Heart Diseases Classifier

Abstract:

Heart disease is the leading cause of death in the United States. With 1 out of every 4 deaths relating to heart disease, the prediction of heart diseases in people is a prominent clinical data analysis experiment. Since the early prognosis of heart diseases can aid in making better decisions on lifestyle changes in high risk patients and in turn reduce the complications, this research intends to predict the chances of a person having heart disease, given certain factors, while also trying to predict the most important risk factors of heart disease. The data, Cleveland Heart Disease dataset, was split into training and test sets using scikit learn. Since, the target variable contained only two classes, binary classification algorithms were implemented on the training data. The models trained were Naive Bayes, Logistic Regression, KNN, Decision Trees, SVM, Random Forest, NN and XGBoost. Using the AUC scores as the performance measure, the best model obtained was XGBoost model with AUC score of 0.95.

Introduction:

Heart Disease or Cardiovascular Disease, in general, means the functioning of the heart is impaired and is one of the leading causes of death in the US in both men and women. Some people are born with heart disease, congenital heart disease, while some people acquire it, acquired heart disease, during their lifetime. The prediction of heart diseases is regarded as very important in clinical data analysis experiments. With huge amounts of raw data available in the health industry and with data mining and good algorithms, we can try to make some informed decisions and predictions. Using a heart disease dataset, and with the help of statistical and ML techniques, we try to manually predict the odds of a person getting heart disease based on risk factors.

Methodology:

The dataset used for this research is Cleveland Heart Disease dataset. The data was already cleaned without any null values and comprises 14 attributes about 303 individuals. The target variable indicates whether the person has heart disease with 1 being yes and 0 otherwise. The data is balanced with almost the same number of records for both types.

The response variable ‘target’ is slightly positively correlated with ‘cp’, ‘thalach’, ‘slope’ and slightly negatively correlated with ‘exang’, ‘oldpeak’, ‘ca’, ‘thal’. Our data indicates that male people have more chances of getting heart disease than females. Also, the data suggests that the age group of above 50 years has more chance of having a heart disease with peak of the distribution around 58 years.

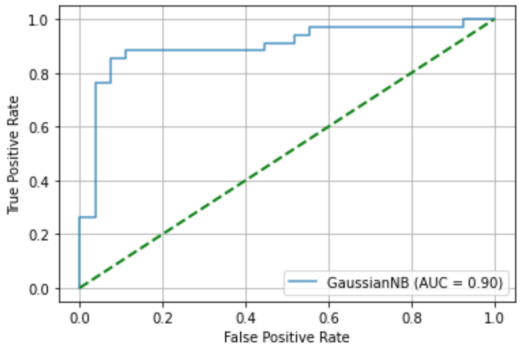
The categorical variables, ‘cp’, ‘thal’ and ‘slope’, were converted into dummy variables. Using min-max normalization, the feature scaling was done on the dataset. Finally, with sklearn, 80% of the data was split into training set and the remaining 20% into test set. On the training set, many machine learning models were trained and tested using the test set.

Findings and Results:

1. **Naive Bayes:**

This is the first classification model that we used, and we got the test accuracy of 86.88%.

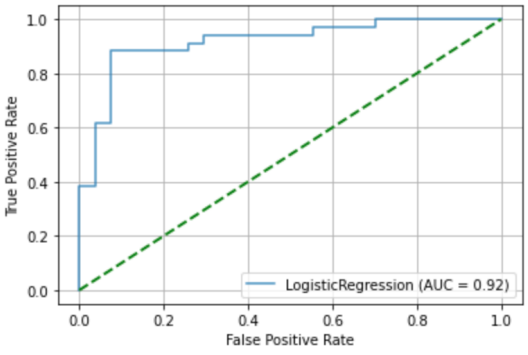
	precision	recall	f1-score
0	0.85	0.85	0.85
1	0.88	0.88	0.88
accuracy			0.87
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.87	0.87



2. **Logistic Regression:**

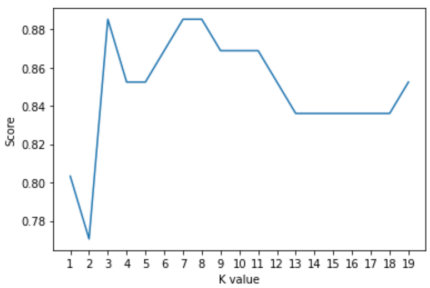
After Naive Bayes, we tried Logistic Regression and since it is a classification problem we believed it would perform better. Our first approach was to create the model from scratch, without using any ML libraries. We decided to tune our hyperparameters like the weights for the sigmoid function using Gradient Descent. We used Forward Backward propagation method, for calculating weights and Bias values and then updated their values and passed them along with the test data for prediction. We got the same accuracy of 86%, the same as the one we got using Sklearn library.

	precision	recall	f1-score
0	0.85	0.85	0.85
1	0.88	0.88	0.88
accuracy			0.87
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.87	0.87



3. **KNN:**

This is the next classification algorithm, and initially we decided to take the value of n = 2, and fitted the train data in the model. We got a test accuracy of 77.05%, which is pretty low. In order to find the best K value, we decided to iterate from 1 to 20. For each iteration we are fitting the train data and train label and appending the score in our array so that in future we can compare the scores and decide the best value of n in KNN.

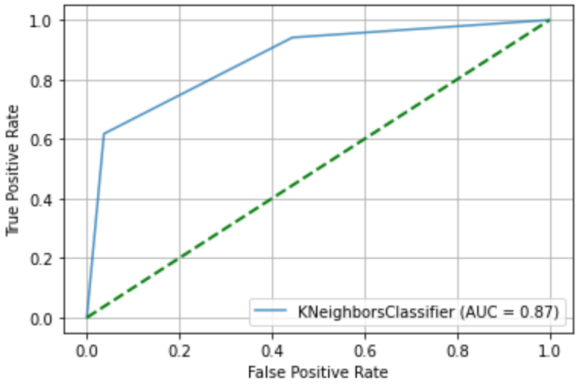



```
Maximum KNN Score is 88.52%
```

As you can see above if we define k as 3-7-8 we will reach maximum score.

We plot the K value versus Score to see which value is better for our model fitting. We saw that using 3,7, or 8 reaches the maximum score of 0.88. Next we fit our model with K=3. We got a test accuracy of 88%.

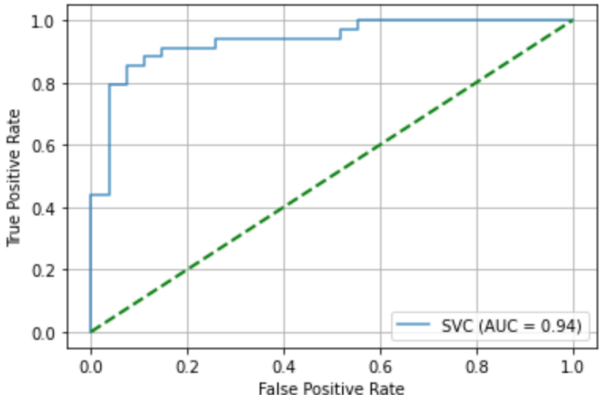
	precision	recall	f1-score
0	0.88	0.85	0.87
1	0.89	0.91	0.90
accuracy			0.89
macro avg	0.89	0.88	0.88
weighted avg	0.89	0.89	0.88



4. SVM:

In the Support Vector Machine, we are using the default kernel or the Radial Basis function. We got a test accuracy of 88%

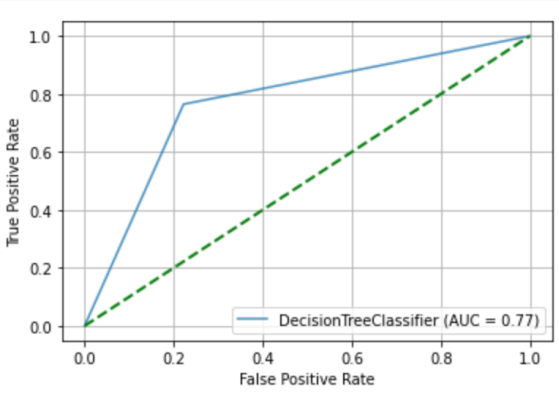
	precision	recall	f1-score
0	0.88	0.85	0.87
1	0.89	0.91	0.90
accuracy			0.89
macro avg	0.89	0.88	0.88
weighted avg	0.89	0.89	0.88



5. Decision Tree:

The Decision tree model gives a test accuracy of 77.04%.

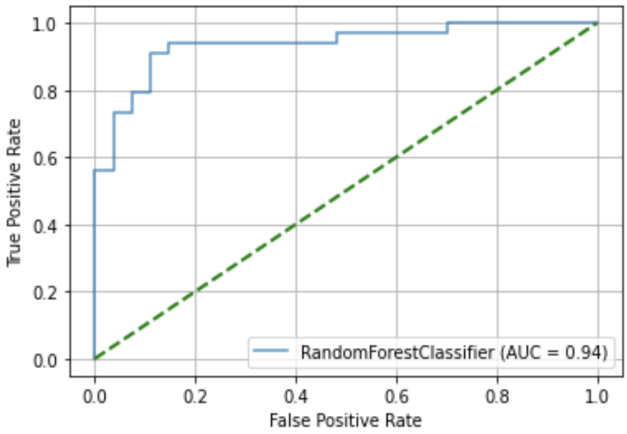
	precision	recall	f1-score
0	0.73	0.81	0.77
1	0.84	0.76	0.80
accuracy			0.79
macro avg	0.79	0.79	0.79
weighted avg	0.79	0.79	0.79



6. Random Forest:

For this model we used 1000 trees. We got a training accuracy of 100% and test accuracy of 88.52%.

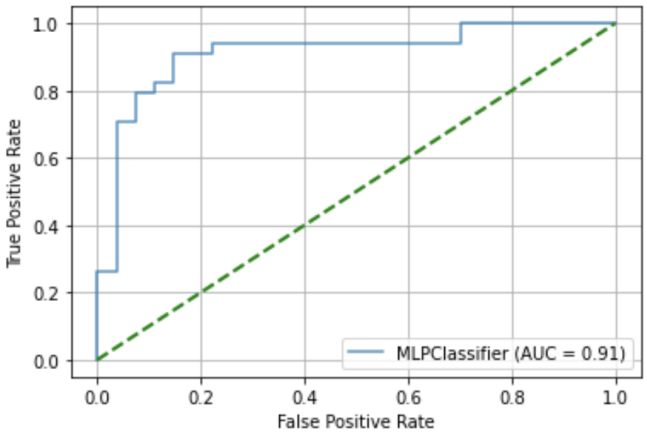
	precision	recall	f1-score
0	0.88	0.85	0.87
1	0.89	0.91	0.90
accuracy			0.89
macro avg	0.89	0.88	0.88
weighted avg	0.89	0.89	0.88



7. Neural Network:

We used a multilayer perceptron model with 2 hidden layers, first layer consisting of 150 units and second layer with 100 units. The activation function we used is the tanh function. This model gave us a training accuracy of 85.95 % and a test accuracy of 86.89%

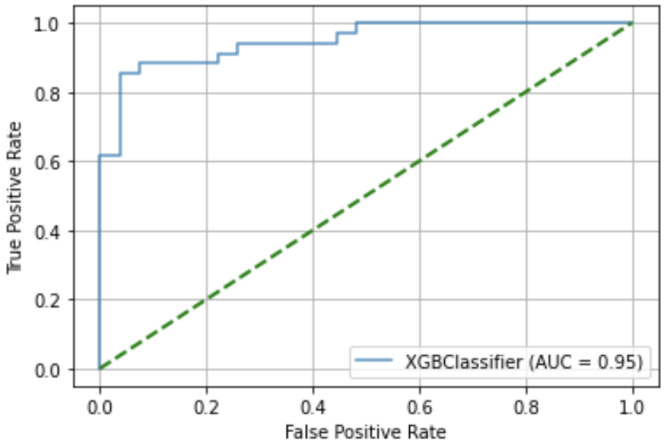
	precision	recall	f1-score
0	0.88	0.81	0.85
1	0.86	0.91	0.89
accuracy			0.87
macro avg	0.87	0.86	0.87
weighted avg	0.87	0.87	0.87



8. XGBoost:

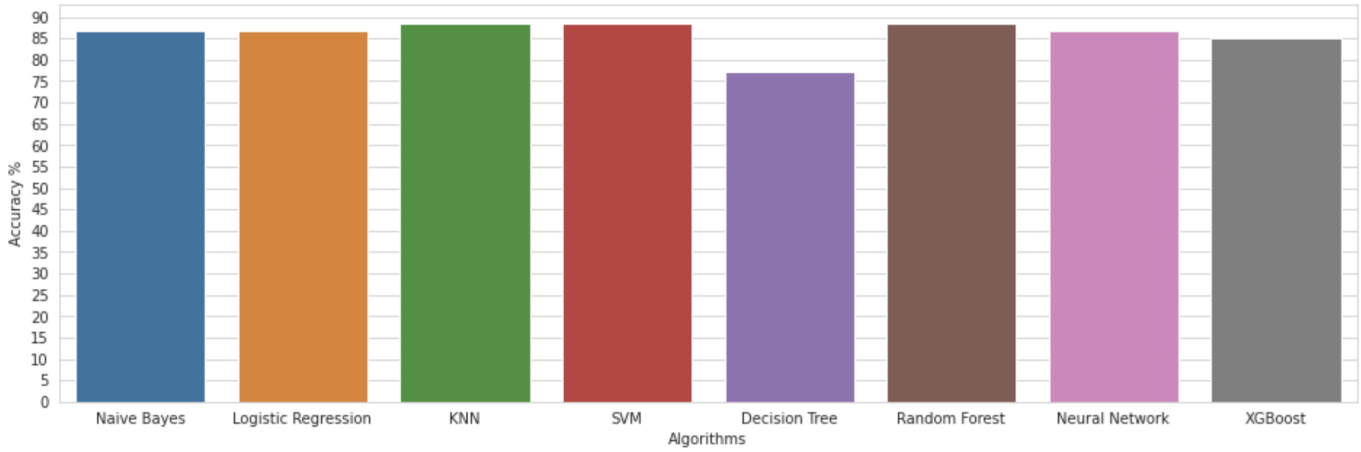
We implemented this model with the default parameters. It gave a training accuracy of 98.35% and a test accuracy of 85.25%.

	precision	recall	f1-score
0	0.85	0.81	0.83
1	0.86	0.88	0.87
accuracy			0.85
macro avg	0.85	0.85	0.85
weighted avg	0.85	0.85	0.85



Conclusion:

We first compared our models based on the test set accuracies. However, KNN, SVM and Random Forest models all have the same highest value 88.52%, followed by Logistic Regression, Naive Bayes and Multilayer Perceptron models with test accuracy 86.89% and the remaining two models, XGBoost and Decision Tree have accuracies of 85.25% and 77.05% respectively.



We also compared our models based on their AUC scores. Since we are doing a binary classification, the area under the ROC curve is a good metric to assess the performance of a model as it measures the performance of a binary classifier averaged across all possible decision thresholds. The table on the right shows the corresponding AUC values of the models. XGBoost classifier performed the best while the decision tree gave the lowest performance compared to other models.

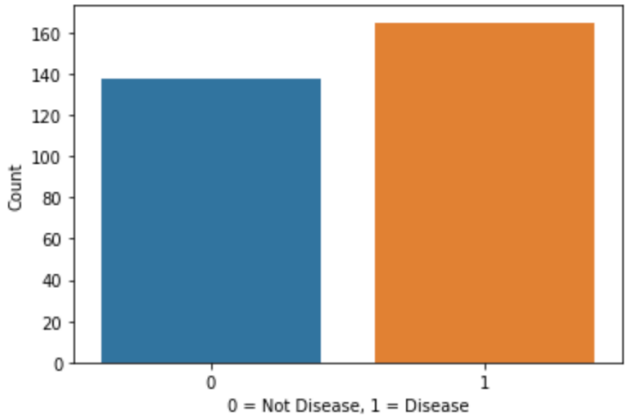
Model	AUC
Naive Bayes	0.90
Logistic Regression	0.92
KNN	0.87
SVM	0.94
Decision Tree	0.79
Random Forest Classifier	0.94
Neural Network	0.91
XGBoost	0.95

Appendix:

Section 1: EDA

Following is the exploratory data analysis of our data

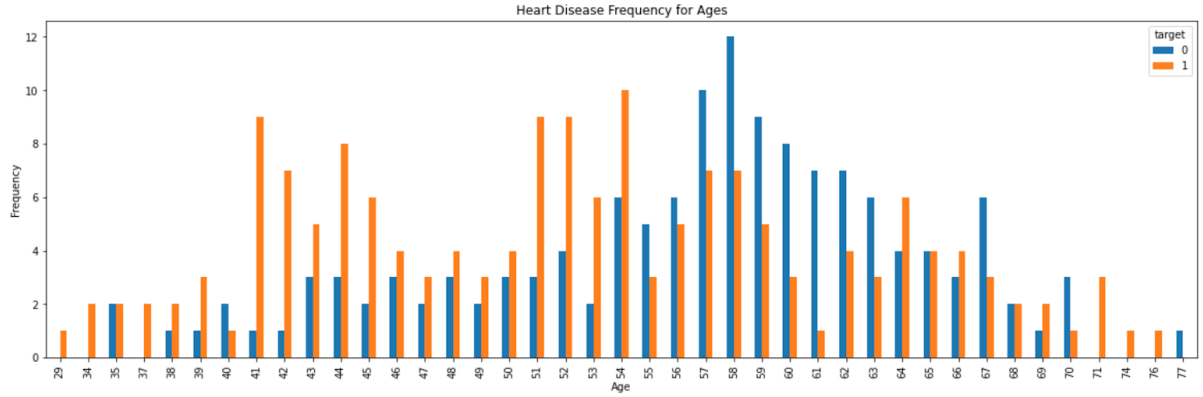
1. The distribution of number of records having or not having a disease



0 = Not Disease
1 = Disease

Analysis:
The data contains almost same amount of both types of records

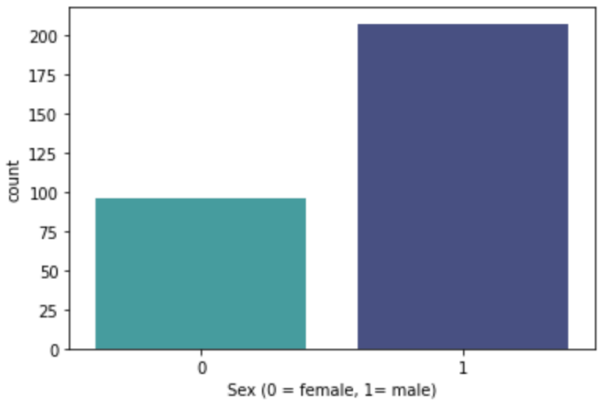
2. Heart Disease Frequency across the ages



0 = No Disease
1 = Disease

Analysis:
As we can see, the decade of 50 has the most records in it with 58 on peak of having a heart disease

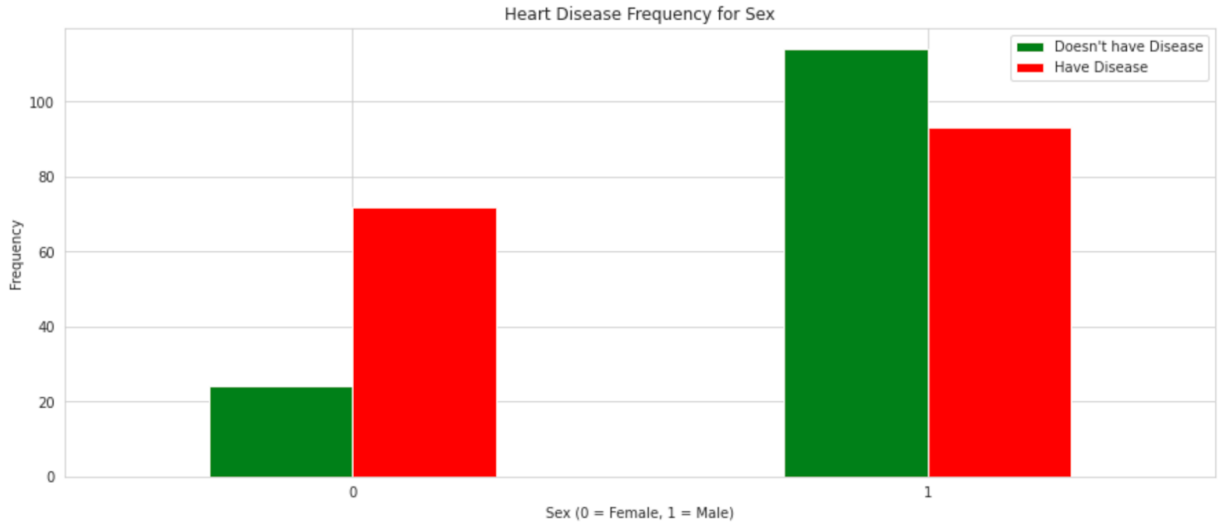
3. Ratio of Male and Female records



0 = Female
1 = Male

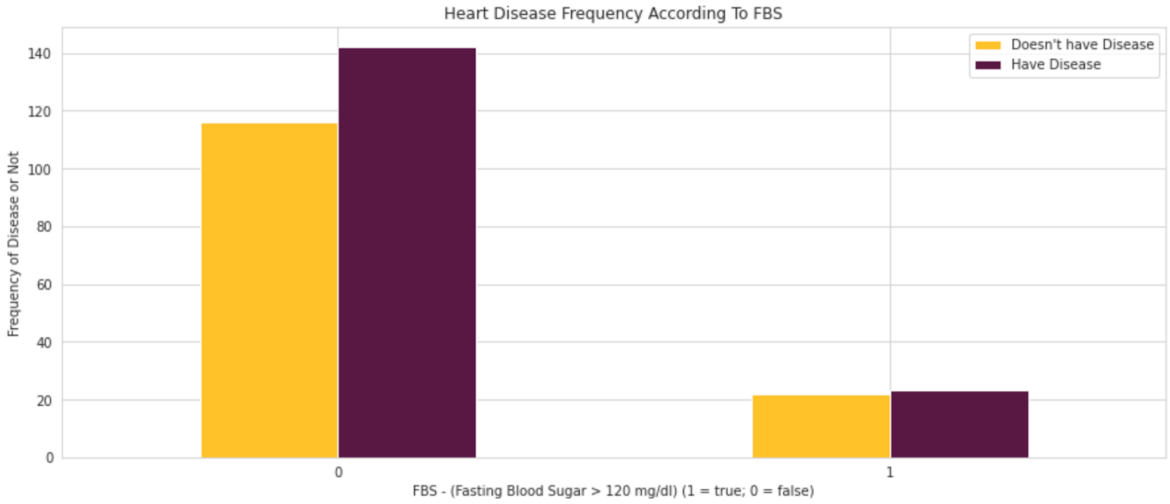
Analysis:
We have more records for Male population than Female

4. Heart Disease Frequency across gender



Analysis:
Males have more chances of having heart disease compared to females

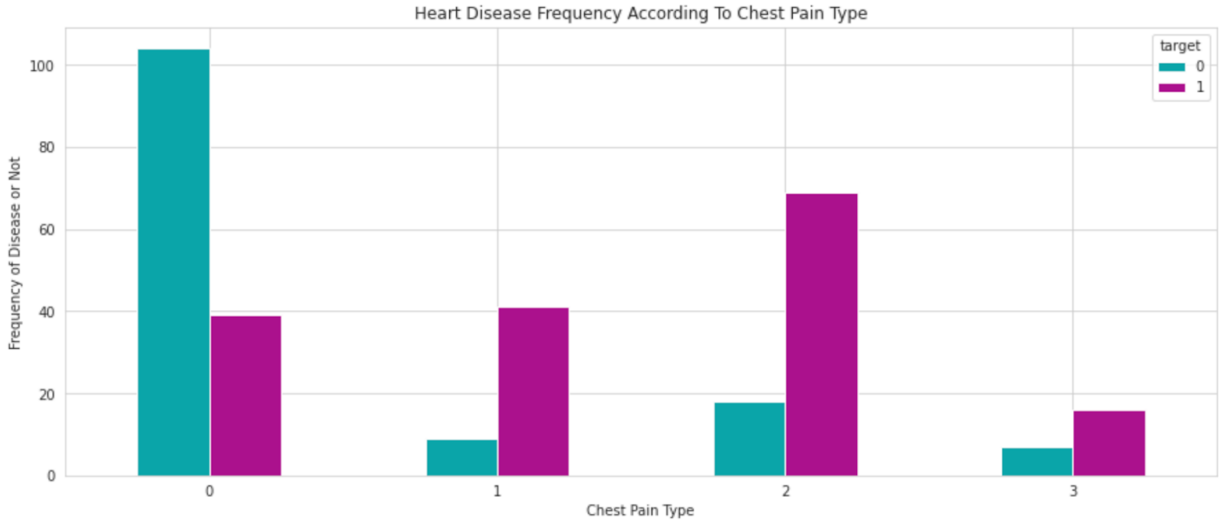
5. Heart Disease Frequency based on Fasting Blood Sugar



Analysis:

We can see that fasting blood sugar go hand in hand with heart disease

6. Heart Disease Frequency according to chest pain type



0 = Typical Angina
1 = Atypical Angina
2 = Non-anginal pain
3 = Asymptotic

Analysis:

We can see that chances of having a heart disease are higher for the type 2 pain in chest, i.e. Non-anginal pain