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. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. This project aims to predict future heart disease by analyzing data of patients that classify whether they have heart disease or not using a machine-learning algorithm. Machine Learning techniques can be an advantage in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing them to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Approach

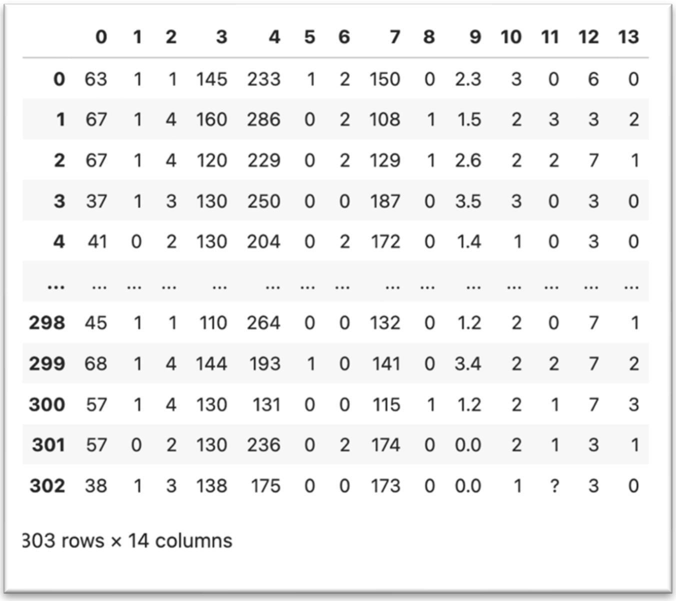
First, will gather data from the UCI machine learning repository and organize it to ensure it is well-defined. Then, I will clean and explore the data before utilizing a machine-learning algorithm to predict heart disease using the Cleveland Dataset. To justify this work, I conducted a comparative study and analysis using various classification algorithms, including Logistic Regression Classifier, Random Forest Classifier, Light GBM Classifier, xgBoost Classifier, and 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

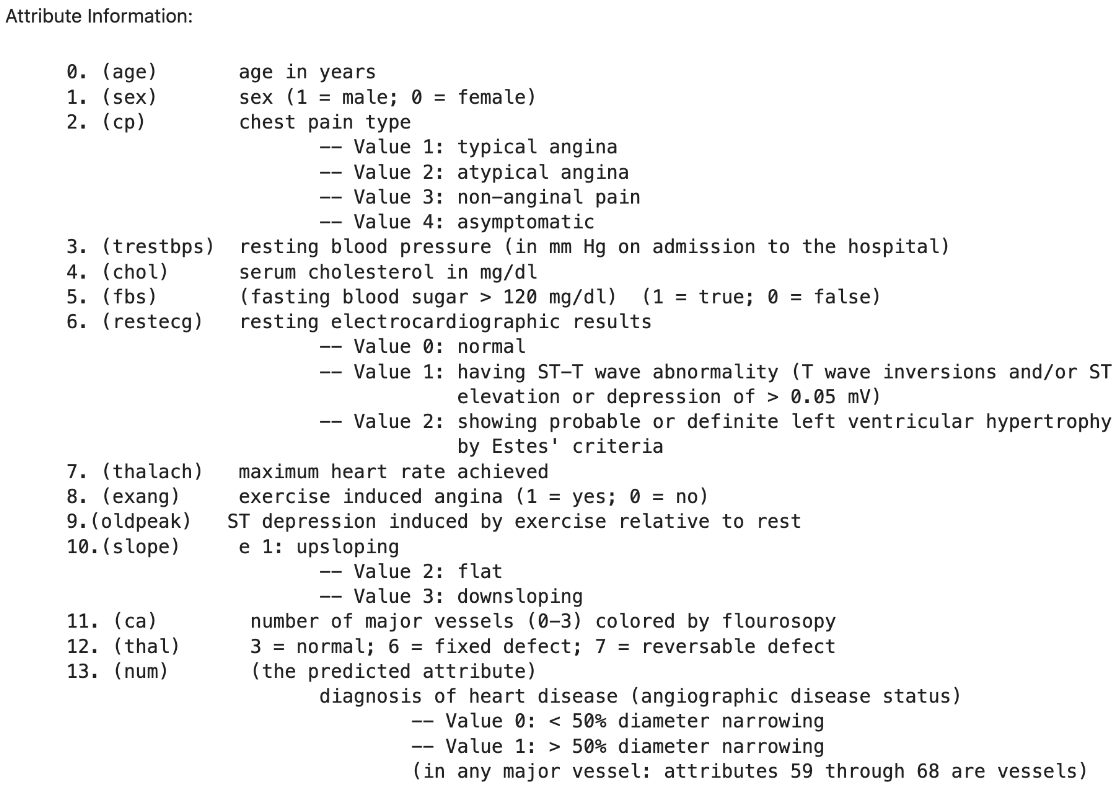
The dataset used for this project is the Heart Disease UCI dataset, which contains 76 attributes. For the system, 14 of these attributes were utilized. Table 1



	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

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

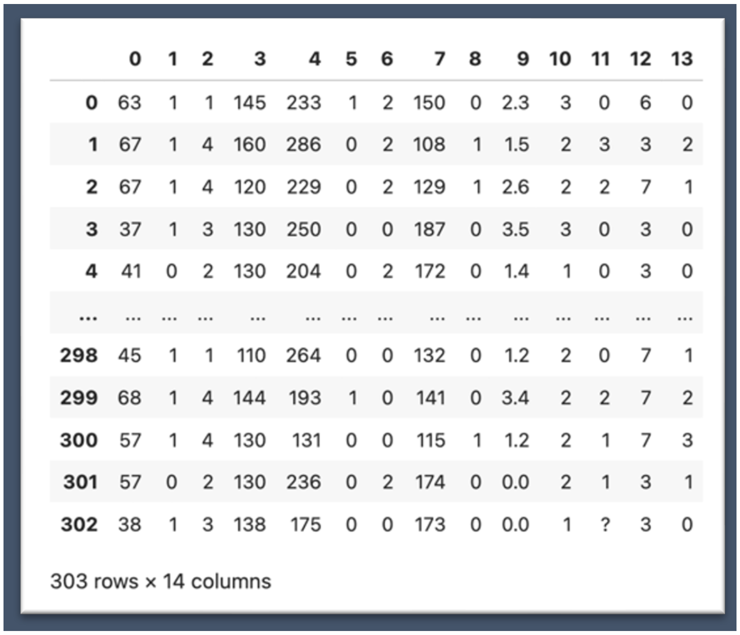


Attribute Information:

```
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 “?”

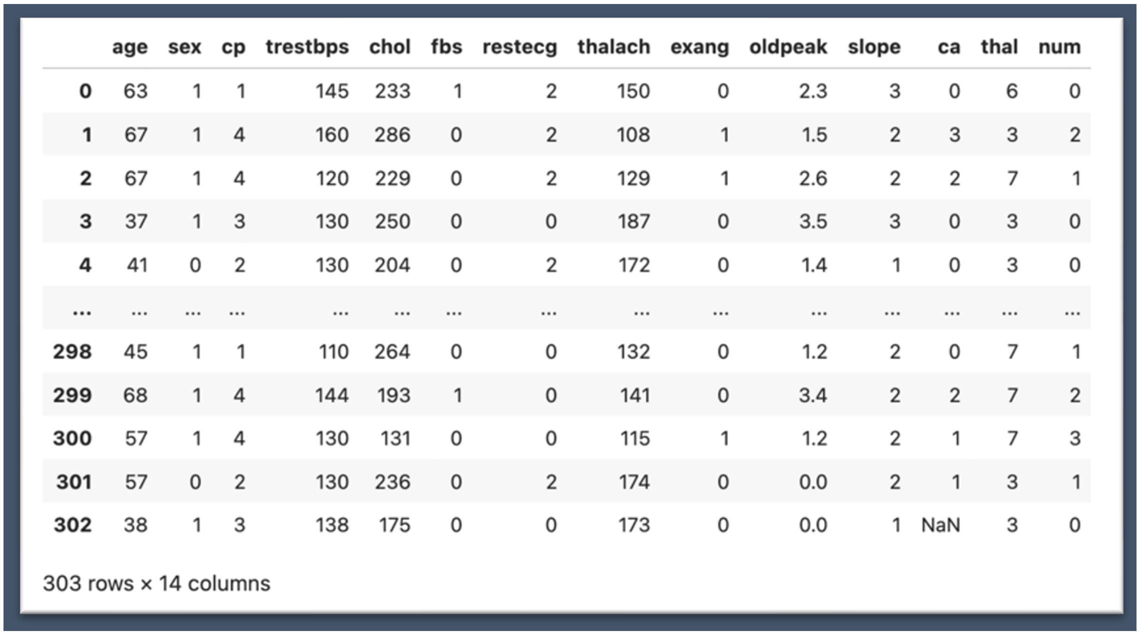


	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	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

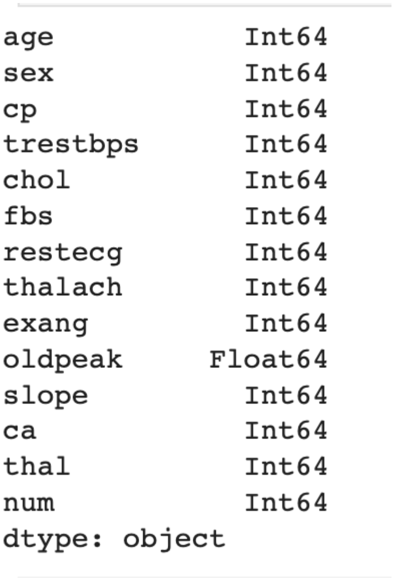
'NaN' was replaced by the most frequent values and stored in the correct format.

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

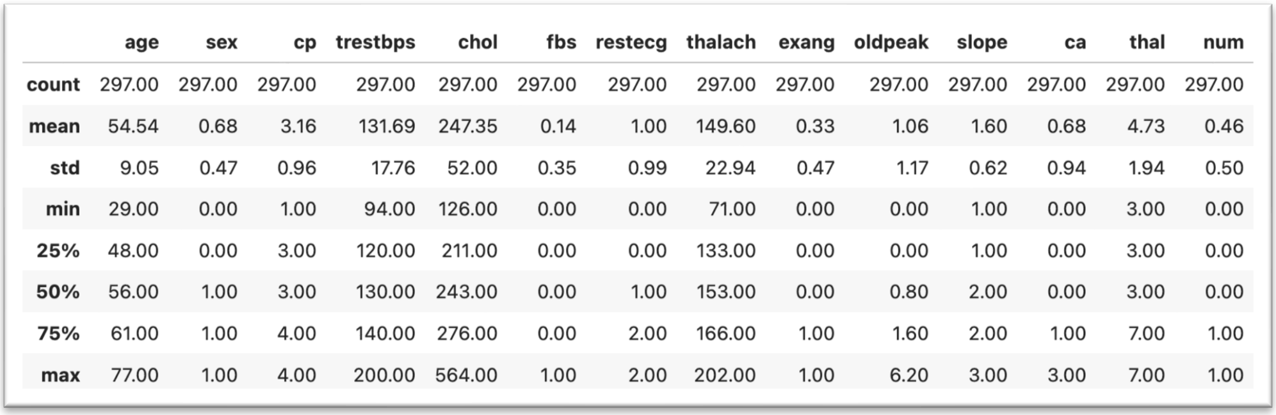
303 rows × 14 columns



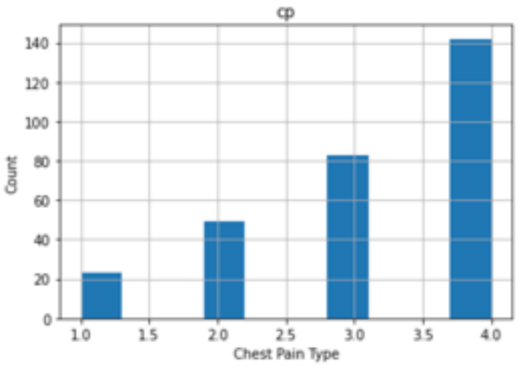
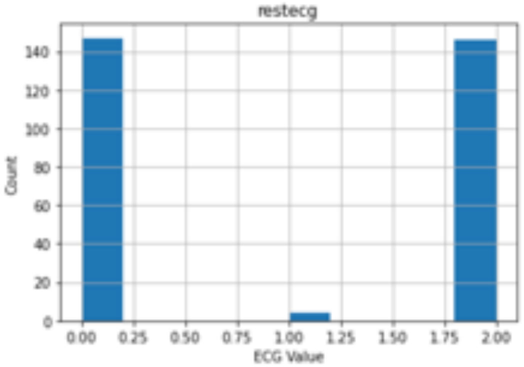
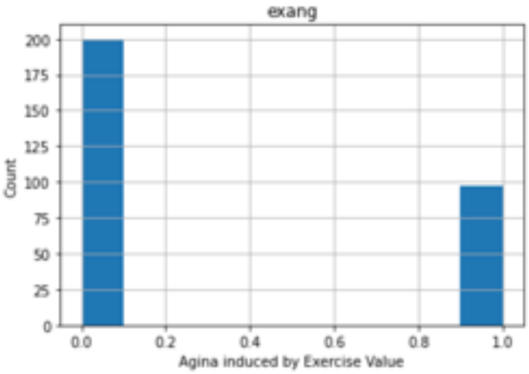
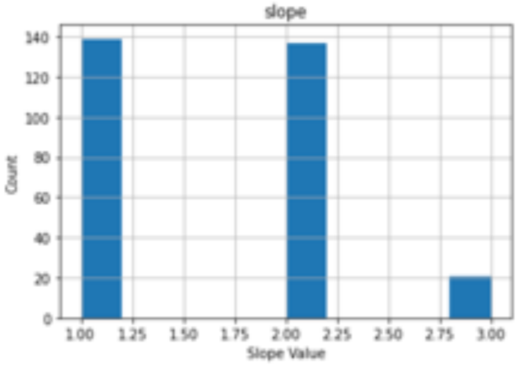
```
age           Int64
sex           Int64
cp            Int64
trestbps      Int64
chol          Int64
fbs           Int64
restecg       Int64
thalach       Int64
exang         Int64
oldpeak     Float64
slope         Int64
ca            Int64
thal          Int64
num           Int64
dtype: object
```

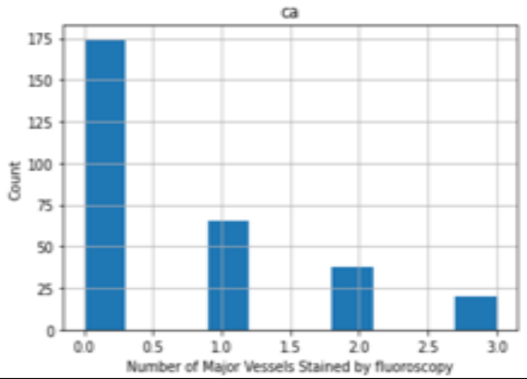
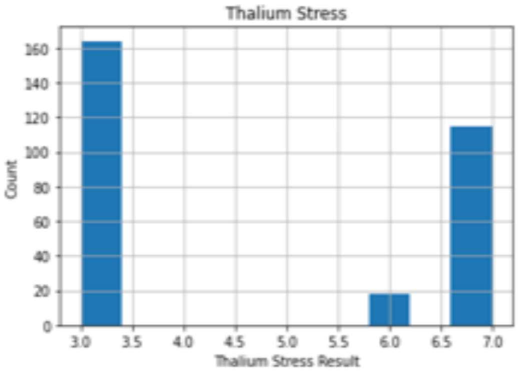
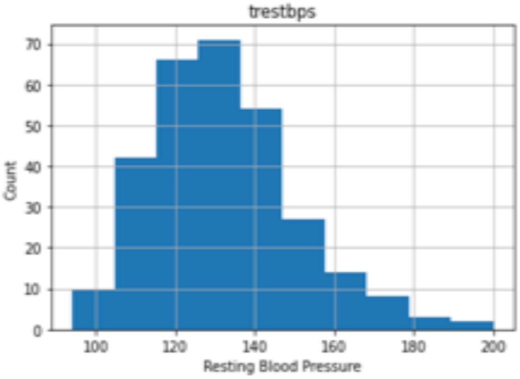
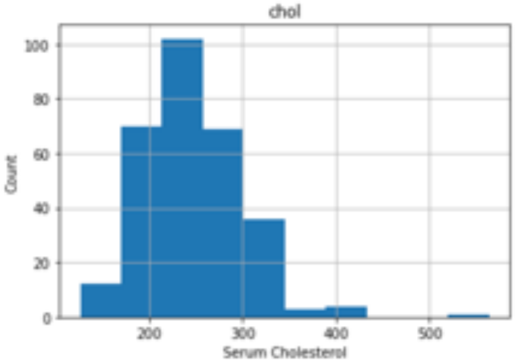
Storytelling and Inferential Statistics

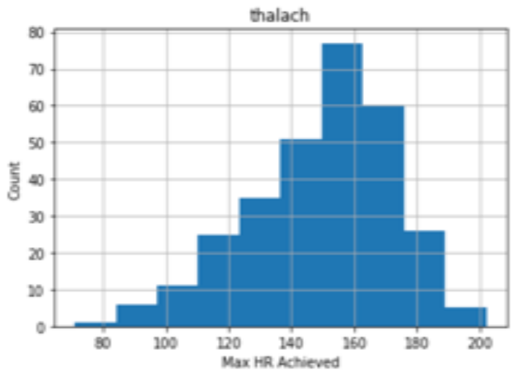
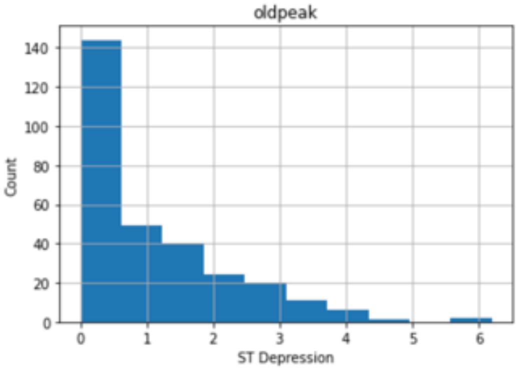
Checking on the min and max value for the categorical variables (min-max). and also observing the mean, std, 25% and 75% on the continuous variables.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00
mean	54.54	0.68	3.16	131.69	247.35	0.14	1.00	149.60	0.33	1.06	1.60	0.68	4.73	0.46
std	9.05	0.47	0.96	17.76	52.00	0.35	0.99	22.94	0.47	1.17	0.62	0.94	1.94	0.50
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00	3.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.00	0.00	0.00	1.00	0.00	3.00	0.00
50%	56.00	1.00	3.00	130.00	243.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00	3.00	0.00
75%	61.00	1.00	4.00	140.00	276.00	0.00	2.00	166.00	1.00	1.60	2.00	1.00	7.00	1.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	3.00	7.00	1.00

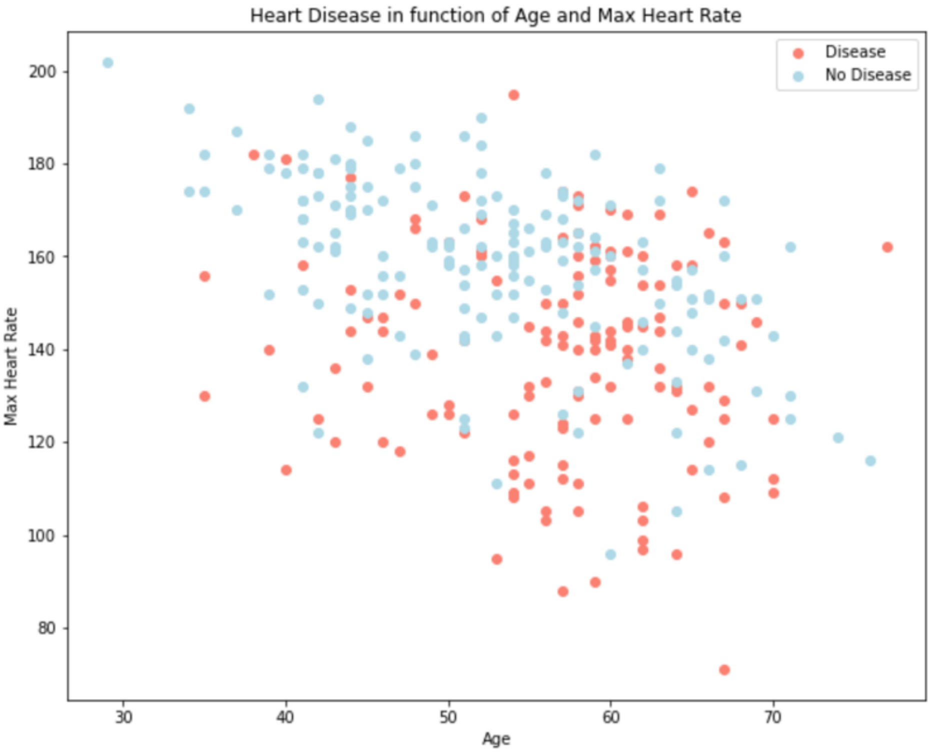
	'cp' {Chest pain}: People with cp 2, 3, 4 are more likely to have heart disease than people with cp 1.
	'restecg' {resting EKG results}: People with a value of 1 having ST-T wave abnormality and with value 2 showing probable or definite left ventricular hypertrophy by Estes' criteria , reporting an abnormal 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.

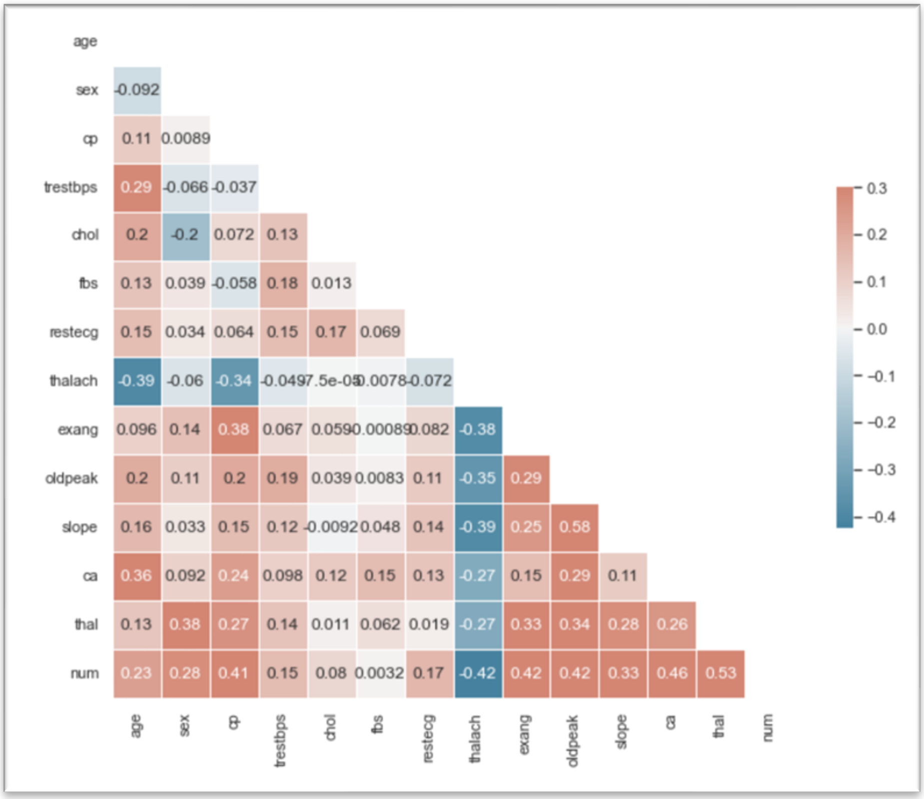
A graph of Scatter plotted points showing the relationship between “Maximum Heart Rate” and “Age” .

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

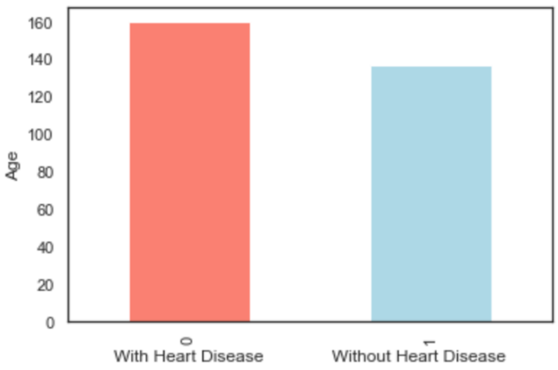


The Correlation matrix is used for attribute selection for this model. Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction.

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.

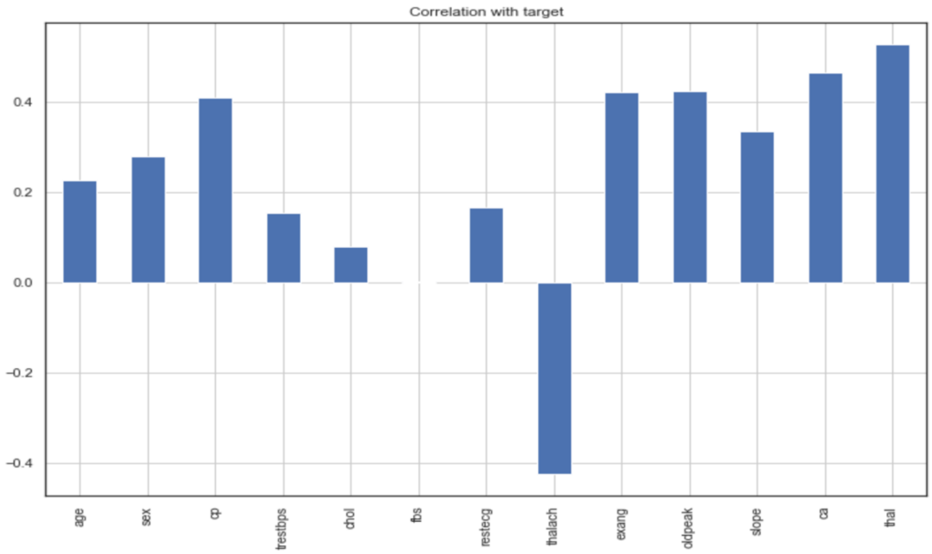


This Bar Graph show people with or without heart disease



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

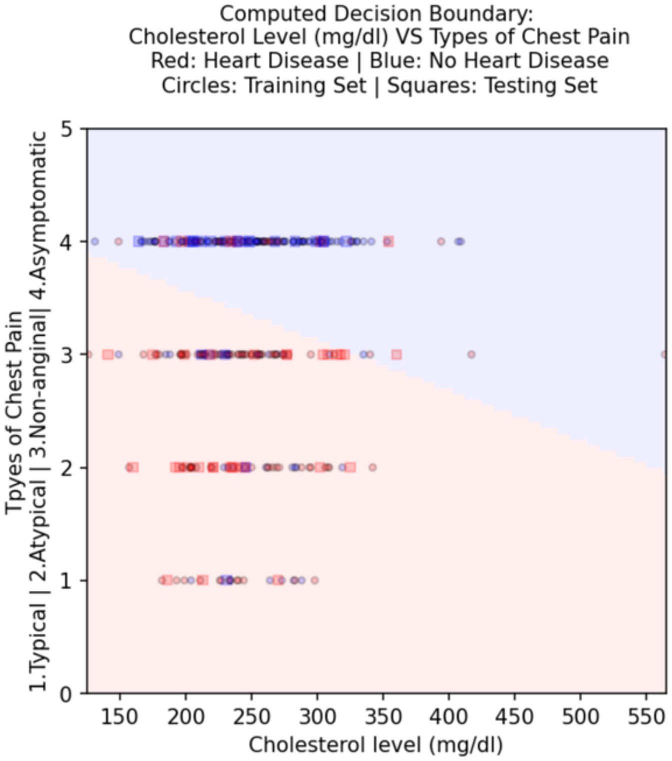
This Bar Graph show the Correlation with the target



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

In order to improve the accuracy of our model, we need to pre-process the data. This involves transforming the dataset into the required format and dealing with issues such as noise, duplicates, and missing values. Activities such as importing datasets, splitting datasets, and attribute scaling are included in data pre-processing.

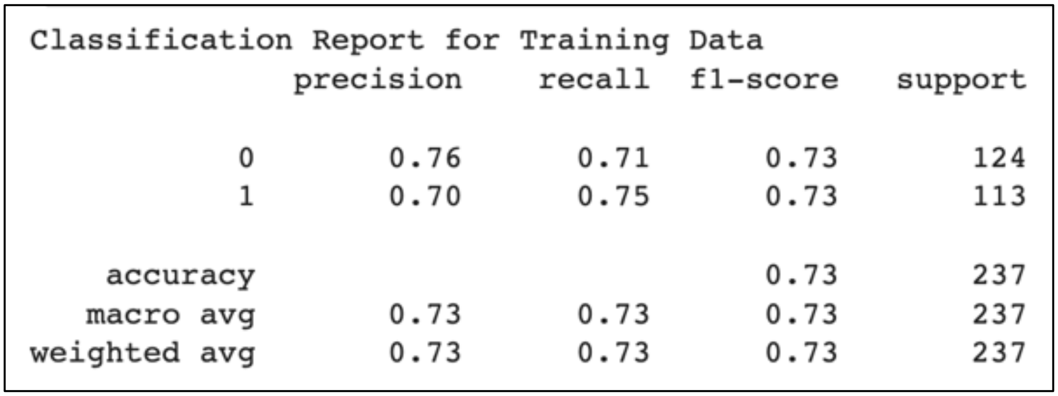


In the figure, class 0 (no heart disease) is shaded red, and class 1 (heart disease) is shaded blue. The train labels are plotted as circles, using the same color scheme, while the test data are plotted as squares.

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

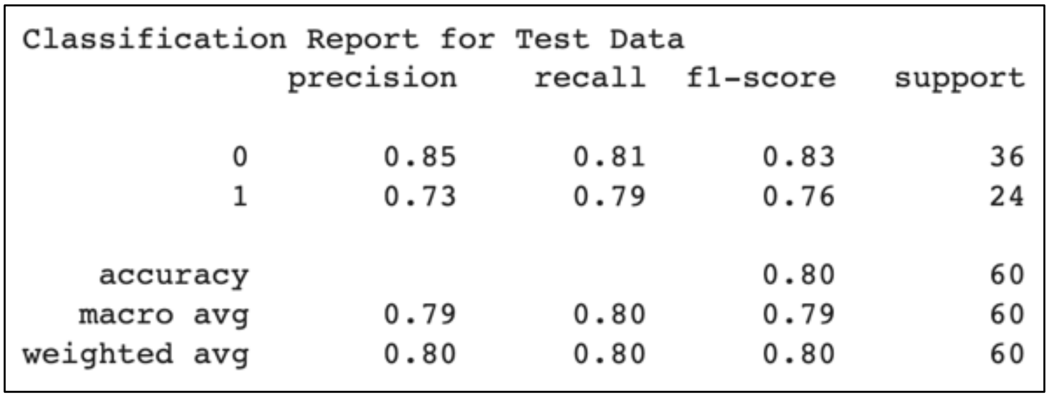
The "decision boundary" is a line. As we add more features, we won't be able to represent the boundary this way. The boundary becomes what is called a hyperplane, which is the generalization of a line into 3 or more dimensions. But here, a patient measured with a combination of cholesterol and Chest 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.

Training Classification Report



```
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
```

Training Classification Report

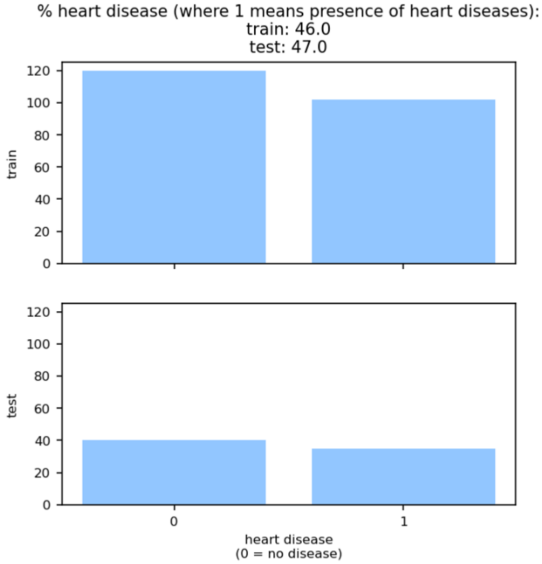


```
Classification Report for Test Data
              precision    recall  f1-score   support

           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
```

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

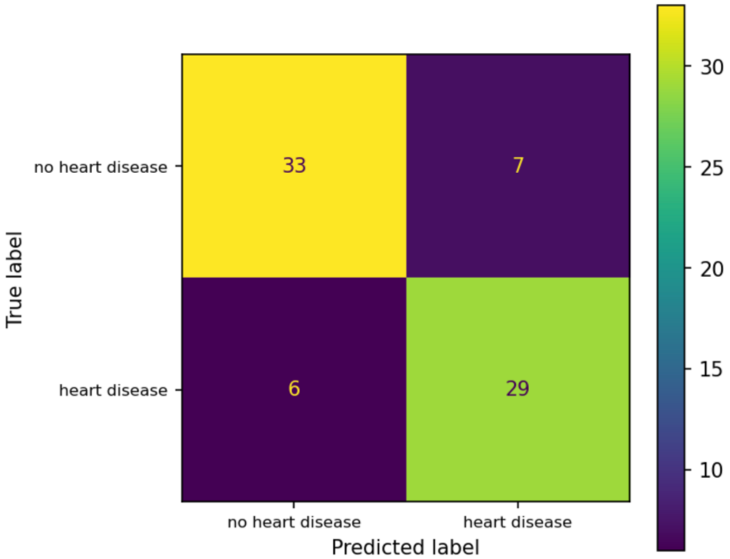
The model performs good when trying to recognize inputs that belong to class 1 (*the class of interest*), we have good values of precision, recall and f1-score for class 1 on training and test set.

*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.

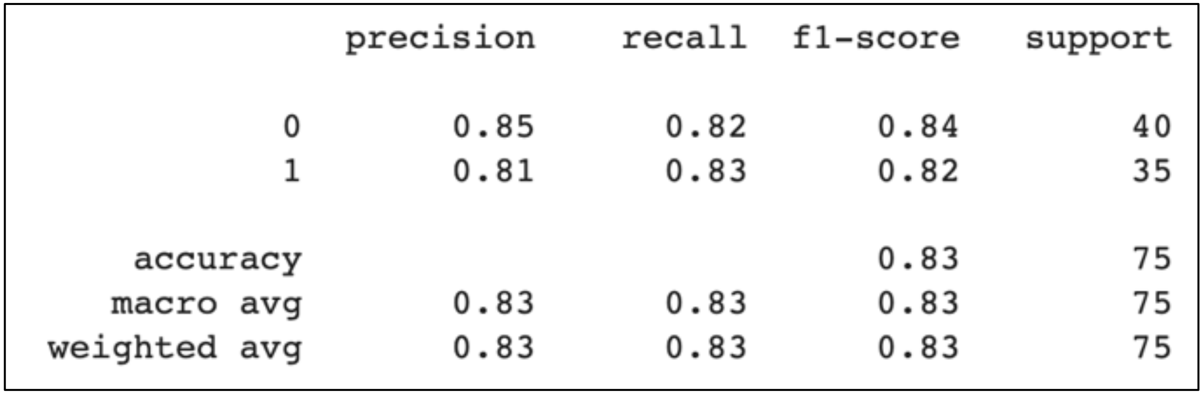
Confusion Matrix



Above Confusion Matrix we have 75 total samples.

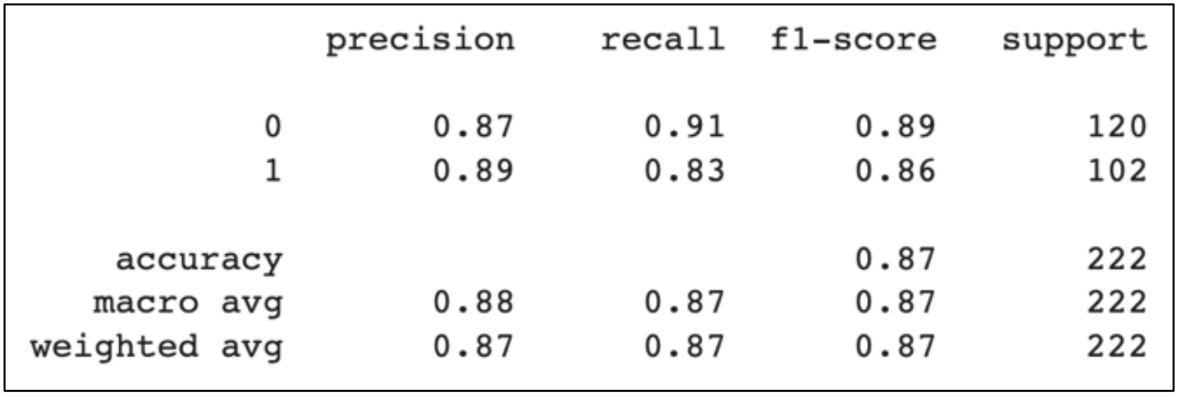
Class 1 (heart disease) has 29 samples we expected to be positive came back positive, these are my True Positive. 6 samples that we expected to be positive came back negative, these are my False Negative and for class 0 (No heart disease) there are 7 samples that we expected to be negative came back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

Training Classificatgion Report



```
              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
```

Test Classification Report



```
              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

The project was completed with five ML models: Logistic Regressions, Random Forest, LGBM, XGBoost and Decision Tree .

LOGISTIC REGRESSION CLASSIFIER

```
Logistic Regression Accuracy: 0.96
               precision    recall  f1-score   support
           0       0.93      1.00      0.97        43
           1       1.00      0.91      0.95        32
    accuracy                           0.96        75
   macro avg       0.97      0.95      0.96        75
weighted avg       0.96      0.96      0.96        75
```

```
Logistic Regression with GridSearchCV Accuracy: 0.96
               precision    recall  f1-score   support
           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29
  accuracy                            1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75
```

Logistic regression: One of the very popular algorithms is considered as logistic regression which is a supervised learning model. It performs categorical predictions which can be ‘true’ or ‘false’. This model provides probabilistic values instead of exact ones. This algorithm works on both continuous and discrete values. A simple S-Shaped curve can elaborate the logistic regression very precisely.

RANDOM FOREST CLASSIFIER

```
RandomForestClassifier Accuracy: 0.96
              precision    recall  f1-score   support
           0       0.98      1.00      0.99        45
           1       1.00      0.97      0.98        30
    accuracy                           0.99        75
   macro avg       0.99      0.98      0.99        75
weighted avg       0.99      0.99      0.99        75
```

```
Random Forest with GridSearchCV Accuracy: 0.96
              precision    recall  f1-score   support
           0       1.00      0.98      0.99        47
           1       0.97      1.00      0.98        28
    accuracy                           0.99        75
   macro avg       0.98      0.99      0.99        75
weighted avg       0.99      0.99      0.99        75
```

Random forest is a flexible, easy-to-use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most-used algorithms, due to its simplicity and diversity (it can be used for both classification and regression tasks).

LIGHT GBM CLASSIFIER

```
LGBM Accuracy: 0.8667
              precision    recall  f1-score   support

           0       0.87      0.91      0.89        44
           1       0.86      0.81      0.83        31
    accuracy                           0.87        75
   macro avg       0.87      0.86      0.86        75
weighted avg       0.87      0.87      0.87        75
```

```
LGBM with GridSearchCV Accuracy: 1.0
        precision    recall  f1-score   support

           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29
    accuracy                           1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75
```

LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine Learning tasks.

XGBOOST CLASSIFIER

```
XGBoost Classifier Accuracy: 1.0
                  precision    recall  f1-score   support
               0       1.00      1.00      1.00        46
               1       1.00      1.00      1.00        29
        accuracy                           1.00        75
       macro avg       1.00      1.00      1.00        75
    weighted avg        1.00      1.00      1.00        75
```

```
XGBoost Classifier with GridSerchCV Accuracy: 0.92
precision    recall  f1-score   support
           0       0.87      1.00      0.93        40
           1       1.00      0.83      0.91        35
    accuracy                           0.92        75
   macro avg       0.93      0.91      0.92        75
weighted avg       0.93      0.92      0.92        75
```

.

XGBoost: It is a decision tree classifier which has been implemented on gradient boosting framework. This model works on the principle that weak learners should be combined to produce best predictions. Ensembling is performed in sequential manner.

DECISION TREE

```
Decision Tree Accuracy: 1.0
               precision    recall  f1-score   support
           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29
    accuracy                           1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75
```

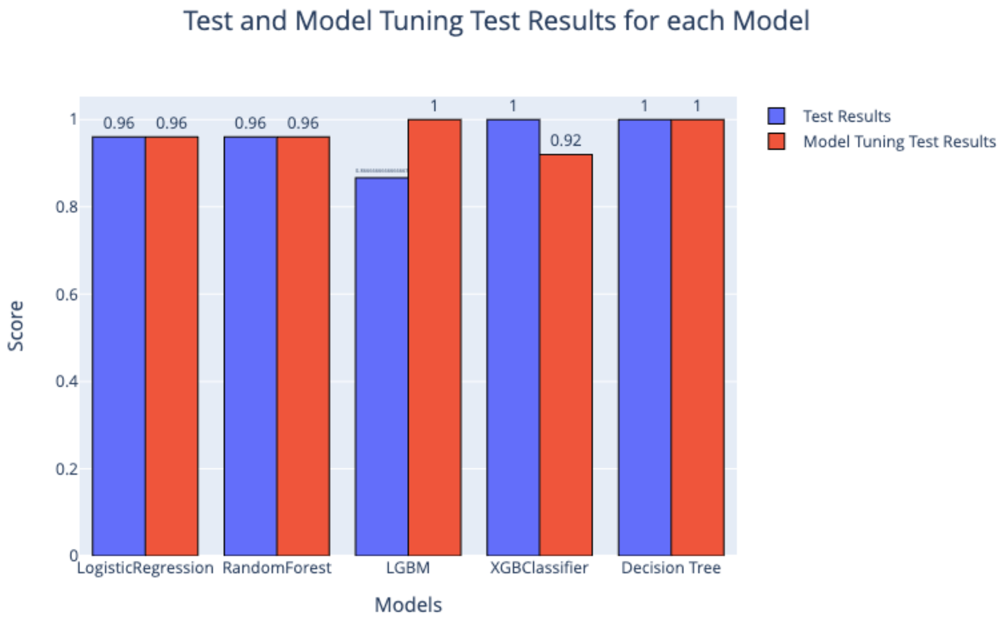
```
Decision Tree with GridSerchCV Accuracy: 1.0
               precision    recall  f1-score   support
           0       1.00      1.00      1.00        46
           1       1.00      1.00      1.00        29
    accuracy                           1.00        75
   macro avg       1.00      1.00      1.00        75
weighted avg       1.00      1.00      1.00        75
```

Decision Tree creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

MODEL COMPARISON

LGBM and Decision Tree provided the best prediction among the five. The LGBM is 1.00 and Decision Tree is 1.00. The prediction with ML models in identifying heart attack symptoms is highly efficient, especially with boosting algorithms. The prediction was done to evaluate accuracy, precision, recall. ML models are being trained to perform optimized predictions.

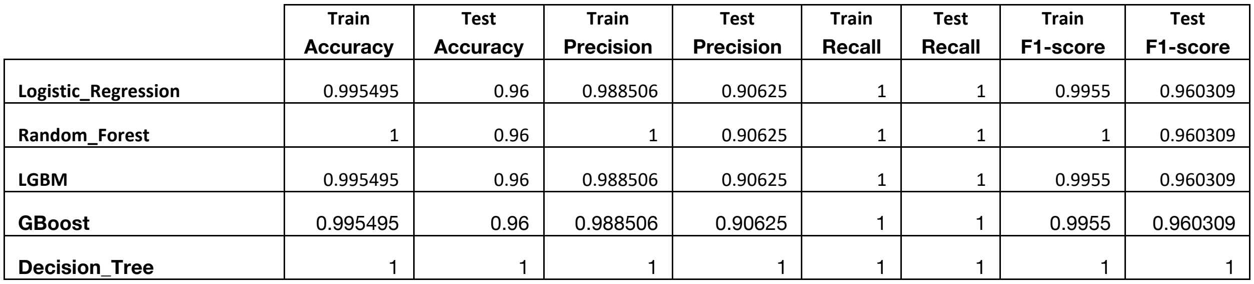
MODEL	Default	with GridSearchCV
LogisticRegression	0.960000	0.96
RandomForest	0.960000	0.96
LGBM	0.866667	1.00
XGBClassifier	1.000000	0.92
Decision Tree	1.000000	1.00



Findings

In this work, the evaluation of the performance metrices are being done with five machine learning classifiers Logistic Regressions, Random Forest, LGBM, XGBoost and Decision Tree .

Decision Tree classifier provided best training and test scores of 1 and 1 along with the 1 accuracy. The results achieved are discussed below presents the interface for taking input from users and predicting using machine learning.



	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1-score	Test F1-score
Logistic_Regression	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
Random_Forest	1	0.96	1	0.90625	1	1	1	0.960309
LGBM	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
GBoost	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
Decision_Tree	1	1	1	1	1	1	1	1

Conclusions and Future Work

The comparative evaluation of five machine learning algorithms for the heart disease prediction was carried out in this study, with promising outcomes. In this investigation, the performance of ML approaches has been better. When data pre-processing was used, LGBM and Decision Tree performed better in the ML technique for the 13 features in the dataset.

Deep learning algorithms are essential in application for the healthcare industry. Therefore, using deep learning techniques to forecast heat disease may produce superior results. In order to determine the severity of the sickness, we are also interested in category it as a multiclass problem.

Recommendations for the Clients

- It is recommended to have Additional data from many sources could be taken so that the models would be able to predict for different conditions for the patients.

- More features that help determine whether a person would suffer from heart disease could be considered.

- Use Decision Tree model, which had the best performance, could be deployed in real-time to provide doctors with faster inference results. This could aid in the diagnosis of whether a person is suffering from heart disease or not.

`

Consulted Resources

Heart Disease Data Set. Creators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Fatima M, Pasha M: Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* 2017; 09: 1–16. Publisher Full Text