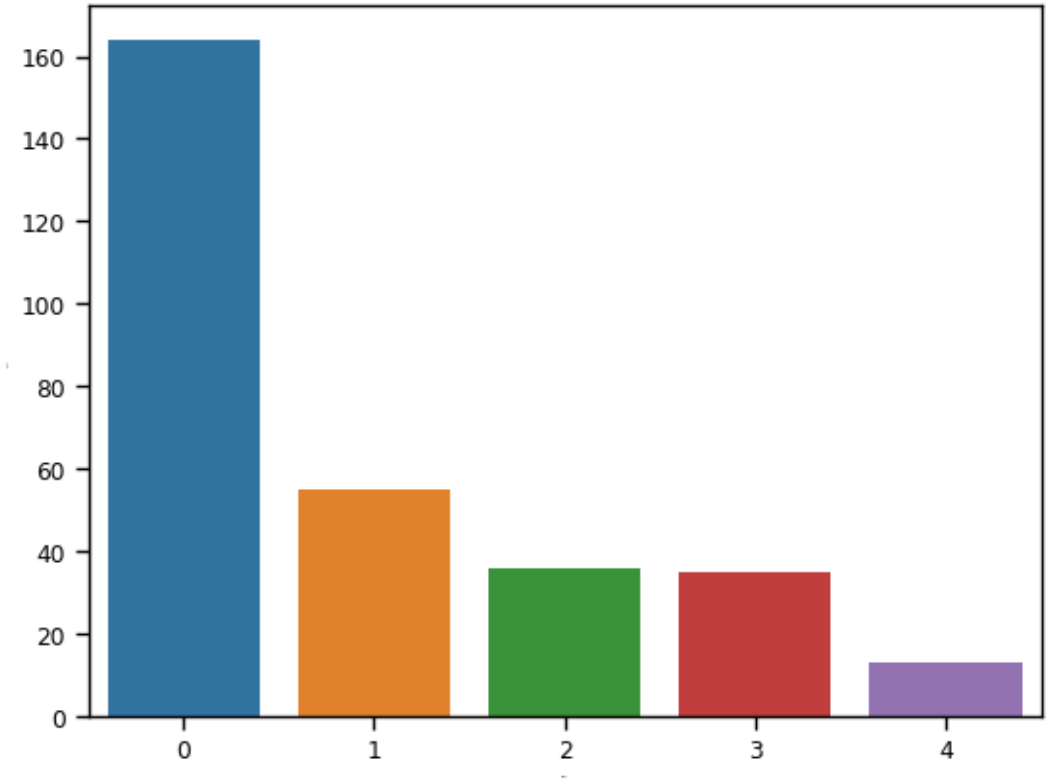
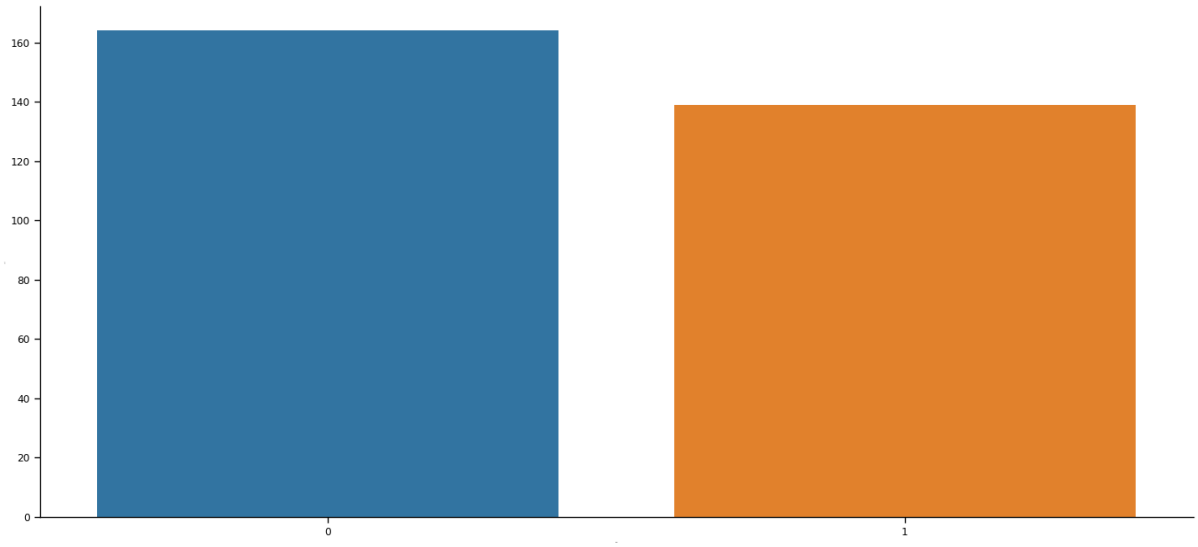
Predicting Heart Disease

1. Data Exploration

- There are 5 heart diseases labels. 0 – no disease, 1-4 varying degree of disease.



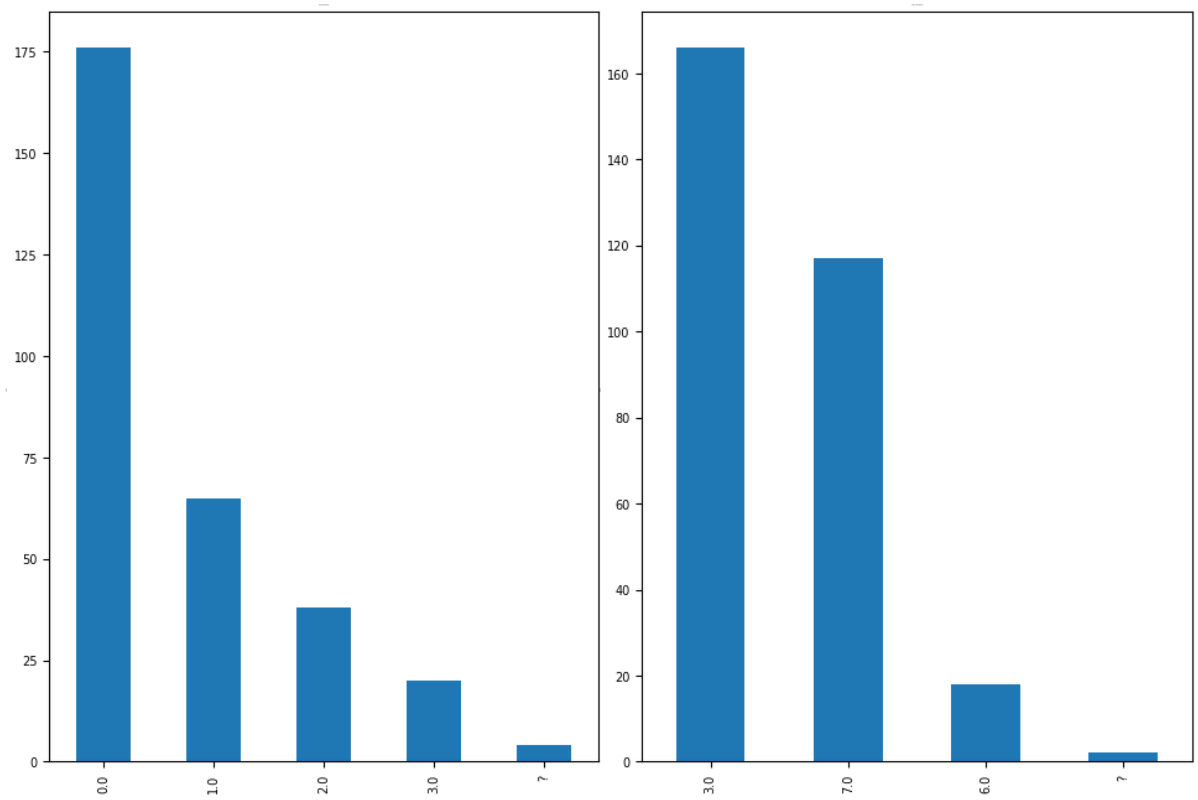
Then Convert it into binary classification 0 – no disease, 1 – patient has a disease.



There is a label imbalance in the dataset. Then resample the dataset.

Before resampling there are 164 0 samples and 139 1 samples. After resampling there are 164 0 samples and 164 1 samples.

- There was unknown (?) data in categorical data.



Then replace them with most frequent data.

- Use One Hot encoding for categorical data.
- Use standard Scaling for scale data.

2. Data Analysis

- After plotting boxplot there were multiple outliers in chol, trestbps columns.

- Plot scatter plots for
    - age vs thalach
    - age vs oldpeak
    - age vs chol
    - age vs trestbps . There were not trends or cycles or clusters to identify.

- Plot correlation matrix to identify correlation between features. Then there was not high correlated features (correlation rate like 0.7 or 0.8).

- In thalach feature there is a skew to positive side.

3. Data Modeling

Yes, we can. After converting this problem to binary classification, we can use logistic regression to solve this problem. In the end of the notebook, I provided. I tried to train the logistic regression model. It provided 96.97% accuracy on the test data.

4. Report

In the model training has tried following algorithms,

1. Sklearn Ensemble RandomForest

   Train accuracy - 0.95

   Test accuracy – 0.96

2. Gradient Boost

   Train accuracy - 0.98

   Test accuracy – 0.96

3. CatBoost

   Train accuracy - 0.95

   Test accuracy – 0.96

4. Tensorflow RandomForest

   Train accuracy - 0.91

   Test accuracy – 0.87

5. Tensorflow Neural Network.

   Train accuracy - 0.97

   Test accuracy – 0.96

After Standard Scaling, one hot encoding and resample imbalance labels accuracy was improved. Especially after resample imbalance labels precision and recall was improved.

Example:

Random Forest

Confusion matrix before resampling is,

	True Positive	True Negative
Predicted Positive	27	4
Predicted Negative	2	28

After resampling,

	True Positive	True Negative
Predicted Positive	30	0
Predicted Negative	2	34

False Positives became 0

False Negative did not change. Then accuracy improved.

Hyper Parameter tuning.

- Use Grid SearchCv to find best hyperparameters for Sklearn Ensemble RandomForest
- For Tensorflow RandomForest - Training a model with automated hyper-parameter tuning and automatic definition of the hyper-parameters
- For Neural network, Catboost and GBBoost used manual hyperparameter tuning.