Final Report: Heart Disease Prediction
Springboard Capstone 2 | Aakash Shah

1) Introduction

a) Dataset: Heart Disease dataset

The heart disease dataset was taken from the UCI's ML Dataset archive from Kaggle. The 13 attributes of the heart disease dataset follow below:

1. age
2. sex
3. Angina - chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Going in depth on some features:

- Age: Overtime, the risk of heart disease increases surpassing the age of 55
- Angina: Correlation between where the pain or discomfort lays with the age or a person or fifty-five. Average rate of people ages 55-65 more likely to experience some type of pain.
- Resting Blood Pressure: Heart disease also deals with any high blood pressure such as diabetes and high cholesterol.

This dataset was already prepared and cleaned exception to a duplicate row which was found and removed in data wrangling.

b) Problem Statement

With leading averages on heart disease staying at a high rate and leading cause for death for both men and women.

c) Background

In this dataset, we were presented with 302 patient attributes who might have been suffering from heart disease. Out of seventy-six parameters, only fourteen features chose because they were classified to be the most important ones which was directly related to heart disease.

2) Project Goal

The project goal is to provide a more thorough understanding how and which features impacted the death of heart disease for both men and women.

3) Data cleaning & Wrangling

For this Data cleaning and wrangling, we have created a Pandas HTML overview report, which has illustrated the general overview, warnings, and the reproduction for some of the models. In the overview, it gives us all of the key information such as number of variables, number of observations, missing cells (in which case was 0 in the dataset), number of duplicate rows, and the memory size. It then broke down into each of our 14 individual features to give us more depth of the data we are looking at.

The Pandas Profiling Report also has shown us the four types of correlations of heat maps to have a better understanding of the correlations and outliers.

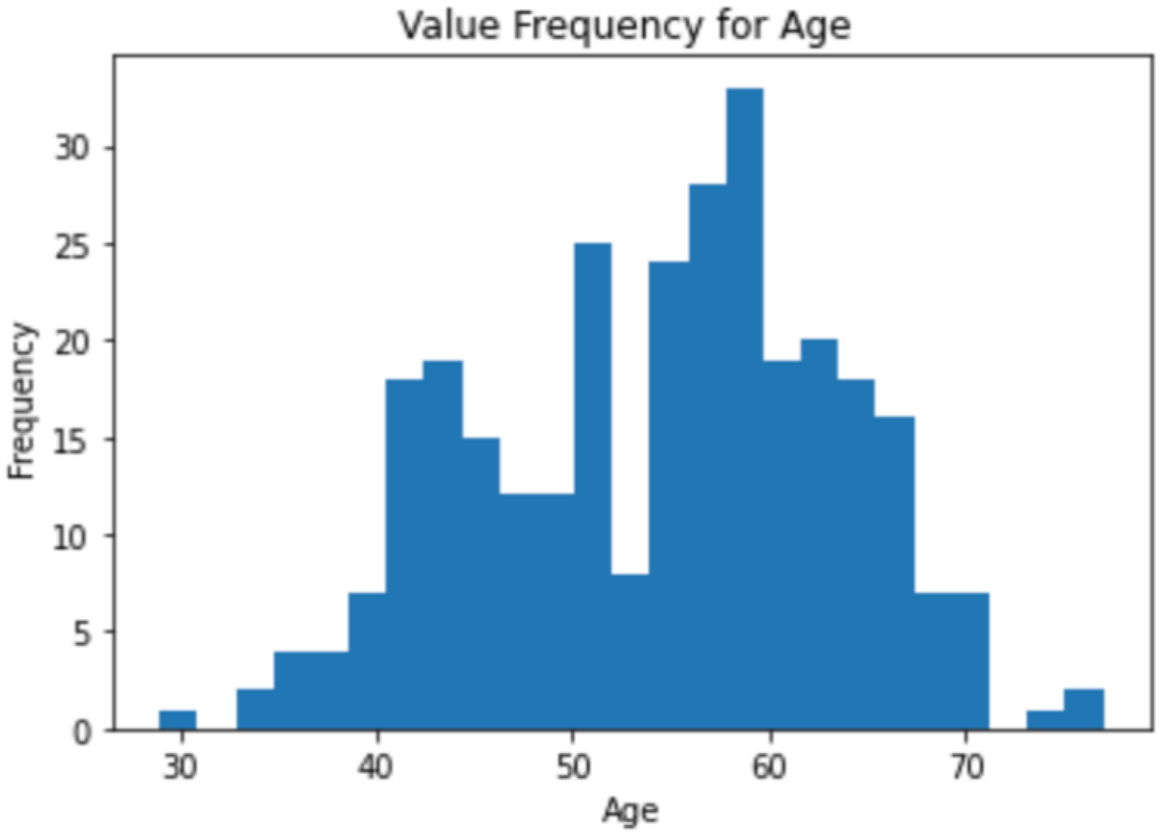
In terms of cleaning the data, we removed the duplicate row of age 38. As stated earlier, most of the dataset was clean and ready to use.

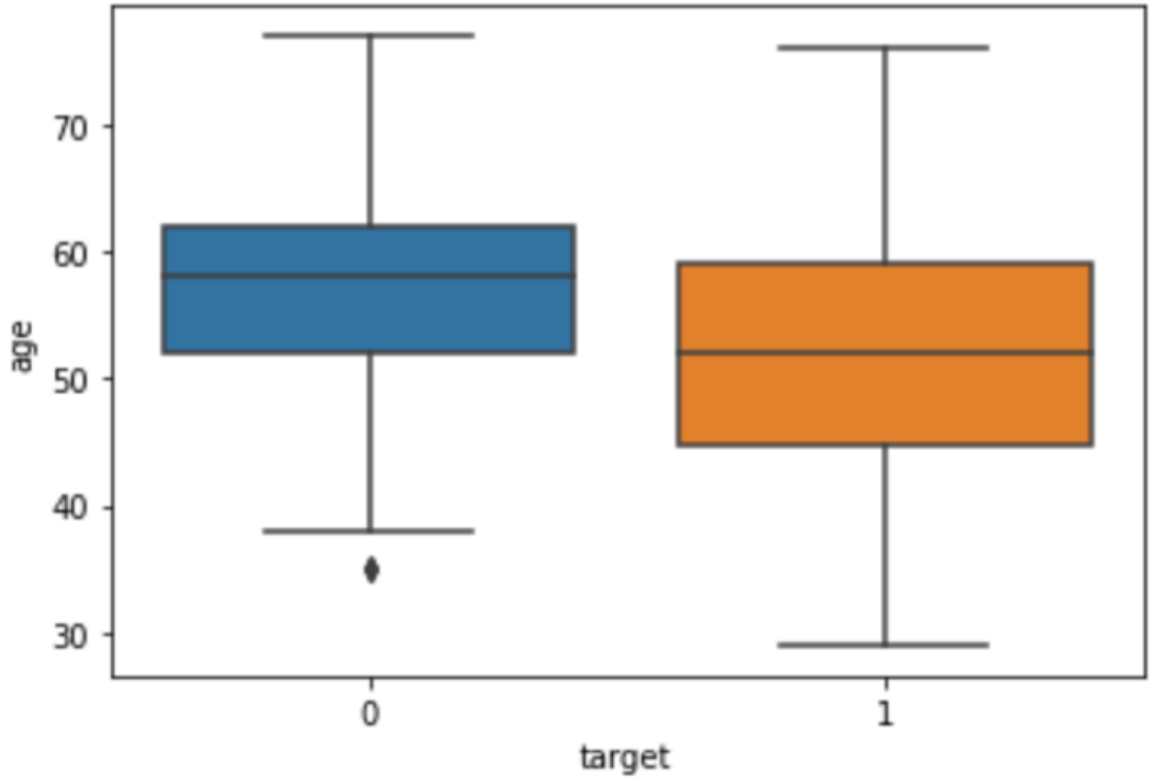
4) Exploratory Data Analysis

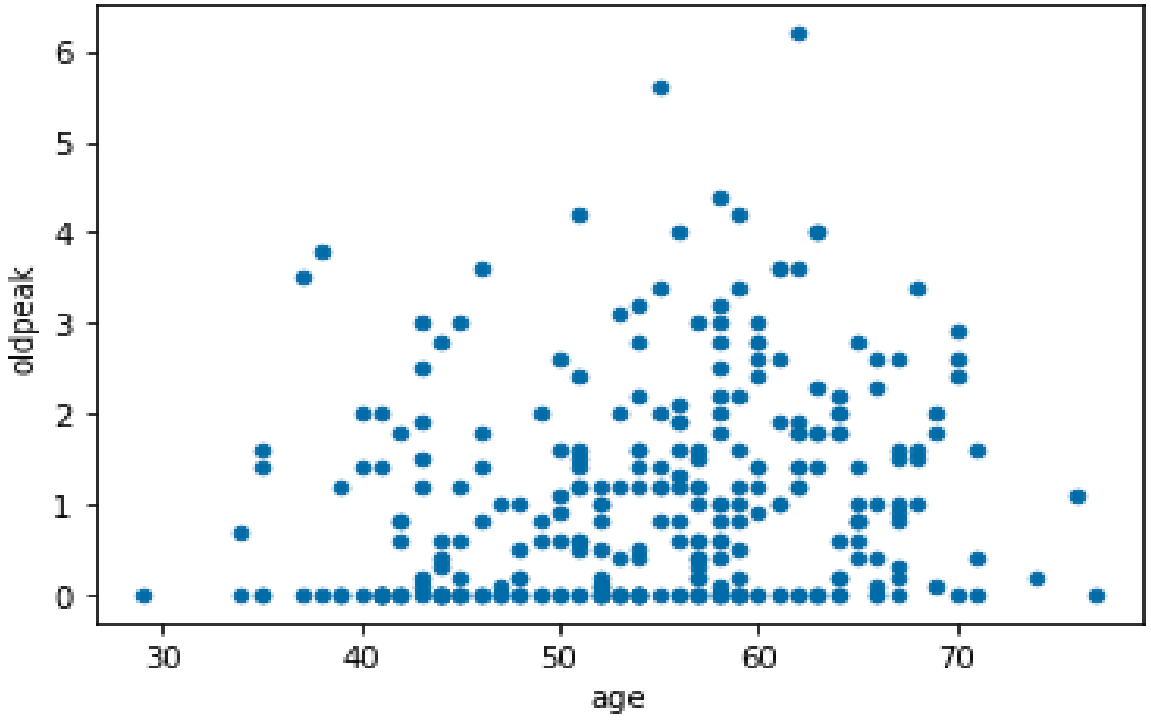
Starting off for the Exploratory Data Analysis, I was looking for various trends and correlations in the data, which would help identify those key features which directly impacted the risk of heart disease.

a) Age- frequency, mean, and scatterplot

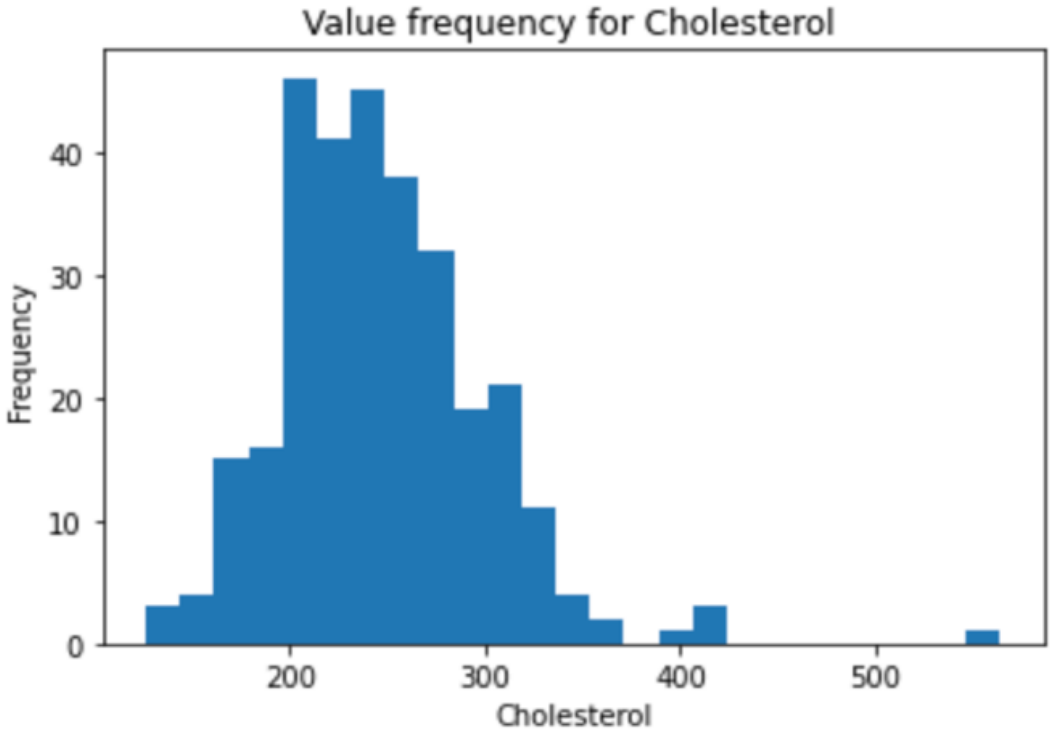
We see the distribution of the value frequency on age much higher on the peak between the ages 55 and 65. This shows that the median is pretty close in terms of the second figure that is shown in the boxplot. We do have small outliers around the age of 30 and surpassing the age of 73. The 3rd image: the scatter plot suggests that the peak exercise is on a positive slope, indicating that there is positive stress.

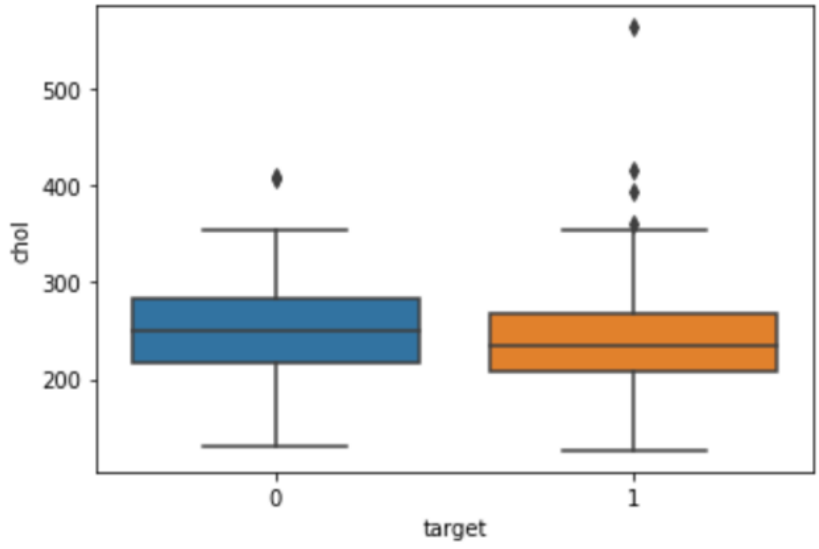


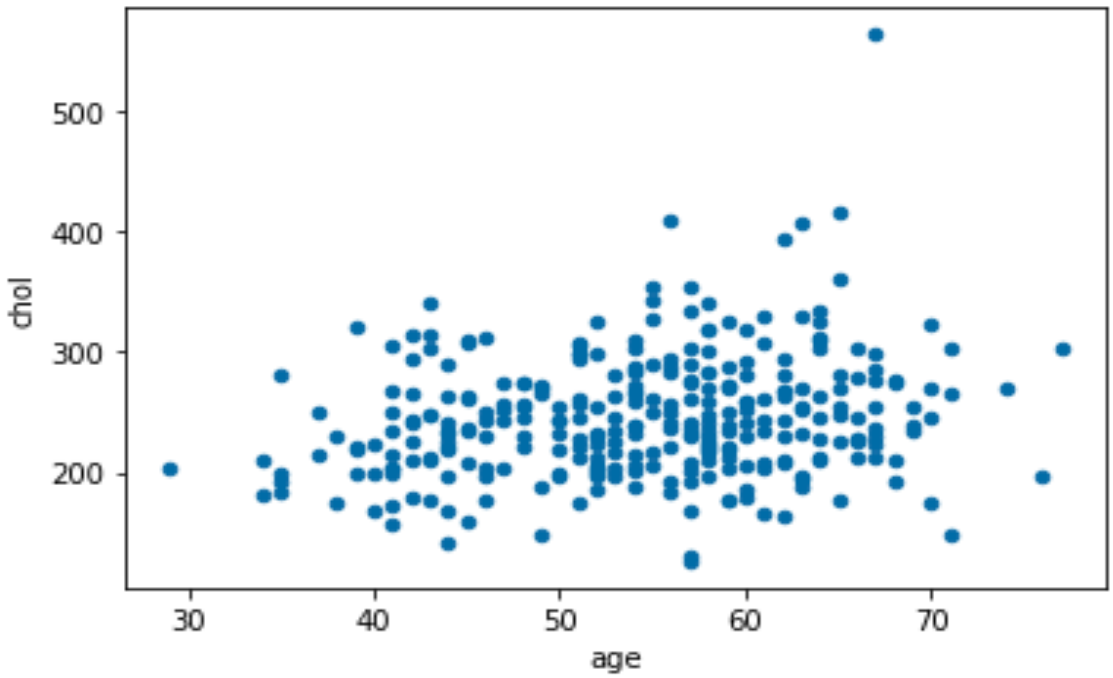

70
60
50
40
30
age
0
1
target


6
5
4
3
2
1
0
oldpeak
30
40
50
60
70
age

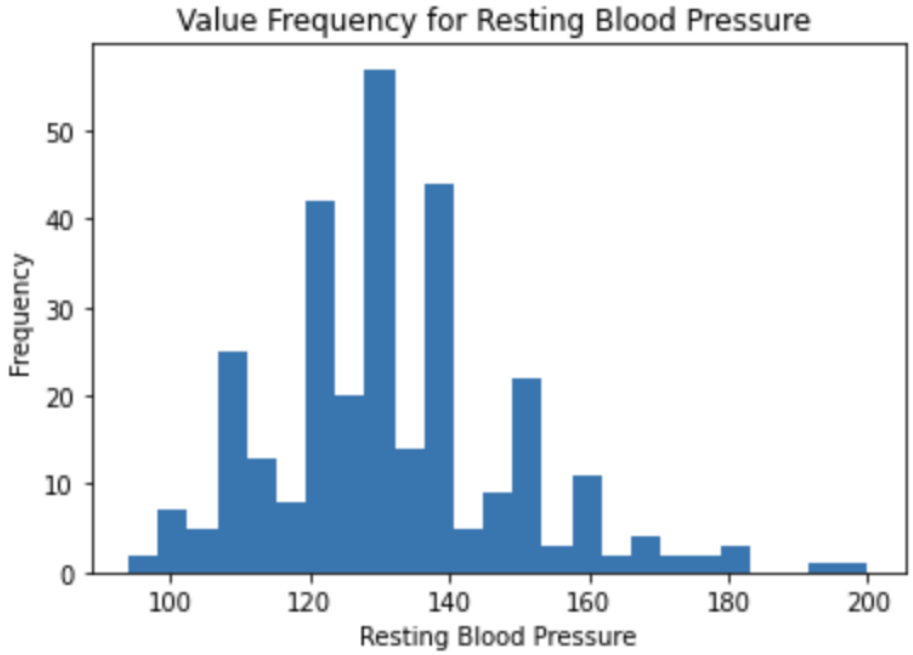
b) Value Frequency & Target for cholesterol: Excluding the outlier in our bar chart, we see that the cholesterol frequency illustration in left tailed. This indicates that we have an alternative hypothesis. As far as the mean, we got pretty close, when looking at the boxplot. On our last illustration, we can clearly state that as the age increases, the cholesterol raises as well, a positive correlation. If these patients have has a high level of the HDL cholesterol, they would have been less at risk.

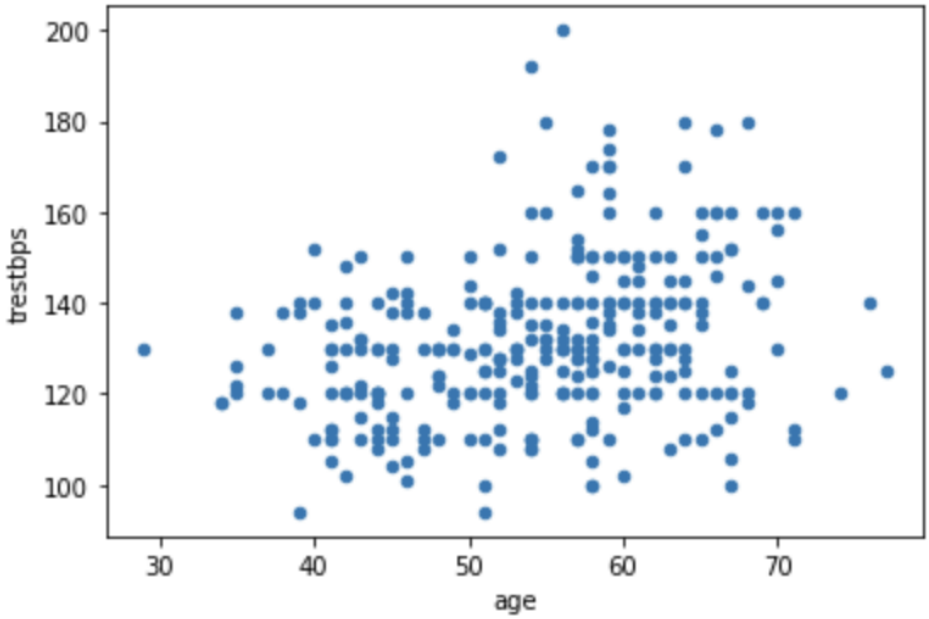






C) Correlation for resting blood pressure: The bar chart illustration below shows that we have a right railed skew in the data, resulting that the average of the resting Blood Pressure site at 130. In figure two (bottom figure), we see that again there is a positive correlation in higher resting blood pressure as age goes up.





5) Preprocessing: Model Selections

a) Logistic Regression: According to the Logistic Regression, after splitting and training the data, we have taken 6 folds and obtained a score of .9868. This tells us that the odds of heart disease are present. After fitting – an accuracy score of .9868 was present. When we were looking for the parameters within the GridSearchCV, for .001 our parameter was at .9913, which confirms that heart disease weas present.

b) Decision Tree: For our decision tree, we took 4 different models: Entropy No Max, Gini No Max, Entropy Max, and Gini Max. Within our first two models our scores remained the same. With Entropy meaning the measure of how uncertain we about which category the data-points fall into at a given point in the tree, we should be able to minimize entropy, and maximize the information gain. We obtained the Model 1 & 2 Entropy model with an accuracy of the following:

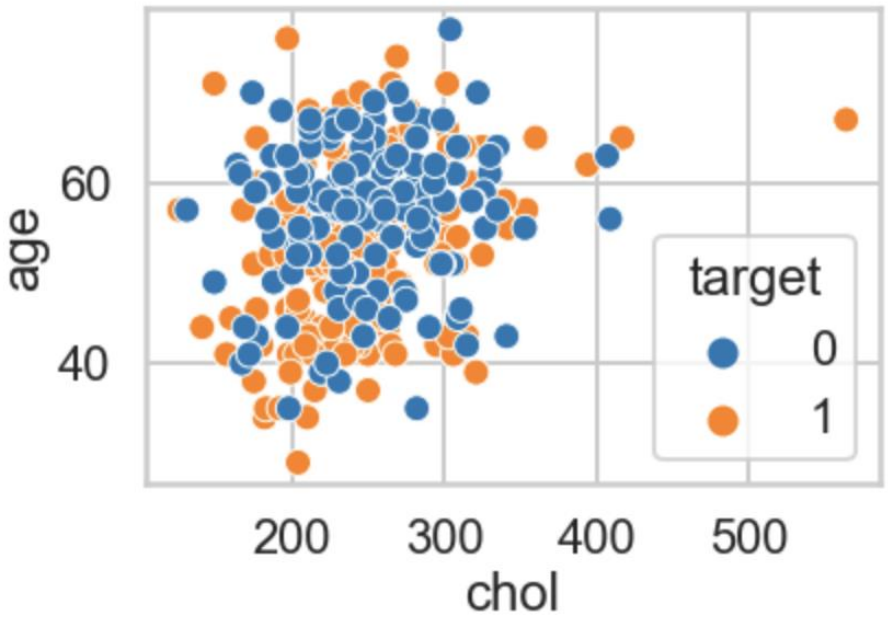
Model Entropy & Gini Models 1 and 2
Accuracy: 0.75
Balanced accuracy: 0.7614517265680056
Precision score: 0.6666666666666666
Recall score: 0.8484848484848485

When testing models 3 and 4, we saw that the Accuracy, Balanced Accuracy, Precision Score, and Recall score all had decreased. On our Model 4: Gini max scores resulted the following:

Accuracy: 0.7236842105263158
Balanced accuracy: 0.7417195207892882
Precision score 0.6304347826086957
Recall score 0.8787878787878788

As the data get tested multiple times, we see with a much higher recall score, our Accuracy score decreases. When looking at the Precision score, we can identify that the weight is fairly high.

Shown below in the boxplot below, we see that the median in age and target is very similar resulting an output of 1 (heart diseases being positive). Our scatterplot illustrates the level of cholesterol level by age (0, no heart disease present, 1, present. This chart made to understand the correlation of high or low cholesterol levels and age.



c) Gradient Boosting: In the output for gradient boosting, we outputted the following:

Learning rate: 0.05
Accuracy score (training): 0.855
Accuracy score (validation): 0.882

At a learning rate of 0.05, our accuracy rate still stayed above a .85%.

d) Random Forest:

In the dataset the fit model shows an overall accuracy of 78.7% which indicated our models was effectively able to identify the status of patients who have heart disease. Our F-1 Score has shown us the balance between precisions and recall of 78.5% in our dataset.

6) Future Research

- Deep dive into what other features impacted a higher risk of heart disease
- Understand how we can get better accuracy scores in our current models

Stakeholders

- Over the course of the project, it can be further implemented by doctors, patients, nurses, and medial companies who are working with the patient to try preventing heart disease or a company making a product which will prevent risk for an individual. This can be used by understanding the features and knowing why it happens at the age which It particularly occurs. All the stakeholders above can generate new feature out of the existing features if there is another big effect to one's health. In addition, patients would be able to keep building upon the models here for better accuracy rates. Doctors can use this information to keep as an additional resource to either walk patient through comparing and contrasting for education or use it to understand the scores and build upon those.