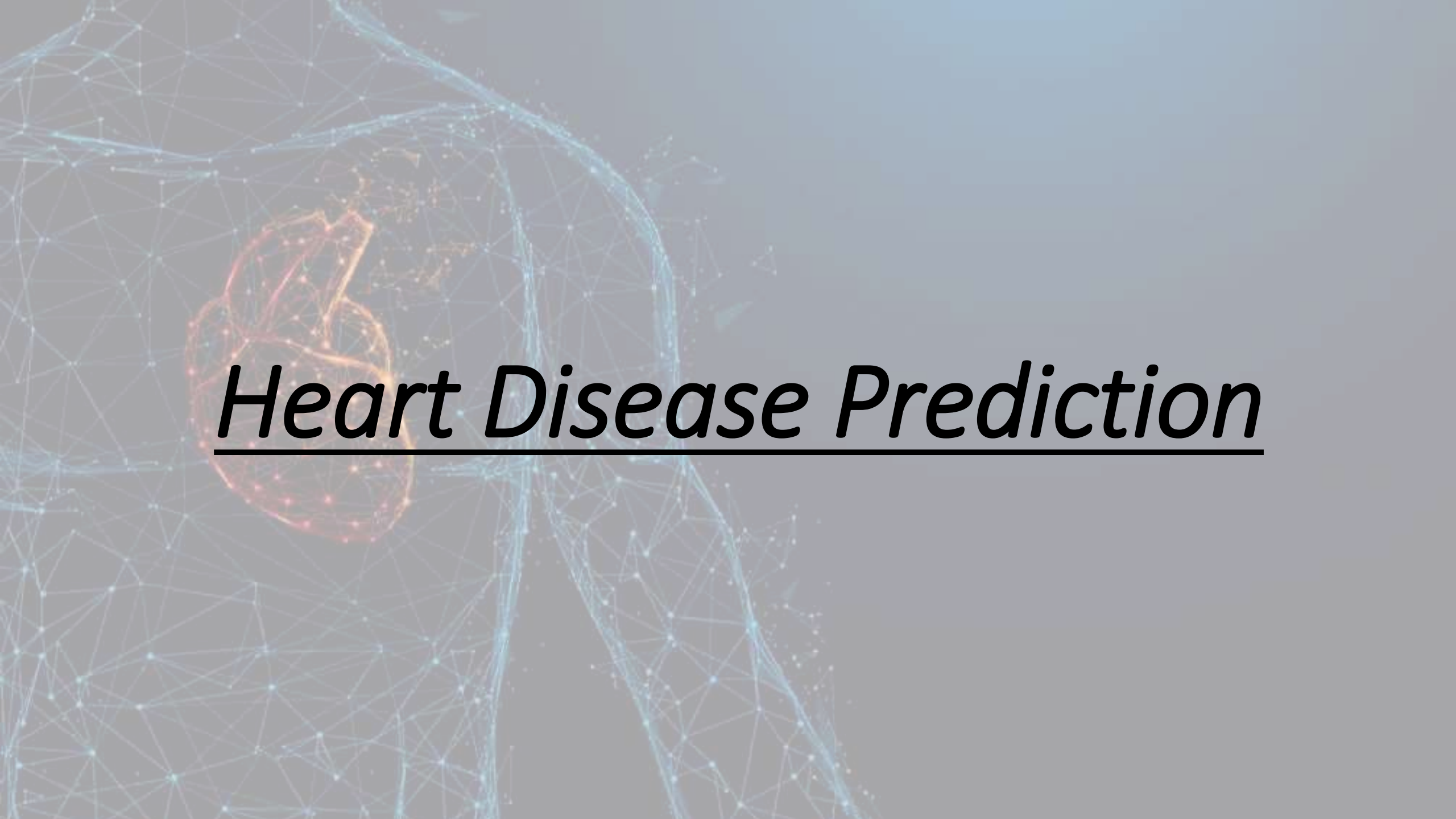


Heart Disease Prediction

Content

- Problem Statement
- Objective
- Exploratory Data Analysis (EDA)
- Transformation of data
- Splitting data
- Fitting Different Model
- Comparison of Model
- Combined ROC curve
- Cross Validation & Hyper parameter Tuning
- Conclusion

Problem Statement

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

Objective

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management where in a machine learning model can be of great help.

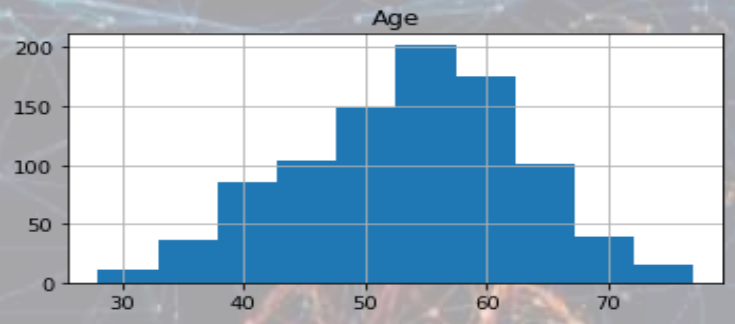
Exploratory Data Analysis(EDA)

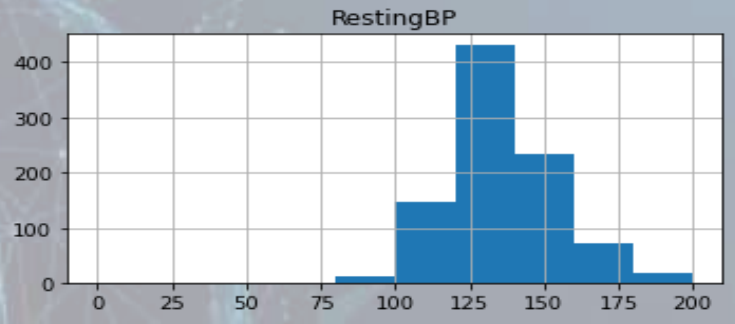
We can see there have some positive & negative correlation, for example

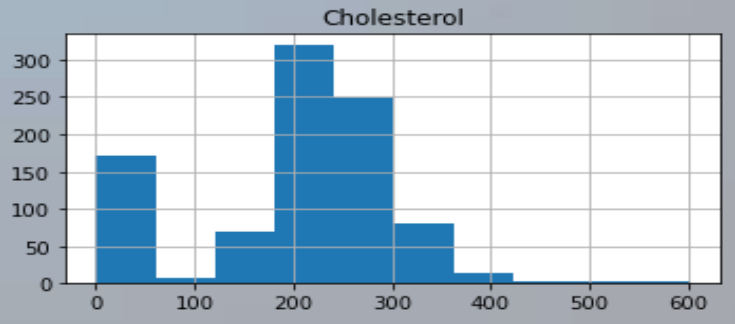
- As age increases, heart disease also increases so there have a positive correlation between them

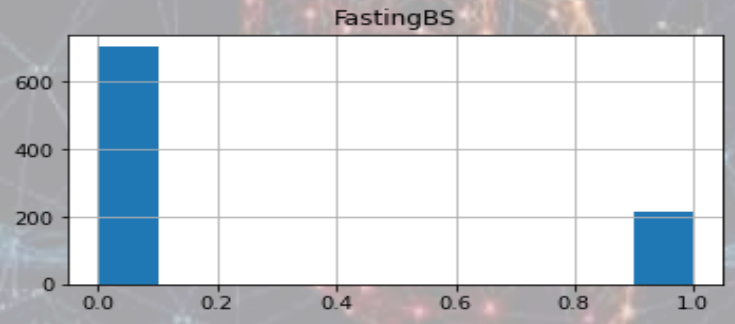
- There have a negative correlation between Fasting Blood Sugar & Cholesterol, maybe both the parameters are not affecting each other directly such as if Fasting Blood Sugar going higher, Cholesterol may goes down or normal

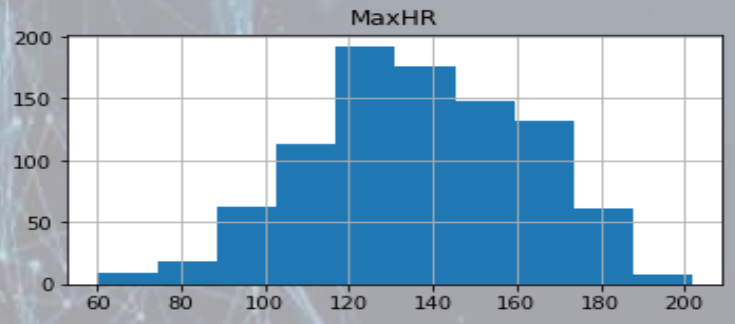
EDA (Continued)

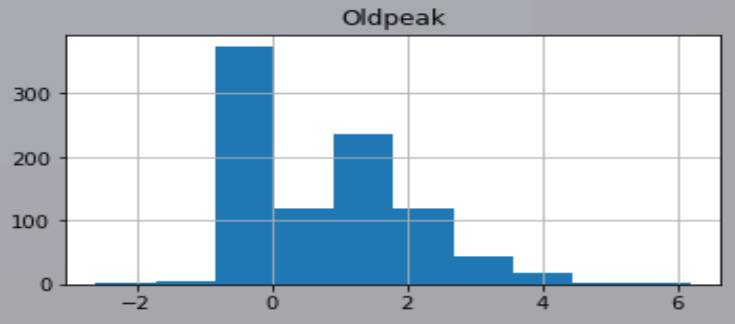


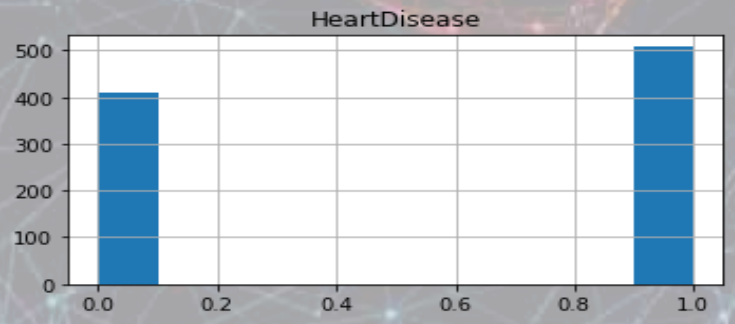






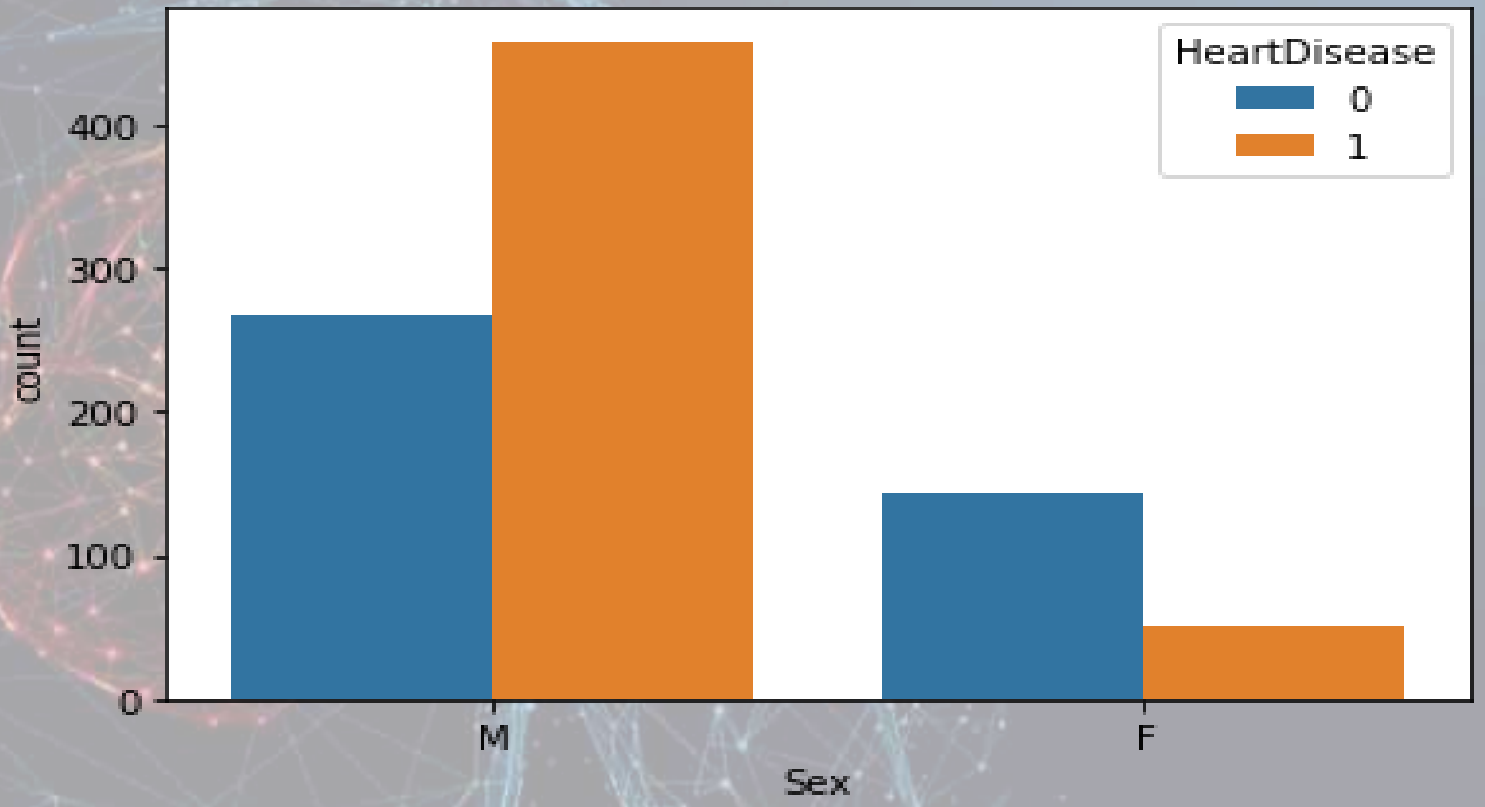






By plotting histogram we can say that only Age column is normally distributed

EDA (Continued)

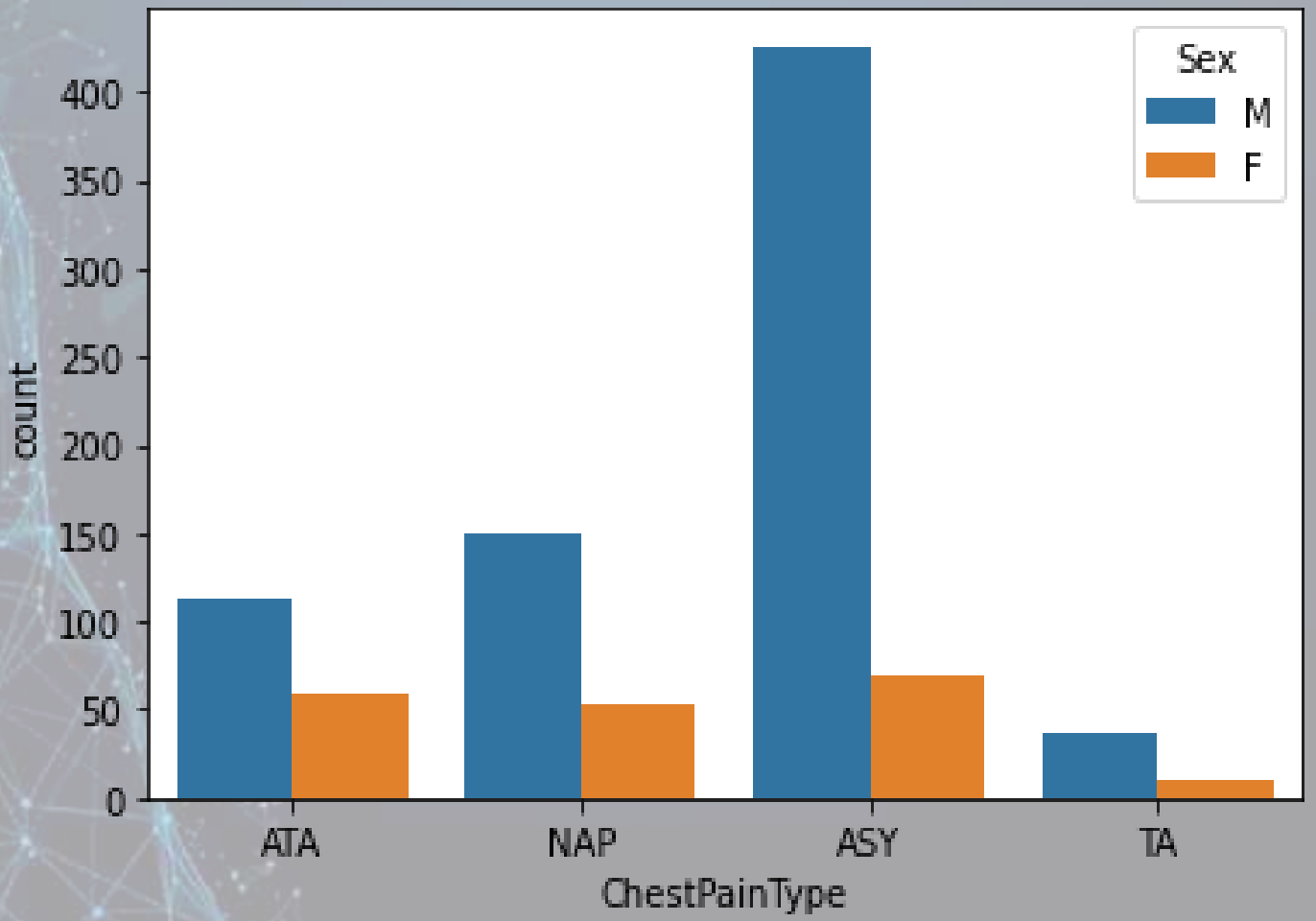


Using count plot we can say males are more affected than females by heart disease

EDA (Continued)

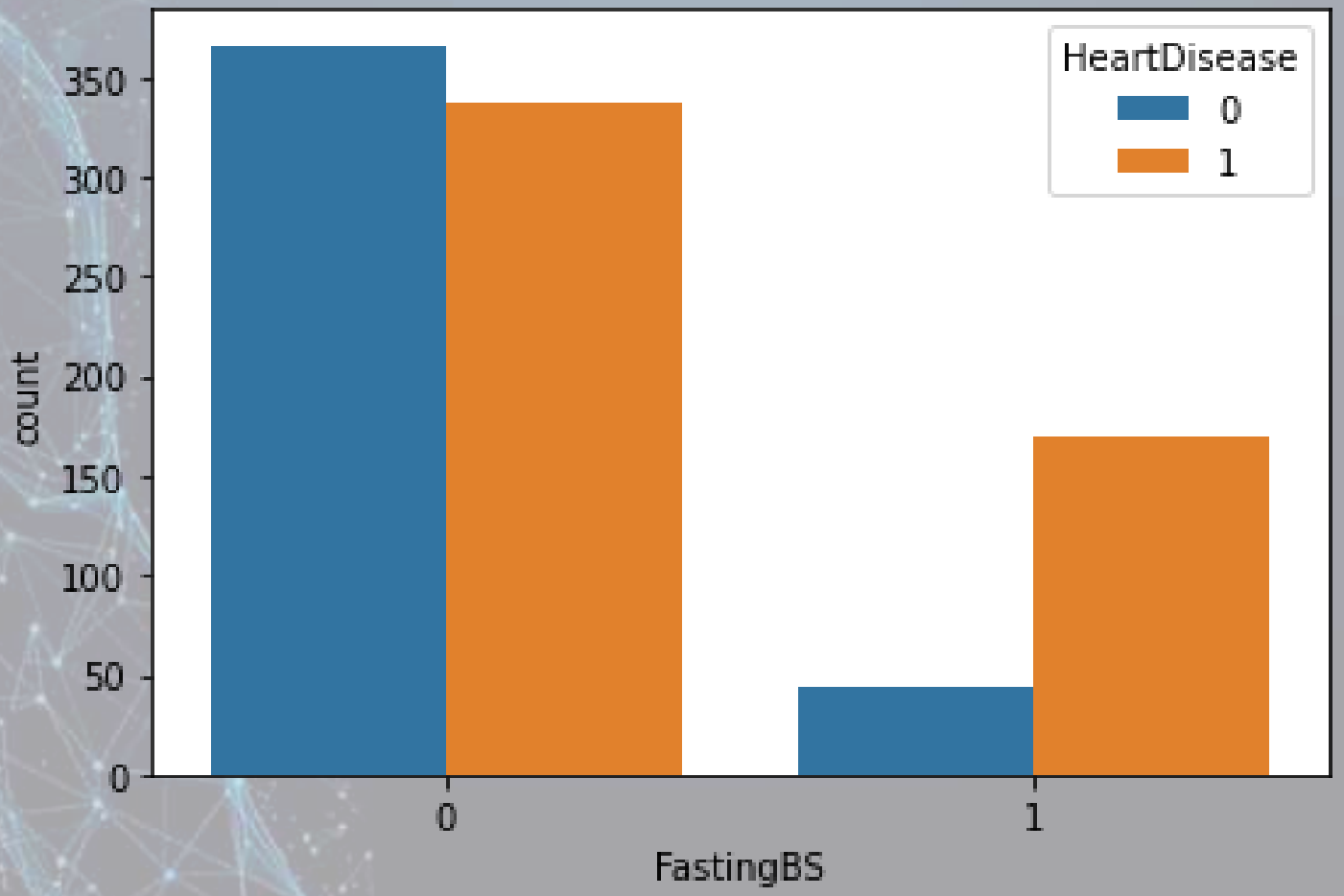
- There have four types of Chest pain, according to our dataset males are hugely affected by the ASY type & lesser affected by the TA type

- For the case of female, this affected rate is very minimum compared with the male



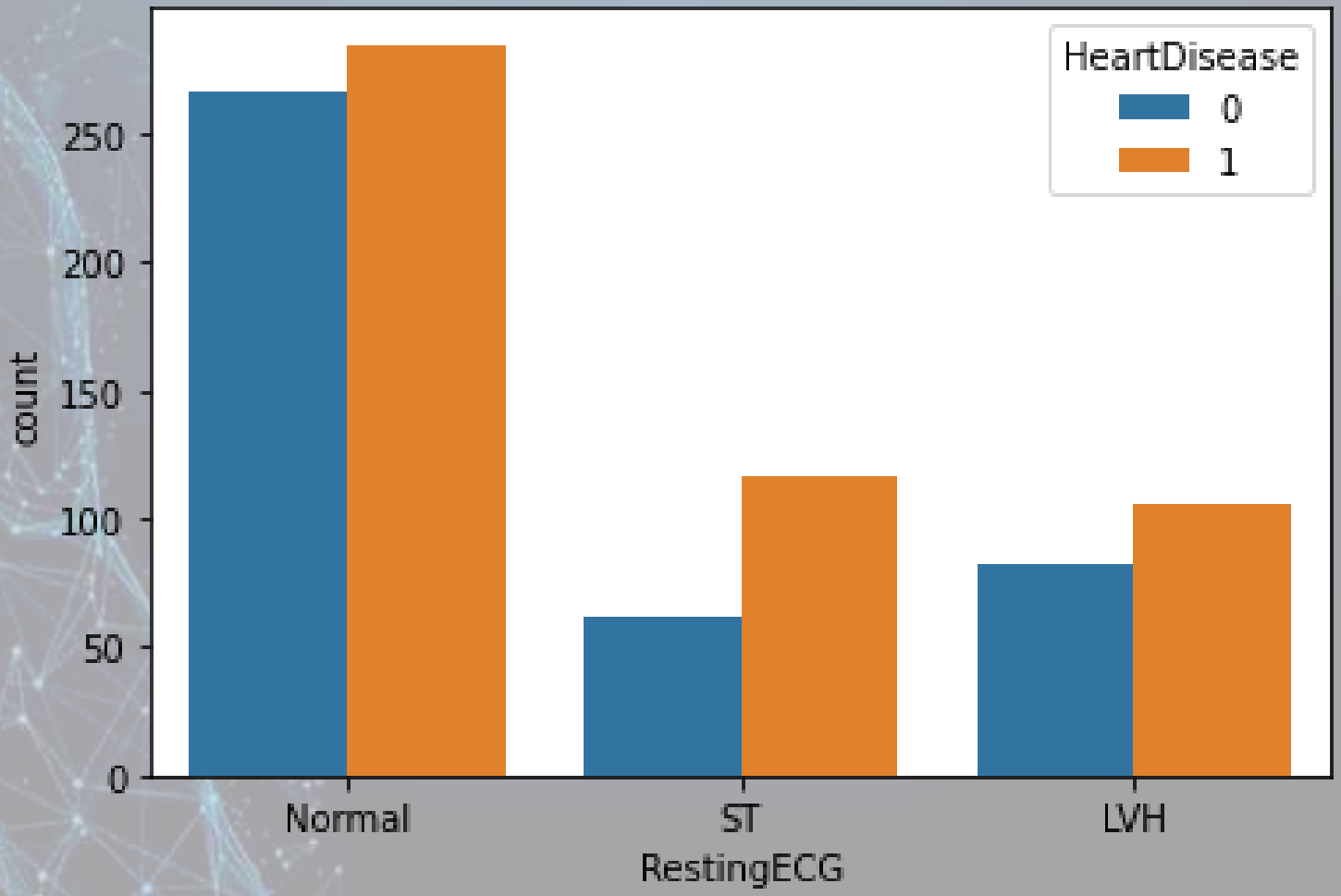
EDA (Continued)

According to our dataset who has FastingBS they are lesser affected by the heart disease, and who has not FastingBS they affected in high amounts by heart disease



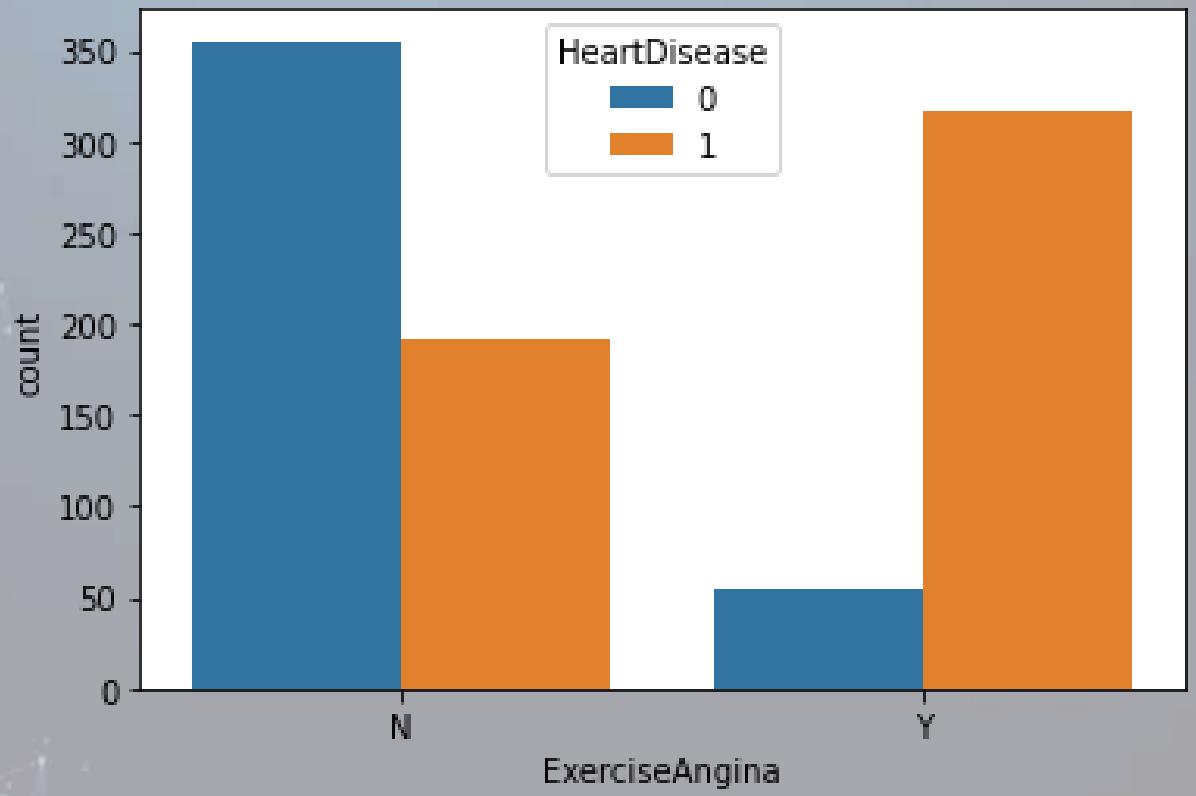
EDA (Continued)

Resting ECG are three types. For who have normal Resting ECG those people has affected more in Heart Disease compared with other two ST & LVH



EDA (Continued)

- Who has Exercise Angina or who do not exercise, they are affected highly in Heart Disease
- Who doing exercise daily, they are affected in lesser amount




```
In [159]: #sorting by age wise who has ExerciseAngina

          df[df.ExerciseAngina == 'Y'].Age.sort_values(ascending = False)

Out[159]: 814    77
          447    77
          556    75
          506    75
          553    74
                 ..
          808    35
          405    35
          696    35
          115    33
          56     31
          Name: Age, Length: 371, dtype: int64
```

Old people has more Exercise Angina than youngers

EDA (Continued)

```
In [148]: #to find max affected people age wise

          df[df.HeartDisease == 1].Age.sort_values(ascending = False)

Out[148]: 814    77
          447    77
          541    76
          506    75
          491    75
                 ..
          119    34
          115    33
          294    32
          76     32
          56     31
          Name: Age, Length: 508, dtype: int64
```

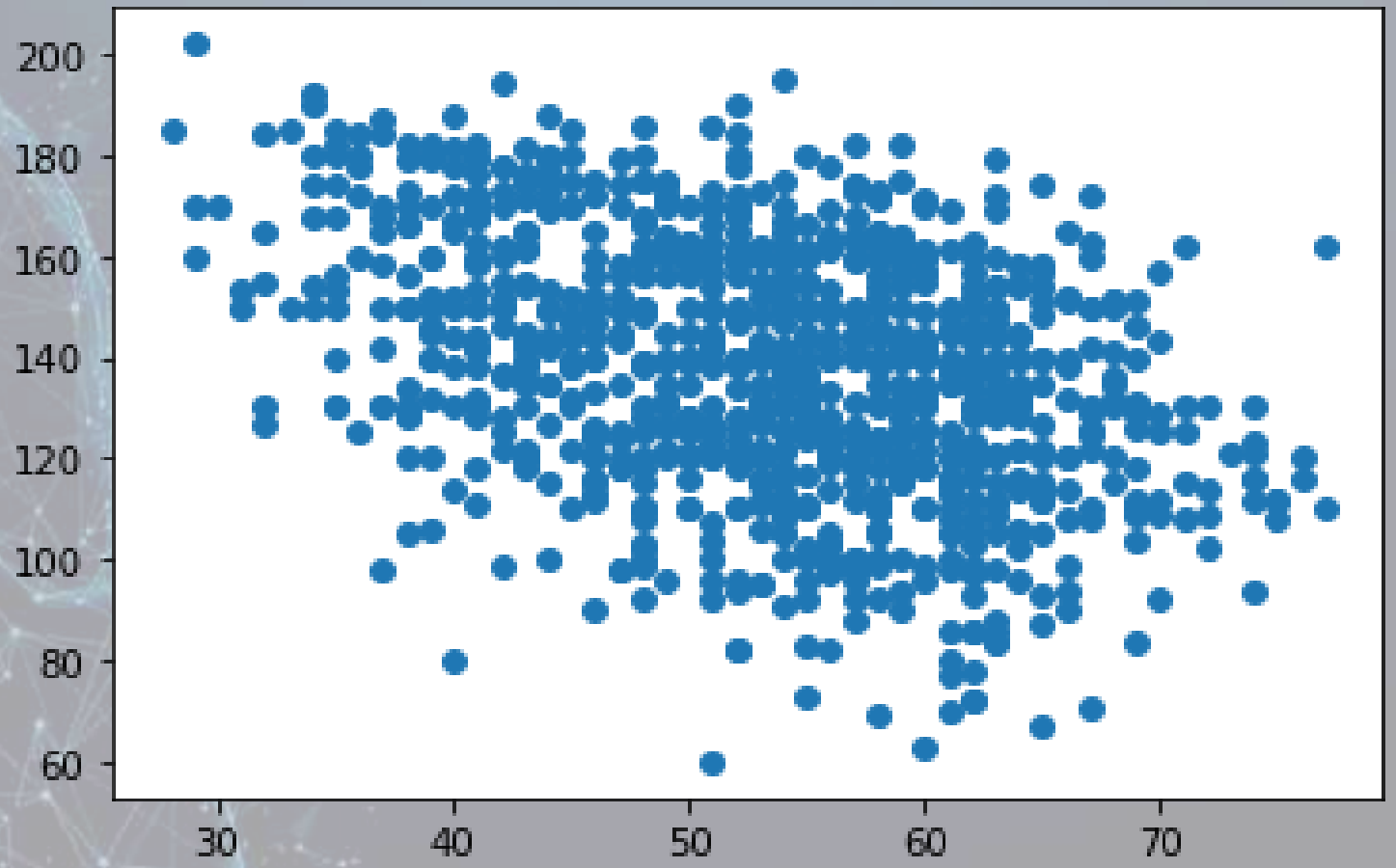
```
In [149]: df[df.HeartDisease == 0].Age.sort_values(ascending = True)

Out[149]: 208    28
          170    29
          219    29
          829    29
          215    30
                 ..
          336    72
          439    74
          619    74
          556    75
          688    76
          Name: Age, Length: 410, dtype: int64
```

- In the age of 77 people got affected more by the heart disease
- On the other hand, who don't have heart disease those age starting from 28

From here we can say that as age increases, people got affected more by the heart disease than younger people.

EDA (Continued)

According to our dataset, we can see that maximum heart rate distributed between the age of 50 to 65



Over all Conclusion from the EDA

- Age wise we can say aged people (more specifically above 50) needs more awareness by changing their lifestyle, food habits, doing exercise everyday, then only they can prevent this Heart Disease

- According to our dataset sex wise males are more affected, so males need more aware about heart disease

- Who do not have blood sugar that's good thing, but they should keep aware about heart disease, because according to our dataset they are affected more in heart disease

- Every person should keep some good habits in their daily life regardless age & sex wise, then only we can prevent heart disease. Because this disease never occurs overnight, this disease occurs depending on the habit of long days

Transformation of data

- To scale data into a uniform format that would allow us to utilise the data in a better way
- For performing fitting and applying different algorithms to it
- The basic goal was to enforce a level of consistency or uniformity to dataset

Splitting data

- Data splits into training dataset and testing dataset
- Training dataset is used to fit the machine learning model
- Test dataset is used to evaluate the fit machine learning model
- Hear 80% of the data taken as training dataset and remaining 20% of dataset used for testing purpose

Fitting Different Model

Following classifiers are used for predicting Heart Disease –

1. K Nearest Neighbor
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Support Vector Machine (SVM)
6. Gradient Boosting
7. XG Boosting

Comparison of Model

```
In [113]: compare_df.sort_values(by=['Test Accuracy'], ascending=False)
```

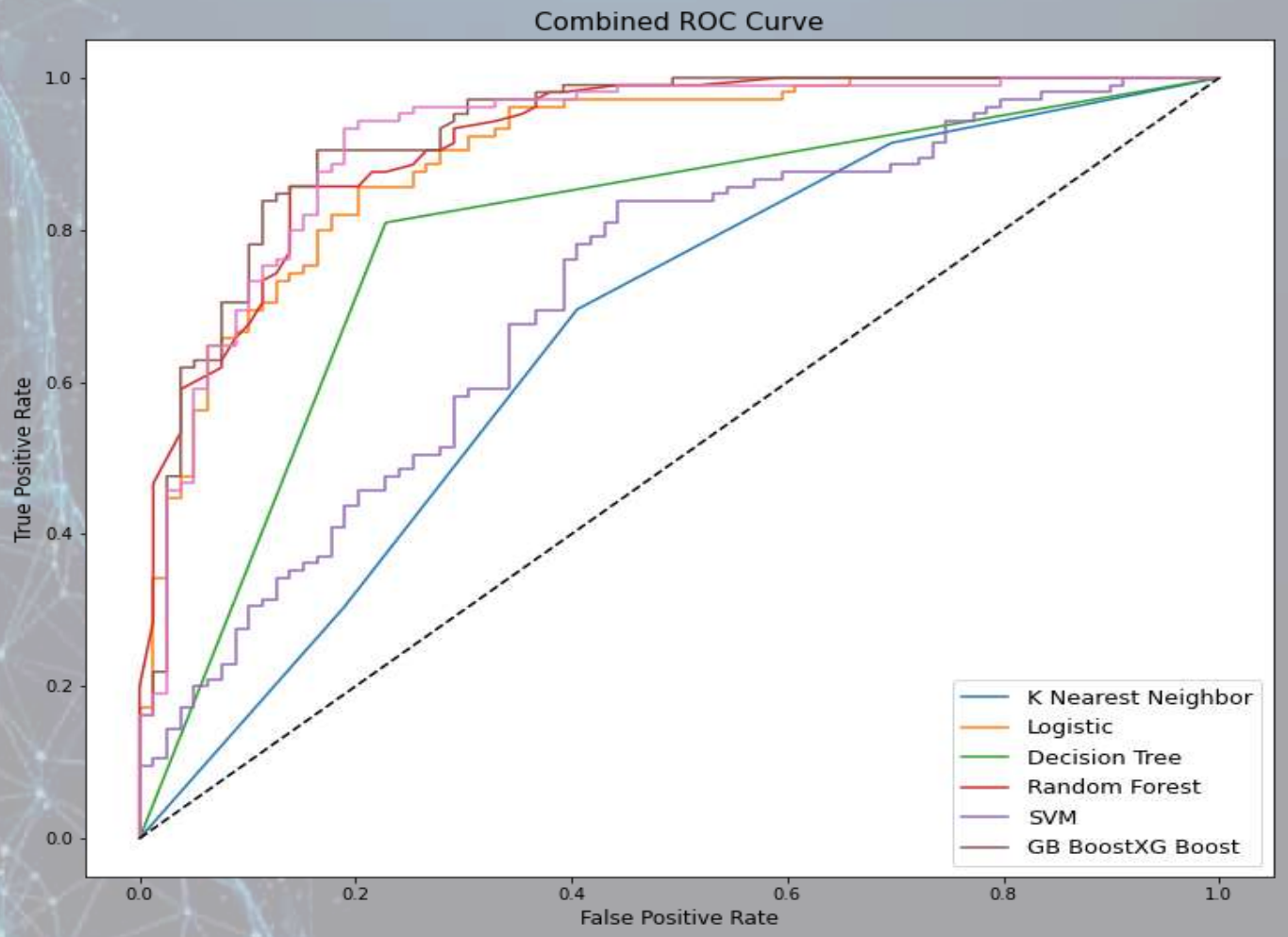
Out[113]:

	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 score	Accuracy
5	Gradient Boosting	0.935	0.870	0.895	0.879	0.887	0.868
3	Random Forest	1.000	0.848	0.857	0.874	0.865	0.844
6	XG Boosting	1.000	0.837	0.838	0.871	0.854	0.833
1	Logistic Regression	0.871	0.826	0.848	0.848	0.848	0.823
2	Decision Tree	1.000	0.793	0.810	0.825	0.817	0.789
4	SVM	0.737	0.658	0.676	0.710	0.693	0.653
0	K Nearest Neighbor	0.820	0.652	0.695	0.695	0.695	0.645

Here we can see that Gradient Boosting classifier shows highest test accuracy and F1 score

Combined ROC curve

- An ROC curve is a graph showing the performance of a classification model at all classification thresholds.
- An ROC curve plots TPR vs. FPR at different classification thresholds.



Cross Validation & Hyper parameter Tuning

- It is a resampling procedure used to evaluate machine learning models on a limited data sample
- Basically, cross validation is a technique using which model is evaluated on the dataset on which it is not trained that is it can be a test data or can be another set as per availability or feasibility
- Tuning the hyper parameters of respective algorithms is necessary for getting better accuracy and to avoid over fitting

Conclusion

I have applied seven different types of classification algorithm on my given dataset to know which algorithm good fit for our dataset & gives us the best accuracy. Before applying Cross validation and hyperparameter tuning Gradient Boosting shows highest test accuracy score of 0.870, F1 score is 0.887 and Accuracy is 0.868, but after applying Cross validation and hyperparameter tuning on the Gradient Boosting algorithm it gives test accuracy score of 0.864, F1 score is 0.88 and Accuracy is 0.861 which is almost same than before. As we know that Cross validation and hyperparameter tuning certainly reduces chances of overfitting and also increases performance of model. So we can conclude that before tuning our model worked well so here no needed to apply tuning as such.


Thank You