Heart Disease Prediction

BY AHMED ALI BAIG

What is heart disease?

- Heart diseases is a term covering any disorder of the heart. Heart diseases have become a major concern to deal with as studies show that the number of deaths due to heart diseases have increased significantly over the past few decades in India, in fact it has become the leading cause of death in India.

Introduction

- A study shows that from 1990 to 2016 the death rate due to heart diseases have increased around 34 per cent from 155.7 to 209.1 deaths per one lakh population in India.
- By leveraging existing data and employing supervised learning algorithms, such as logistic regression, decision trees and etc.
- We can effectively train the model to identify patterns and risk factors associated with heart disease. Once trained, this model can accurately predict the likelihood of individuals developing heart disease based on their unique characteristics and medical history

Problem Statement

- It has become increasingly imperative to prevent heart diseases. Effective data-driven systems for predicting such conditions can significantly enhance research and prevention efforts, thereby ensuring that a greater number of individuals can enjoy healthier lives. This is where the application of Machine Learning becomes essential. Machine Learning facilitates accurate predictions of heart diseases, thereby playing a crucial role in proactive healthcare interventions.

About Dataset

- The dataset consists of 303 individuals data with 14 columns in the dataset.

1. Age : displays the age of the individual.

2. Sex : displays the gender of the individual using the following format : 1 = male 0 = female.

3. Chest-pain type : displays the type of chest-pain experienced by the individual using the following format : 1 = typical angina 2 = atypical angina 3 = non - anginal pain 4 = asymptotic

4. Resting Blood Pressure : displays the resting blood pressure value of an individual in mmHg (unit)

5. Serum Cholestrol : displays the serum cholestrol in mg/dl (unit)

6. Fasting Blood Sugar : compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false)

7. Resting ECG : 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy

8. Max heart rate achieved : displays the max heart rate achieved by an individual.

9. Exercise induced angina : 1 = yes 0 = no

10. ST depression induced by exercise relative to rest : displays the value which is integer or float.

11. Peak exercise ST segment : 1 = upsloping 2 = flat 3 = downsloping

12. Number of major vessels (0-3) colored by flourosopy : displays the value as integer or float.

13. Thal : displays the thalassemia : 3 = normal 6 = fixed defect 7 = reversable defect

14. Diagnosis of heart disease : Displays whether the individual is suffering from heart disease or not : 0 = absence 1,2,3,4 = present.

Matrics

- Our data is quit balanced in terms of target variable so we going to approach are model with accuracy score instead of F1 score.

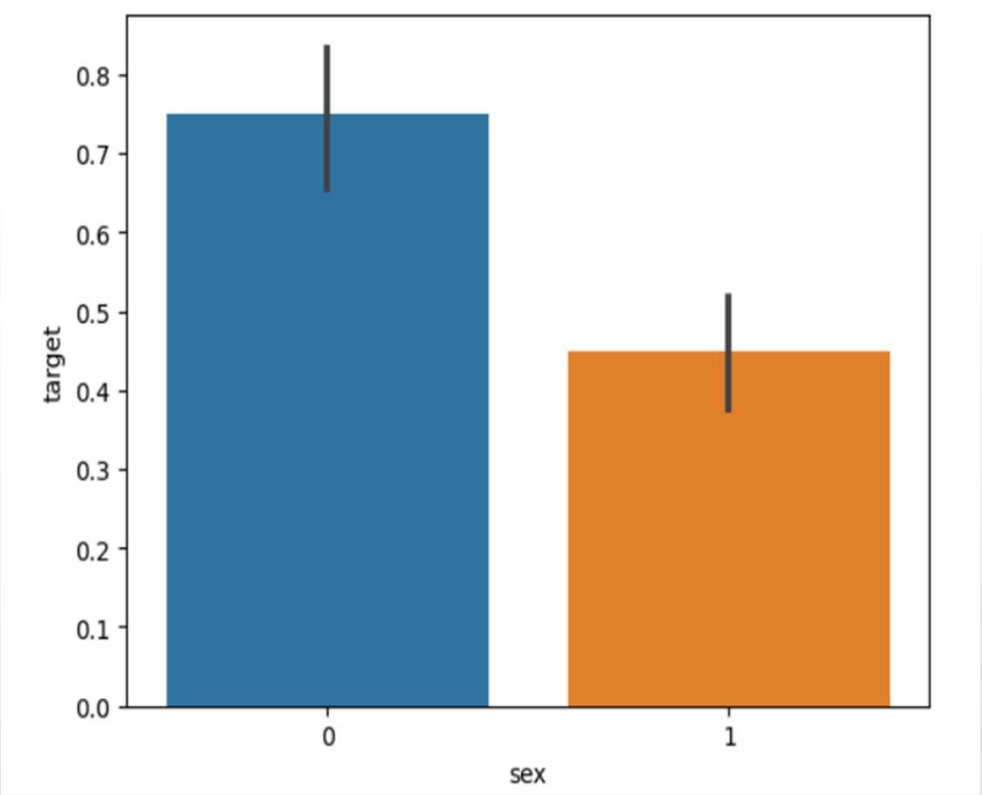
```
Patience without heart problems:  138
Patience with heart problems:  165
```

```
target      1.000000
exang       0.436757
cp          0.433798
oldpeak     0.430696
thalach     0.421741
ca          0.391724
slope       0.345877
thal        0.344029
sex         0.280937
age         0.225439
trestbps    0.144931
restecg     0.137230
chol        0.085239
fbs         0.028046
Name: target, dtype: float64
```

This shows that most columns are moderately correlated with target, but 'fbs' is very weakly correlated.

Exploratory data analysis

Correlation with respect to dependent variable (target)

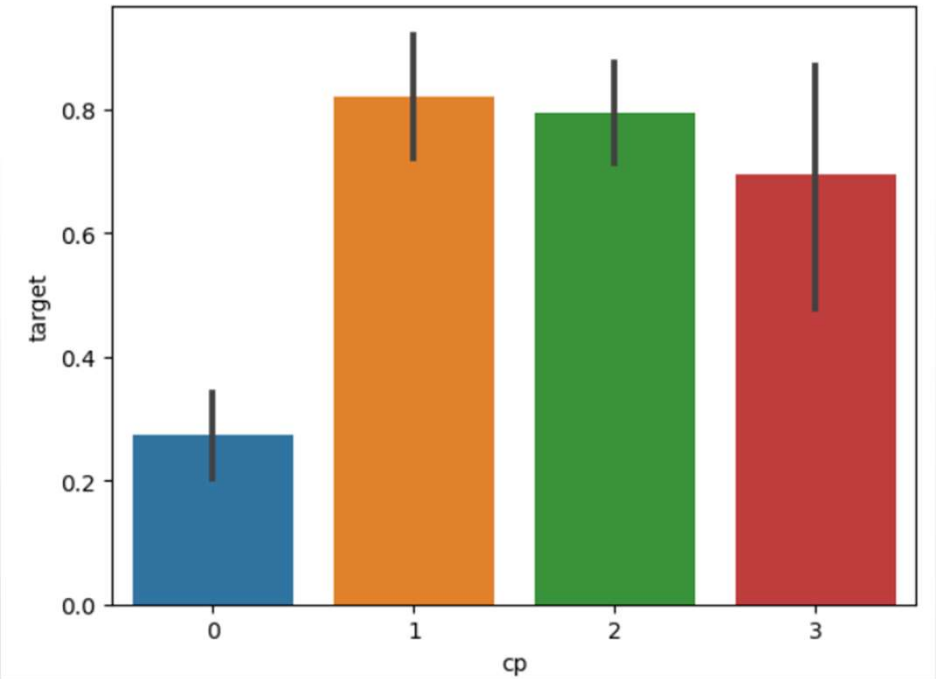


We notice, that females are more likely to have heart problems than males

Exploratory data analysis

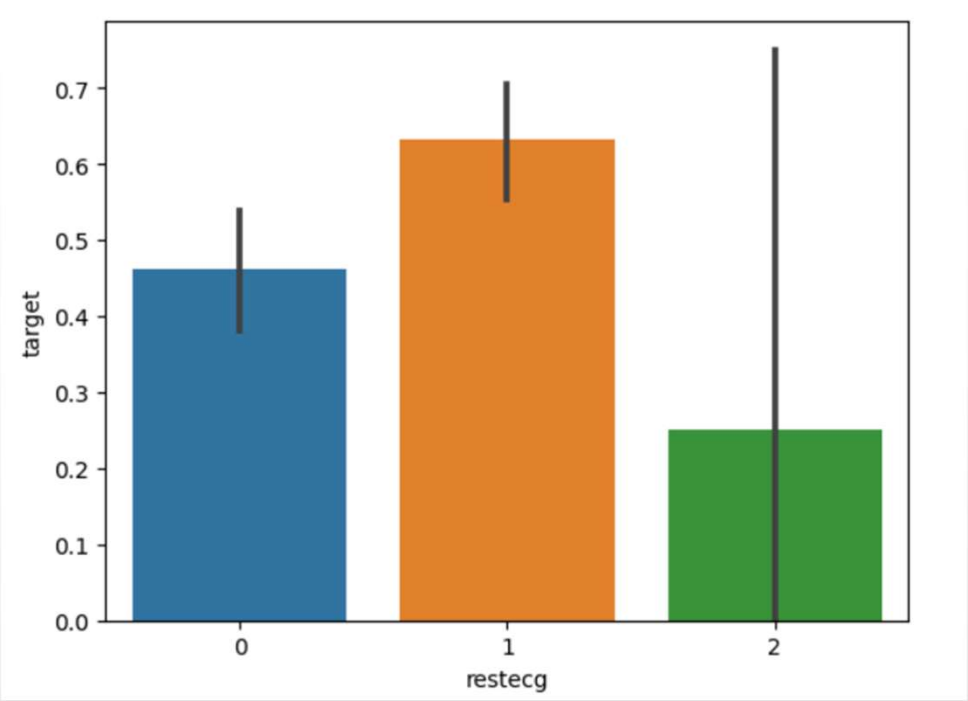
Barplot is used to find the insight between the dependent and indepent variable

Exploratory data analysis



We notice, that chest pain of '0', i.e. the ones with typical angina are much less likely to have heart problems

Exploratory data analysis



We realize that people with restecg '1' and '0' are much more likely to have a heart disease than with restecg '2'

Model training

- We train our model by analyzing existing data, as we already have information on each patient suffering from heart disease. This process is commonly referred to as supervised learning. The trained model can then be utilized to predict whether users are likely to suffer from heart disease.
- First, data is divided into two parts using component splitting. In this experiment, data is split based on a ratio of 80:20 for the training set and the prediction set. The training set data is used in the logistic regression component for model training, while the prediction set data is used in the prediction component.
- The following classification models are used - Logistic Regression, Random Forest Classfier, SVM, Naive Bayes Classifier, Decision Tree Classifier, XGBoost

Model Evalution

- XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. which attempts to accurately predict a target variable by combining the estimates of a set of model.
- In the final conclustion xgboost showed the awesome result of trained dataset which is know to the model.

```
Accuracy for training set for XGBoost = 1.0
Accuracy for test set for XGBoost = 0.8360655737704918
```

- Where as random forest is much better in both cases (training and test data) with an accuracy of 98% and test dataset accuracy with 86% which is quit better then xgboost

```
Accuracy for training set for Random Forest = 0.987603305785124
Accuracy for test set for Random Forest = 0.8688524590163934
```

Conclusion

- The development of a heart disease prediction model using machine learning techniques offers promising prospects for improving healthcare outcomes.
- Such predictive capabilities hold significant potential for early detection, personalized intervention, and ultimately, better patient outcomes in the realm of cardiovascular health aka health disease.