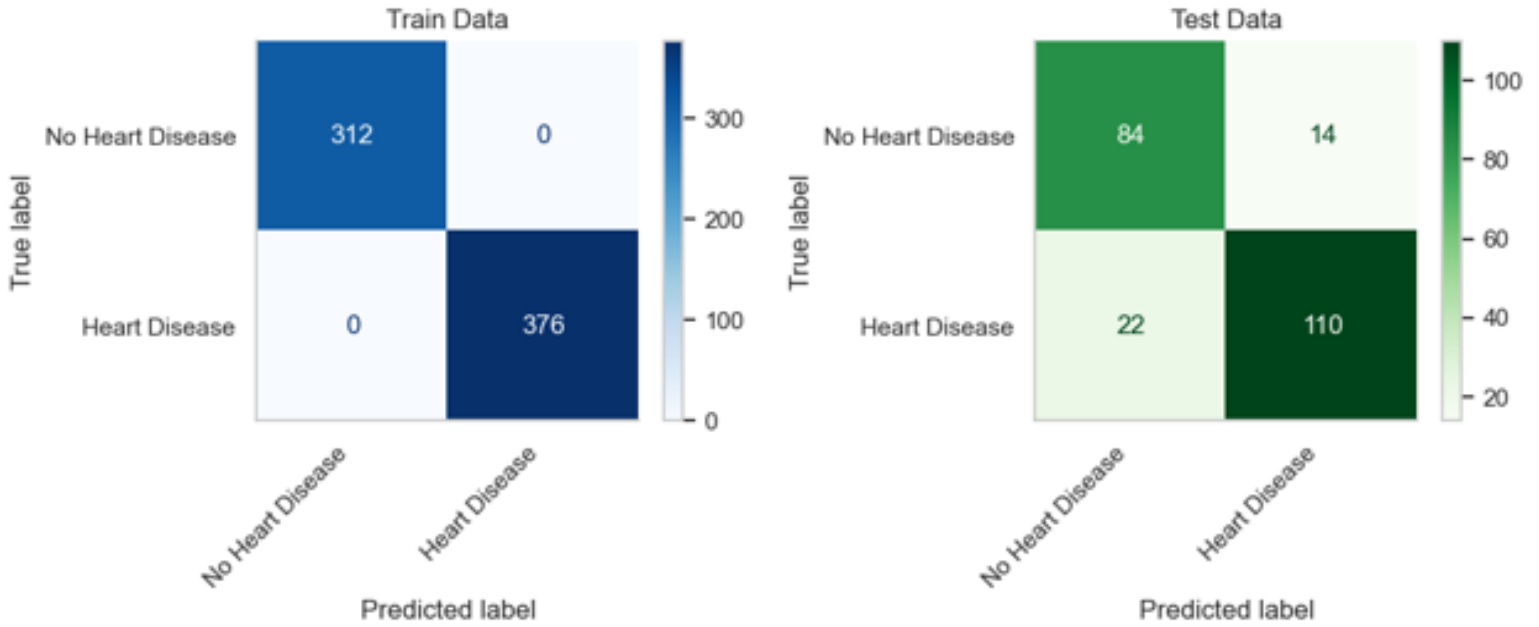


Best Model

XGBoost with PCA

XGBoost with PCA:

```
Accuracy score on XGBoost with PCA train set:  0.9200581395348837
Accuracy score on XGBoost with PCA test set:  0.8913043478260869
```



Study Observations

The two factors of age and Oldpeak have a strong correlation with the occurrence of Heart Disease and can be further examined to better understand the onset.

This study was heavily weighted towards patients in their 50's and did not take into account genetic or hereditary disorders that may have skewed the data or provided greater insights.

This was a categorical investigation into how to use clinical data to detect the onset of heart disease and inform preventive care.

https://www.cdc.gov/heartdisease/facts.htm

Final Recommendations

Given the weight of the study and the exclusion of race, genetics, and heredity I would recommend a larger sample size composed of equal numbers of people for each age group.