Springboard Data Science Course
Capstone Project 2

Classification of Heart Disease

By: Joe Anson R. Aquino
April 28, 2023

Introduction

Western countries face a major problem with heart disease. According to the US government, a heart attack occurs every 36 seconds. Our health is affected by many factors, including cholesterol, blood sugar levels, etc. My goal is to predict heart disease based on the following 13 attributes.

Approach

Gather data from the UCI machine learning repository and will organize make sure it is well-defined before cleaning and exploring it further. Will use the following machine-learning algorithm as follows for the prediction using the Cleveland Dataset for heart disease prediction.

- Logistic Regression Classifier
- Random Forest Classifier
- Light GBM Classifier
- xgBoost Classifier
- Decision Tree

Here is the link for Cleveland Dataset from the UCI machine learning repository

UCI Link: https://archive.ics.uci.edu/ml/datasets/heart+disease

Data Acquisition and Wrangling

After importing the Dataset we visualized the data in tabular format. Table 1

Table 1: **Summary of our Cleveland.csv dataset**

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2	7	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1	7	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1	3	1
302	38	1	3	138	175	0	0	173	0	0.0	1	?	3	0

303 rows × 14 columns

After changing the values of the row to the given attributes as follows.

Attribute Information:

```
0. (age)       age in years
1. (sex)       sex (1 = male; 0 = female)
2. (cp)        chest pain type
                      -- Value 1: typical angina
                      -- Value 2: atypical angina
                      -- Value 3: non-anginal pain
                      -- Value 4: asymptomatic
3. (trestbps)  resting blood pressure (in mm Hg on admission to the hospital)
4. (chol)      serum cholesterol in mg/dl
5. (fbs)       (fasting blood sugar > 120 mg/dl)  (1 = true; 0 = false)
6. (restecg)   resting electrocardiographic results
                      -- Value 0: normal
                      -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST
                                  elevation or depression of > 0.05 mV)
                      -- Value 2: showing probable or definite left ventricular hypertrophy
                                  by Estes' criteria
7. (thalach)   maximum heart rate achieved
8. (exang)     exercise induced angina (1 = yes; 0 = no)
9.(oldpeak)   ST depression induced by exercise relative to rest
10.(slope)     e 1: upsloping
                      -- Value 2: flat
                      -- Value 3: downsloping
11. (ca)        number of major vessels (0-3) colored by flourosopy
12. (thal)      3 = normal; 6 = fixed defect; 7 = reversable defect
13. (num)       (the predicted attribute)
                      diagnosis of heart disease (angiographic disease status)
                              -- Value 0: < 50% diameter narrowing
                              -- Value 1: > 50% diameter narrowing
                              (in any major vessel: attributes 59 through 68 are vessels)
```

Found missing entries marked with ‘?' (Table 2) and converted to ‘NaN’ Table 2.1

Table 2: Missing values found marked with “?”

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2	7	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1	7	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1	3	1
302	38	1	3	138	175	0	0	173	0	0.0	1	?	3	0

303 rows × 14 columns

'NaN' was replaced by the most frequent values and stored in the correct format.

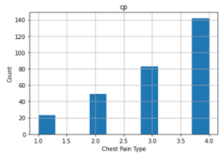
Table 2.1: Missing values converted to “NaN”

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	264	0	0	132	0	1.2	2	0	7	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2	7	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1	7	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1	3	1
302	38	1	3	138	175	0	0	173	0	0.0	1	NaN	3	0

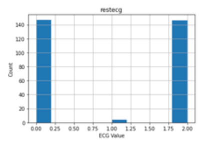
303 rows × 14 columns

```
age           Int64
sex           Int64
cp            Int64
trestbps      Int64
chol          Int64
fbs           Int64
restecg       Int64
thalach       Int64
exang         Int64
oldpeak     Float64
slope         Int64
ca            Int64
thal          Int64
num           Int64
dtype: object
```

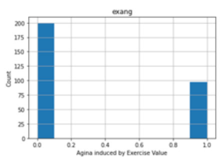
Storytelling and Inferential Statistics



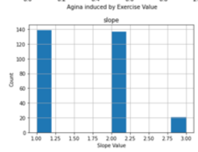
'cp' {Chest pain}: People with cp 2, 3, 4 are more likely to have heart disease than people with cp 1.



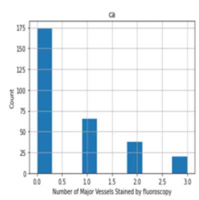
'restecg' {resting EKG results}: People with a value of 1 having ST-T wave abnormality and with value 2 showing probable or definite left ventricular hypertrophy by Estes' criteria , reporting an abnormal heart rhythm, which can range from mild symptoms to severe problems are more likely to have heart disease.



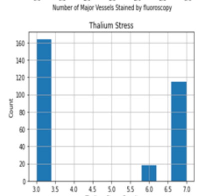
'exang' {exercise-induced angina}: people with a value of 0 (angina induced by exercise) have more heart disease than people with a value of 1 (angina induced by exercise)



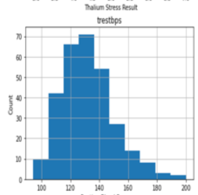
'slope' {the slope of the ST segment of peak exercise}: People with a slope value of 3 (Downslopins: signs of an unhealthy heart) are more likely to have heart disease than people with a slope value of 1 slope (Upsloping: best heart rate with exercise) or 2 (Flat sloping: minimal change (typical healthy heart)).



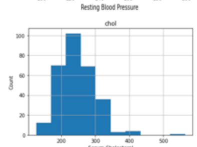
'ca' {number of major vessels (0-3) stained by fluoroscopy}: the more blood movement the better, so people with 'ca' equal to 1 are more likely to have heart disease.



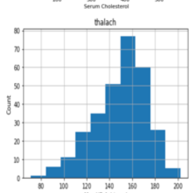
'thal' {thalium stress result}: People with a thal value of 3 with no blood flow in some part of the heart are more likely to have heart disease.



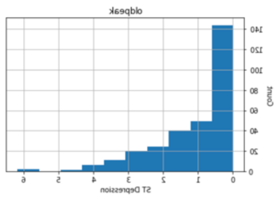
'trestbps': resting blood pressure anything above 130-140 is generally of concern



'chol': greater than 200 is of concern.



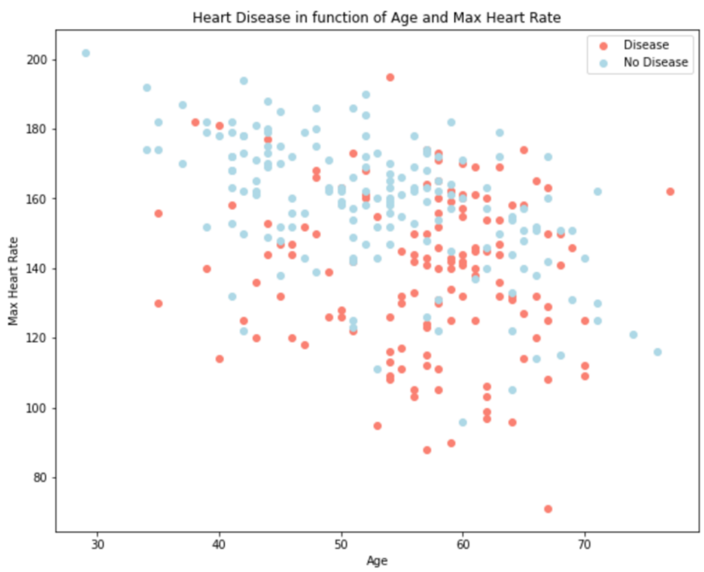
'thalach': People with a maximum of over 140 are more likely to have heart disease.



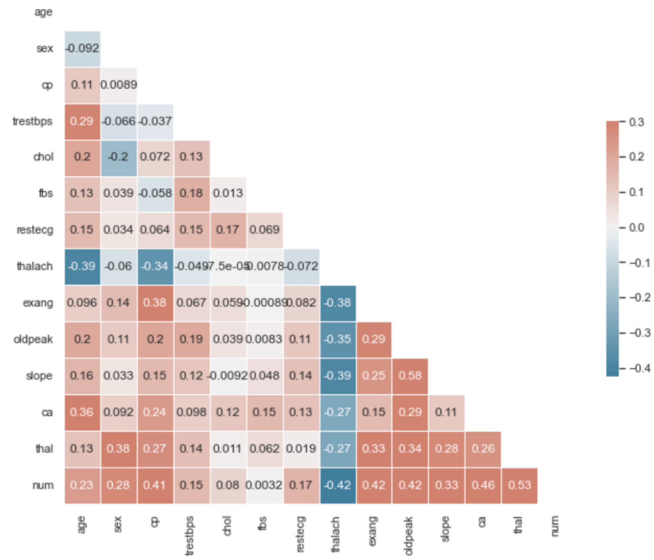
'oldpeak' of exercise-induced ST depression vs. rest looks at heart stress during exercise an unhealthy heart will stress more.

A graph of Scatter plotted points showing the relationship between “Maximum Heart Rate” and “Age” .

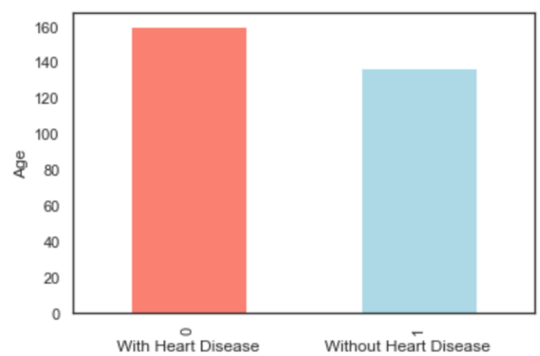
'Maximum Heart Rate' versus 'Age' showed 50 years of age and higher mostly have heart disease.



Heatmap shows between 'oldpeak' and 'slope has are highly positively correlated. Our target 'num' is mostly correlated to all our features except 'num' and 'thalach' with a negatively correlation.

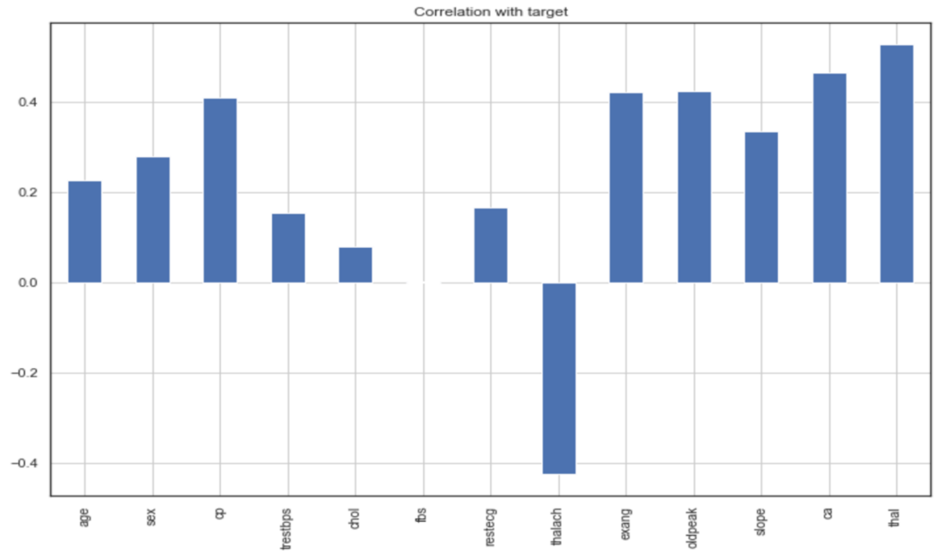


This Bar Graph show people with or without heart disease



We have 160 people with heart disease and 138 people without heart disease, so our problem is balanced.

This Bar Graph show the Correlation with the target



Exploration of the data indicated that patients 'oldpeak' and 'slope has are highly positively correlated. 'fbs' and 'chol' are the least correlated with the target variable. Our target 'num' is mostly correlated to all our features except 'num' and 'thalach' with a negatively correlation.

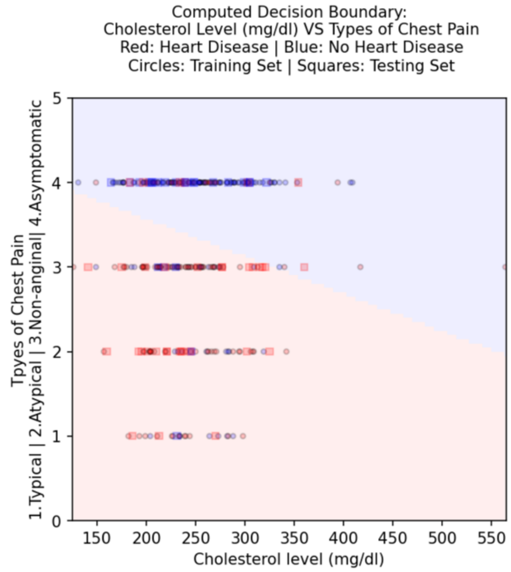
Baseline Modeling

Building classification will count the number of samples per class, proportionally to the total number of samples

In the figure, class 0 (no heart disease) is shaded red, and class 1 (heart disease) is shaded blue. The train labels are plotted as circles, using the same color scheme, while the test data are plotted as squares.

The classifier tends to suggest heart disease either with Chest Pain or cholesterol increase. This seems possibly correct.

The "decision boundary" is a line. As we add more features, we won't be able to represent the boundary this way. The boundary becomes what is called a hyperplane, which is the generalization of a line into 3 or more dimensions. But here, a patient measured with a combination of cholesterol and Chest Pain to the right of the line in the blue region would be classified as likely having heart disease. Asymptomatic patient even without feeling of Chest Pain but lab works, ekg/ecg are remarkable.



The classifier good and realistic! The accuracy on the training data is only 73%, and the accuracy on the testing data is 80%.

The model performs good when trying to recognize inputs that belong to class 1 (*the class of interest*) , we have good values of precision, recall and f1-score for class 1 on training and test set.

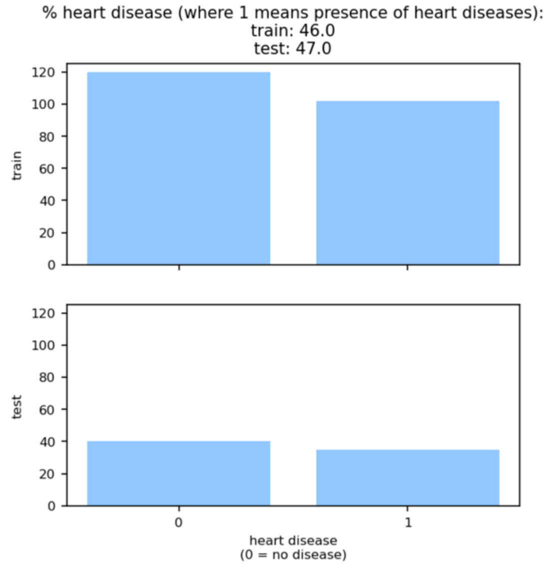
Classification Report for Training Data

	precision	recall	f1-score	support
0	0.76	0.71	0.73	124
1	0.70	0.75	0.73	113
accuracy			0.73	237
macro avg	0.73	0.73	0.73	237
weighted avg	0.73	0.73	0.73	237

Classification Report for Test Data

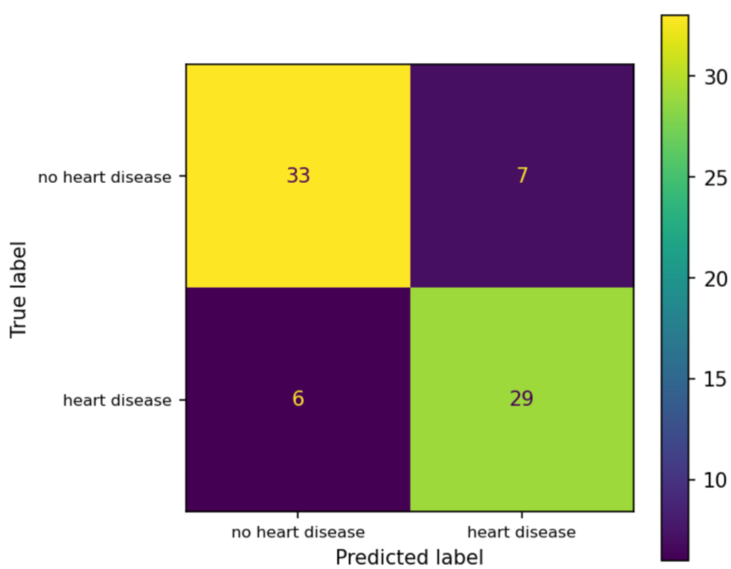
	precision	recall	f1-score	support
0	0.85	0.81	0.83	36
1	0.73	0.79	0.76	24
accuracy			0.80	60
macro avg	0.79	0.80	0.79	60
weighted avg	0.80	0.80	0.80	60

*Perform train/test split on (X,y). Inspect the **Train** response data (ylr) compared to the **Test** response data (ytestlr).*



It turns out that 'train_test_split' provides a way to compute splits that try to preserve the proportions among the classes in the original dataset. ytestlr' has one point higher percentage of heart disease (47%), compared to the percentage in the original dataset (46%). train/test split made the imbalance, we would like to perform a split preserving the original proportions among the classes, so we do not have to worry about the possibility of getting poor results due to this fact.

Confusion Matrix



Above Confusion Matrix we have 75 total samples.

Class 1 (heart disease) has 29 samples we expected to be positive came back positive, these are my True Positive. 6 samples that we expected to be positive came back negative, these are my False Negative and for class 0 (No heart disease) there are 7 samples that we expected to be negative came back positive, these are my False Positive. 33 samples that we expected to negative came back negative, these are my True Negative.

Classificatgion Report for Training.

	precision	recall	f1-score	support
0	0.85	0.82	0.84	40
1	0.81	0.83	0.82	35
accuracy			0.83	75
macro avg	0.83	0.83	0.83	75
weighted avg	0.83	0.83	0.83	75

Classification Report for Test

	precision	recall	f1-score	support
0	0.87	0.91	0.89	120
1	0.89	0.83	0.86	102
accuracy			0.87	222
macro avg	0.88	0.87	0.87	222
weighted avg	0.87	0.87	0.87	222

The model's **training accuracy** (0.87) is pretty good (meaning, close to 1--or 100%), then one says there is only a small "bias" in the model.

The model's **test accuracy** (0.83) is decently close to the training accuracy, we would say that there is a small "variance" between the training accuracy and the test accuracy. This is an indication that the model will "generalize well", which means that the model will be well-behaved when new data is presented to it.

Since the gap between training and testing accuracy is about 4%, one might say that the model is slightly over-fitting the data. Thus, in general, one says that a model is over-fitting (or just overfitting), when there is an important gap between its training performance and its test performance. The model can be improved--repeat as needed with additional algorithms and/or by applying hyper-parameter tuning.

Extended Modeling

- Logistic Regression Classifier

- Random Forest Classifier
- Light GBM Classifier
- xgBoost Classifier
- Decision Tree

Findings

2 Approach

2.1 Data Acquisition and Wrangling

2.2 Storytelling and Inferential Statistics

2.3 Baseline Modeling

2.4 Extended Modeling

4 Conclusions and Future Work

5 Recommendations for the Clients

6 Consulted Resources