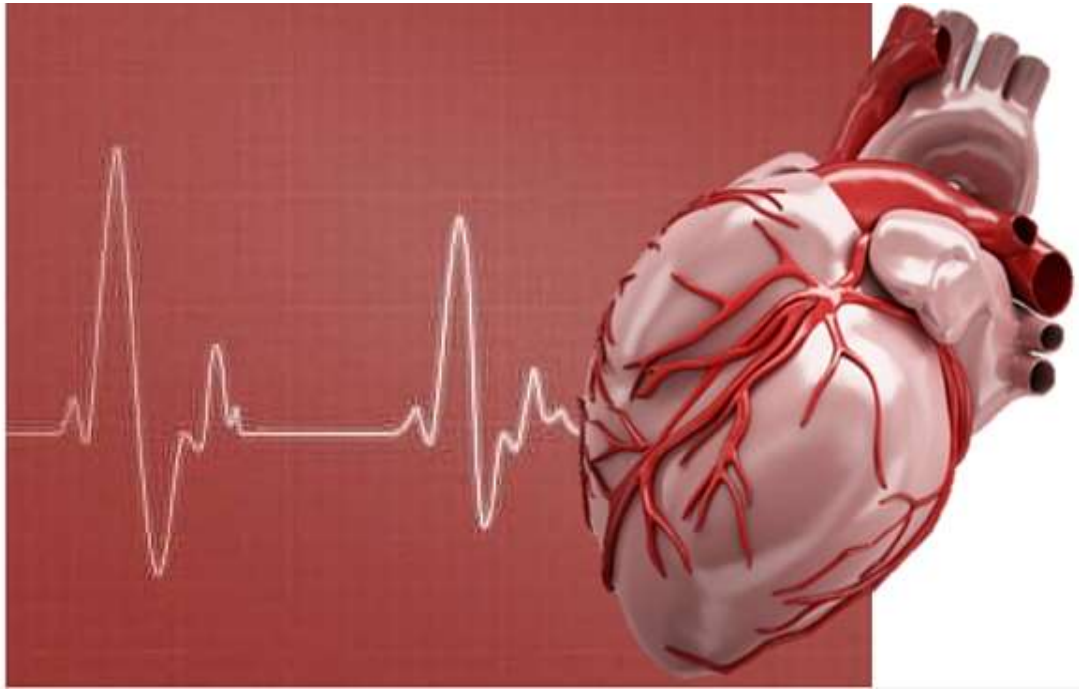
Supervised ML classification Capstone Project

Cardiovascular Risk Prediction

By Sarthak Arora

Contents

- Problem Statement
- Data description and Attributes
- Data Inspection
- EDA
- Level Encoding
- Feature Selection
- Handling Imbalanced Data
- Implementing Algorithms
- Challenges
- Conclusion



Problem Statement

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

- The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease(CHD).

- The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variable Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors

Data Description and Attributes

- **Demographic:**
- • Sex: male or female("M" or "F")
- • Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to
- whole numbers, the concept of age is continuous)

- **Behavioral**
- • is_smoking: whether or not the patient is a current smoker ("YES" or "NO")
- • Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be
- considered continuous as one can have any number of cigarettes, even half a cigarette.)

Data Description and Attributes

Medical(history)

• BP Meds: whether or not the patient was on blood pressure medication (Nominal)

• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)

• Diabetes: whether or not the patient had diabetes (Nominal)

Medical(current)

• Tot Chol: total cholesterol level (Continuous)

• Sys BP: systolic blood pressure (Continuous)

• Dia BP: diastolic blood pressure (Continuous)

• BMI: Body Mass Index (Continuous)

• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

• Glucose: glucose level (Continuous)

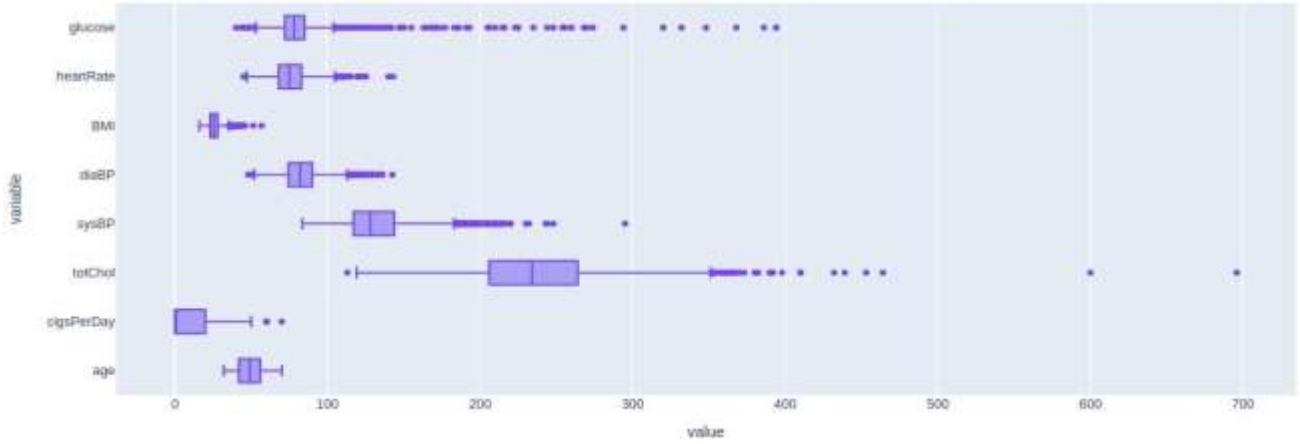
Predict variable (desired target)

Data Inspection:

- This Dataset has contains 3390 rows and 16 columns.
- Six categorical features i.e. sex , is_smoking ,BPMeds , prevalentStroke, prevalent Hyp, diabetes.
- This Dataset also contain missing values around 510 of seven features.

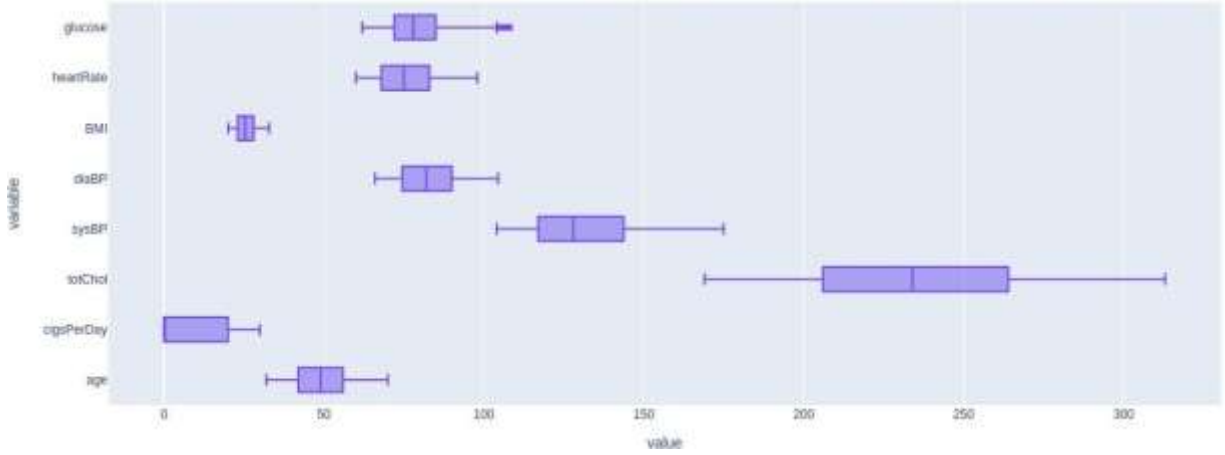
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3390 entries, 0 to 3389
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   id               3390 non-null   int64  
 1   age              3390 non-null   int64  
 2   education        3303 non-null   float64
 3   sex              3390 non-null   object 
 4   is_smoking       3390 non-null   object 
 5   cigsPerDay       3368 non-null   float64
 6   BPMeds           3346 non-null   float64
 7   prevalentStroke  3390 non-null   int64  
 8   prevalentHyp     3390 non-null   int64  
 9   diabetes         3390 non-null   int64  
 10  totChol          3352 non-null   float64
 11  sysBP            3390 non-null   float64
 12  diaBP            3390 non-null   float64
 13  BMI              3376 non-null   float64
 14  heartRate        3389 non-null   float64
 15  glucose          3086 non-null   float64
 16  TenYearCHD       3390 non-null   int64  
dtypes: float64(9), int64(6), object(2)
memory usage: 450.4+ KB
```

Analysis Of Outliers:



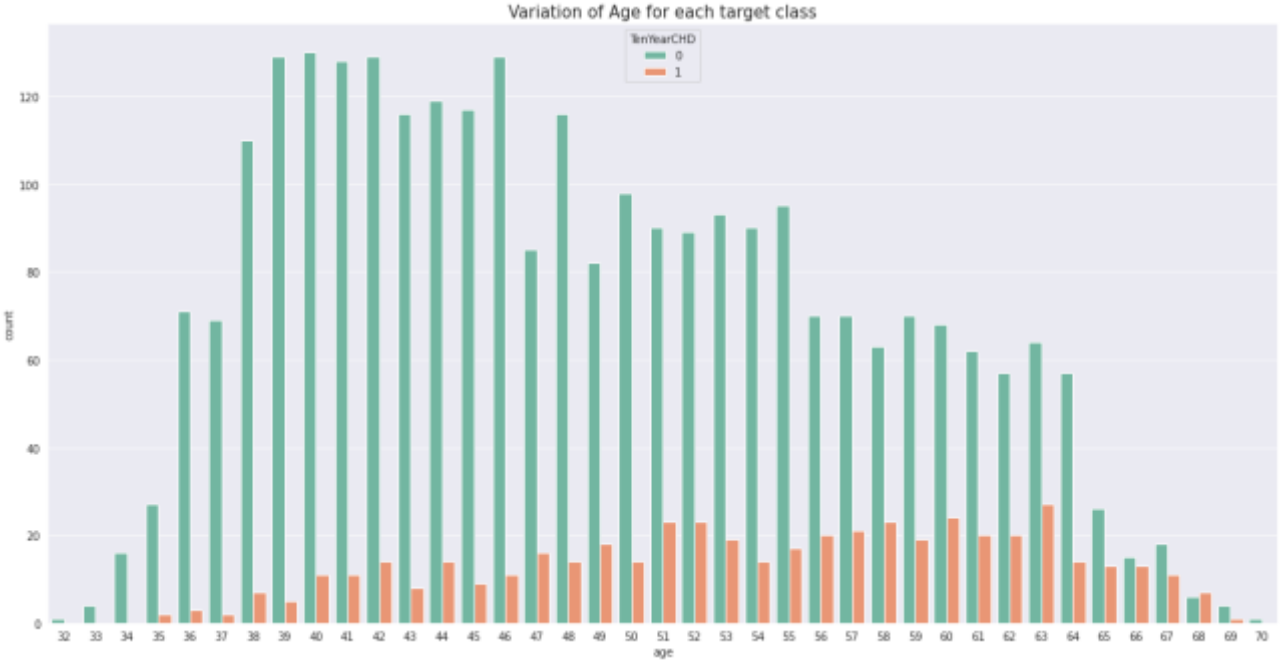
- Age has no outlier.

- Capping the outlier rows with Percentile.



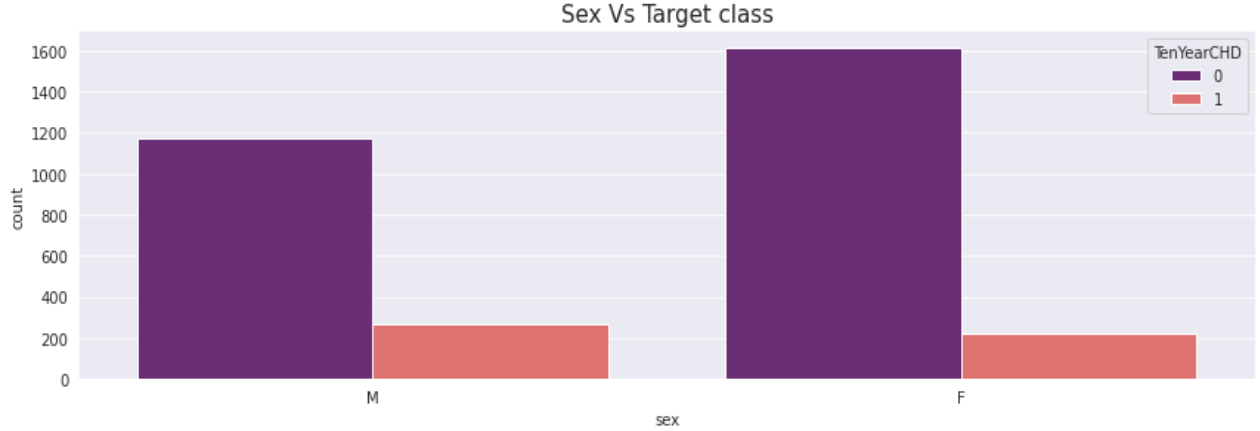
EDA

Analysis Of Age for each Target class:

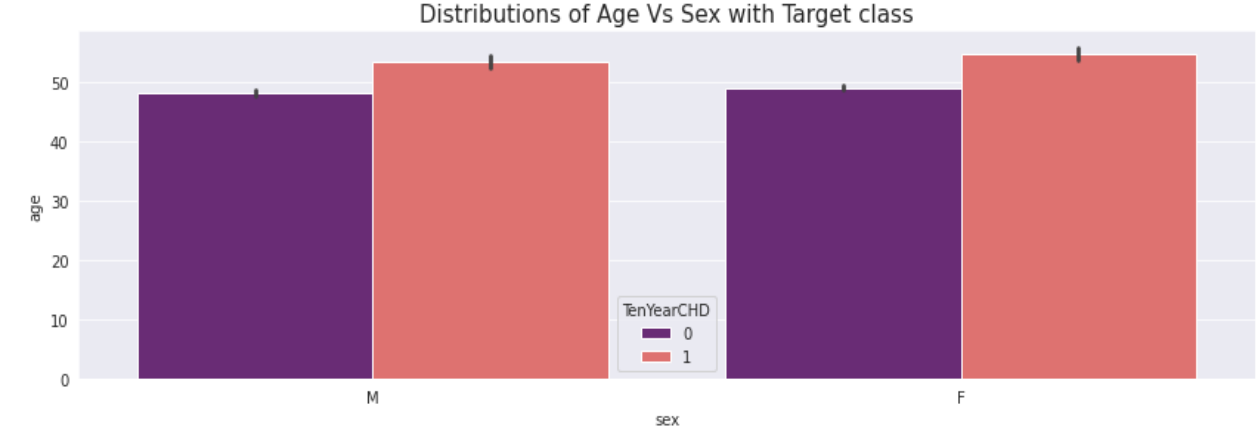


- Coronary heart disease(CHD) increases after age 51.
- Age group (34 < Age < 51) are at lower risk of cardiovascular disease.

Analysis Of Age vs Sex with Target class :



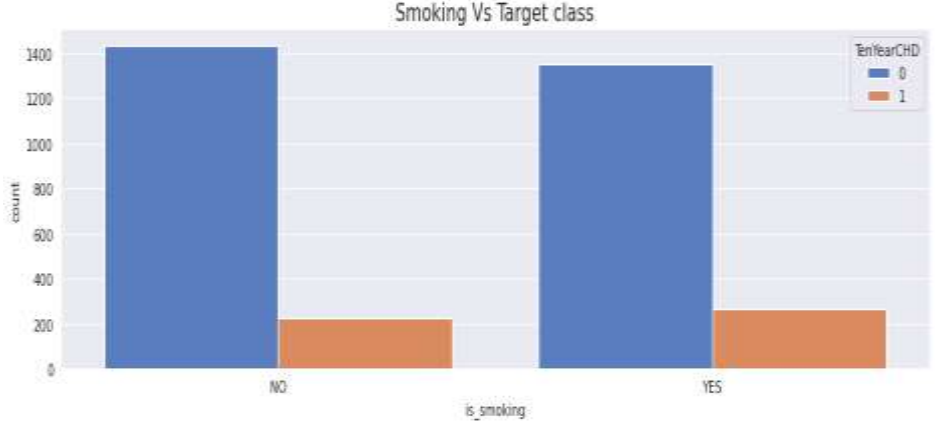
- We can see from the countplot that no. of male heart patient more than female.



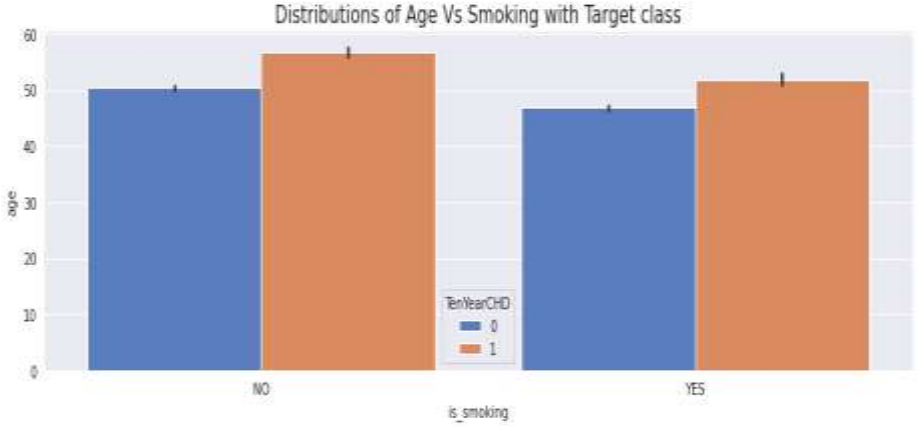
- We can see from the barplot that male get early CHD as compared to female.

Analysis Of Age vs Smoking with Target class :

- We can see from the countplot that no. of patient those who smoke more than as compared to those who don’t.

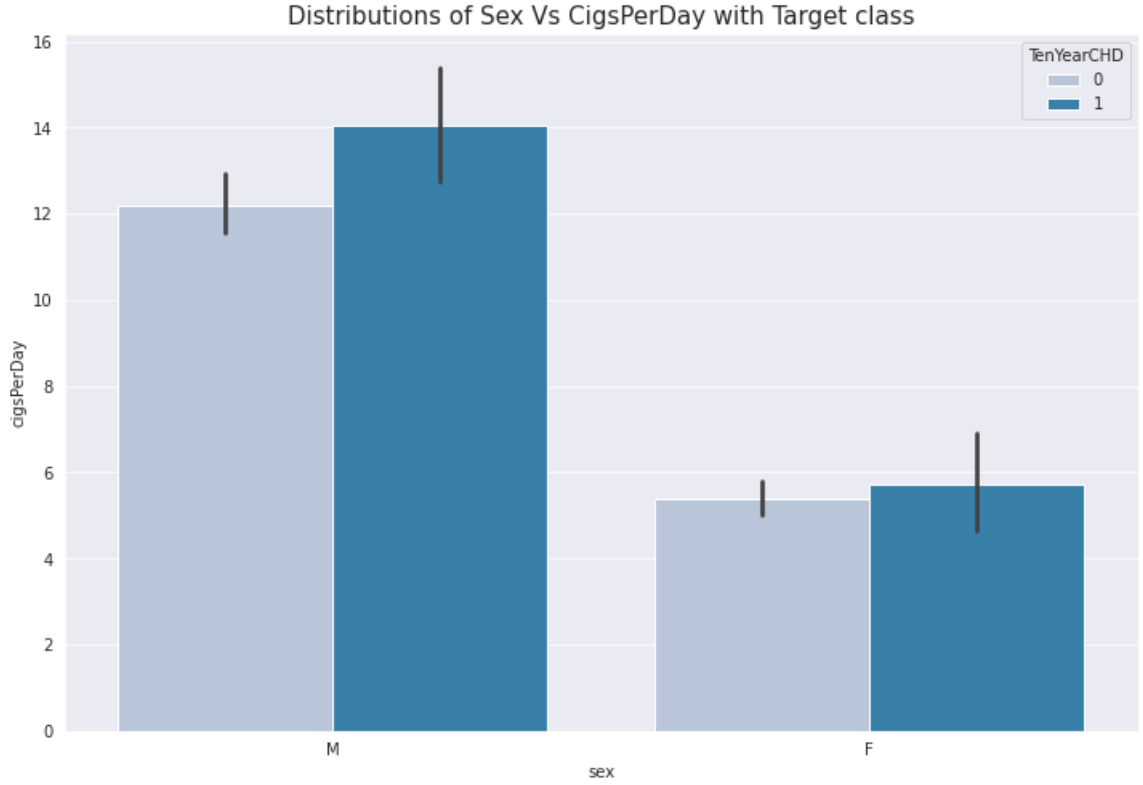


- We can see from the barplot that those who smoke get early heart disease as compared to those who don't.



Analysis Of Cigs per day vs Sex with Target class :

- We can see the barplot and say that no. of cigsperday taken by male is more than female.
- So, male heart patient is more as compared to female.
- In case of male CHD = 1 when he take cigsperday > 12.1 and in case of female CHD = 1 when she take cigsperday > 4.8.



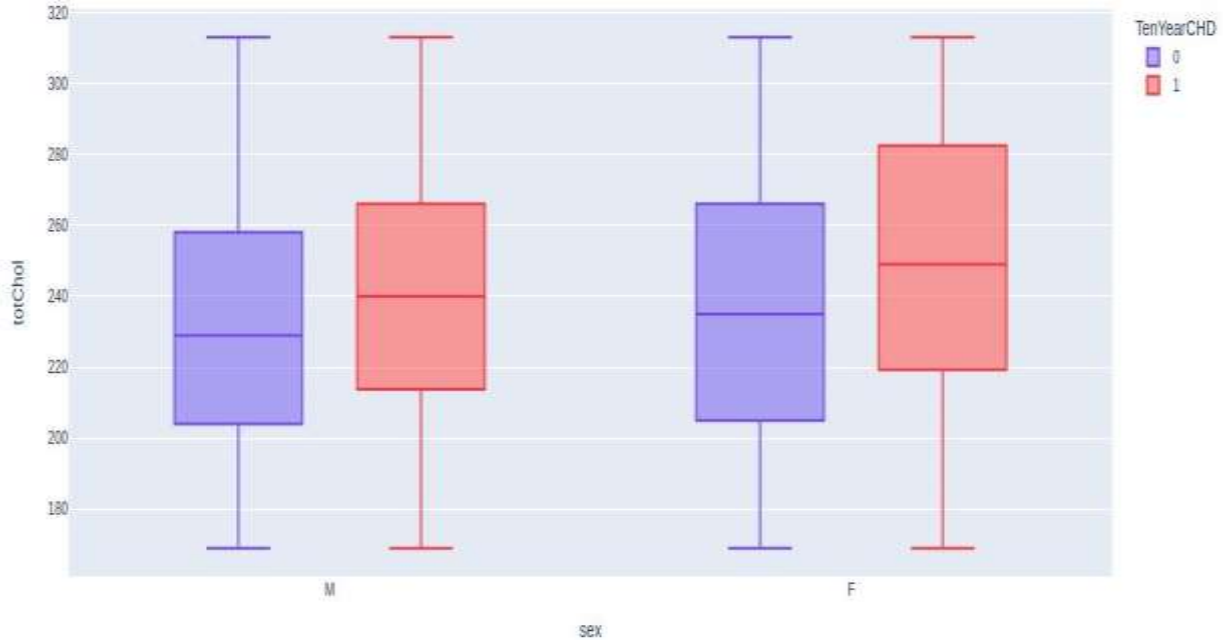
Analysis Of BMI vs Sex with Target class :

- We can see from the boxplot and say that female BMI is more than male BMI. that's leads to OVERWEIGHT.
- So,female CHD patient more than male CHD patient.
- If your BMI is:
  - below 18.5 – you're in the underweight range
  - between 18.5 and 24.9 – you're in the healthy weight range
  - between 25 and 29.9 – you're in the overweight range
  - between 30 and 39.9 – you're in the obese range

Analysis Of Cholesterol vs Sex withTarget class :

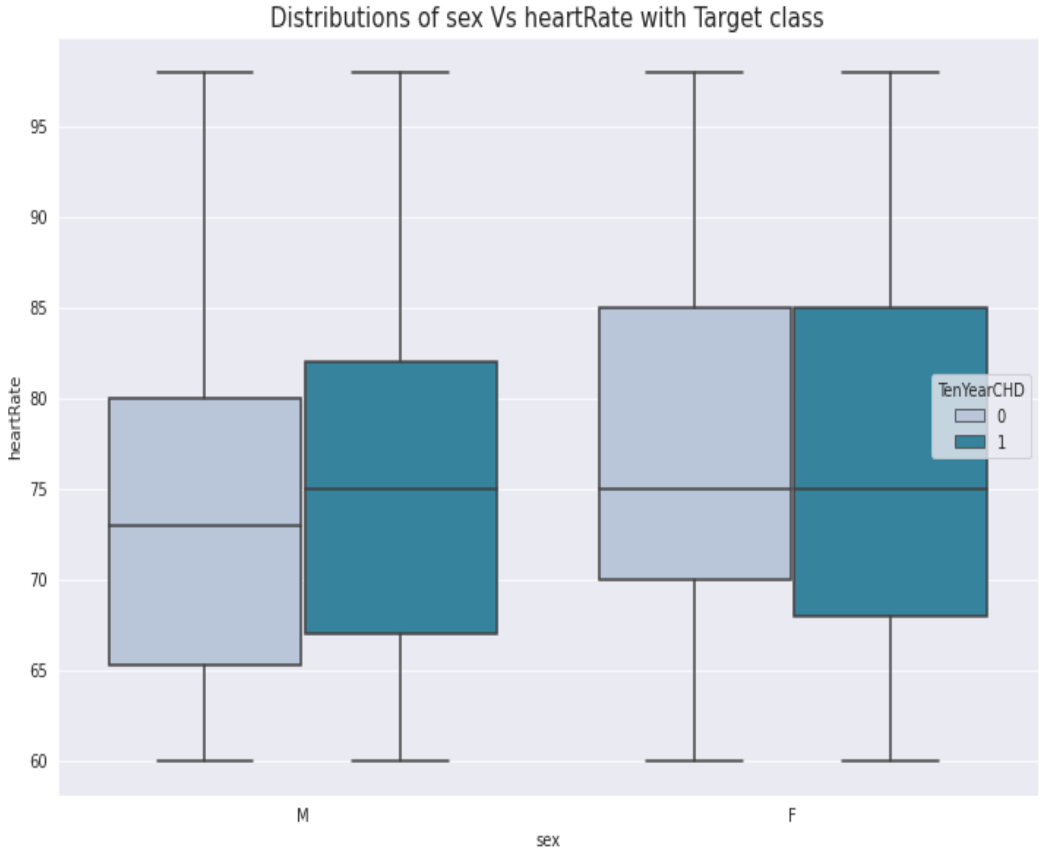
- We can see from the boxplot and say that female cholesterol is more than male cholesterol that's leads to OVERWEIGHT.

- So, In female heart disease is more due to cholesterol.



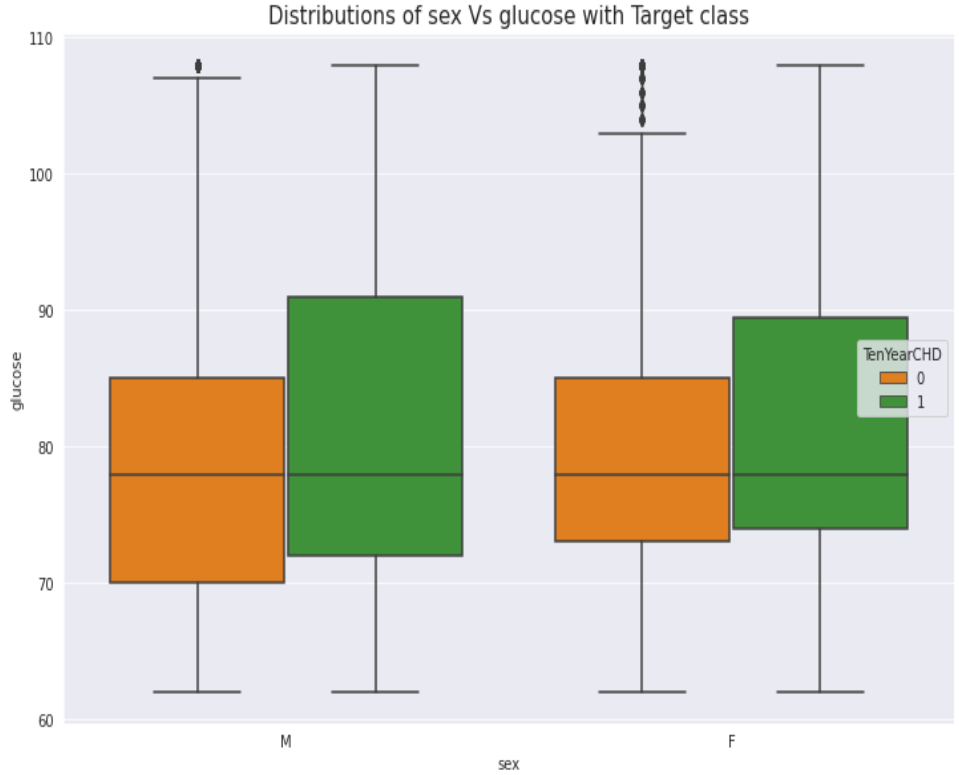
Analysis Of Heart Rate vs Sex with Target class :

- We can see from the box plot and say that for Female heart disease patients has more Heart Rate as compared to male heart disease patients.



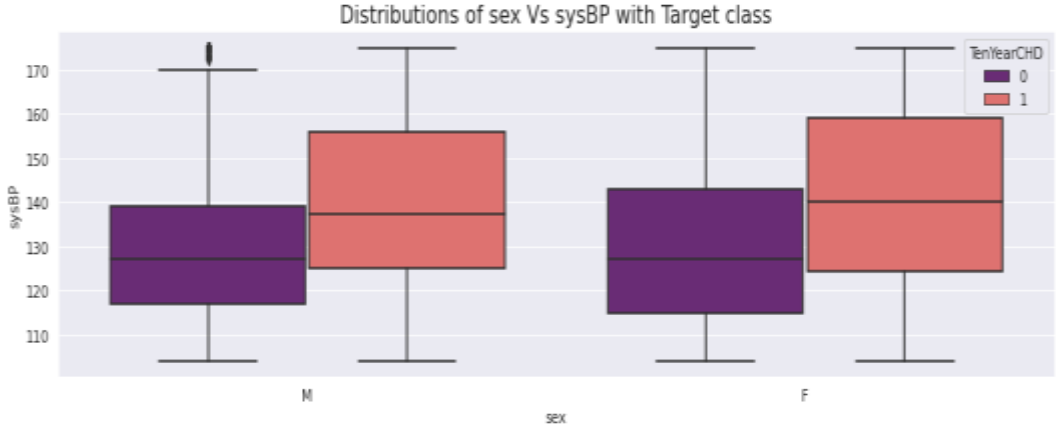
Analysis Of Glucose vs Sex with Target class :

- We can see the box plot that for male heart disease patients has more glucose level as compared to female heart disease patients.

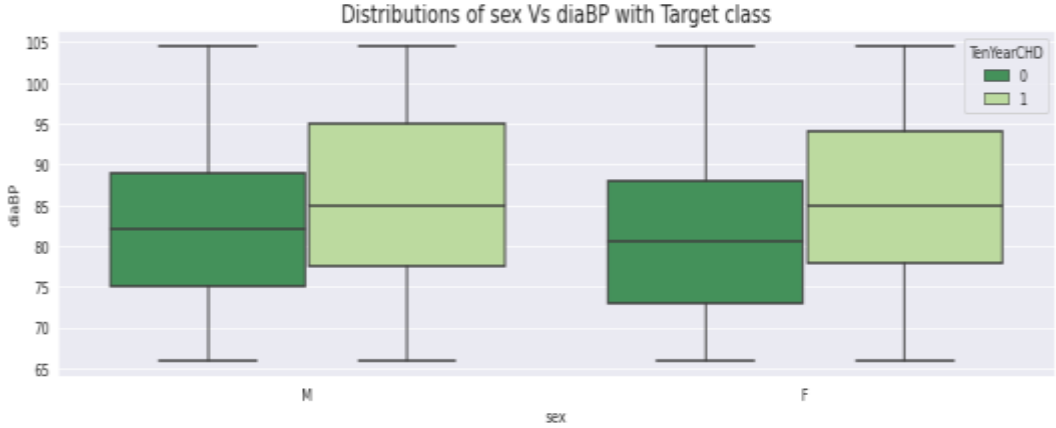


Analysis Of Systolic and Diastolic vs Sex with Target class :

- We can see the box plot and say that for female heart disease patients has more Systolic BP level as compared to male heart disease patients.
- Normal < 120 mmHg.

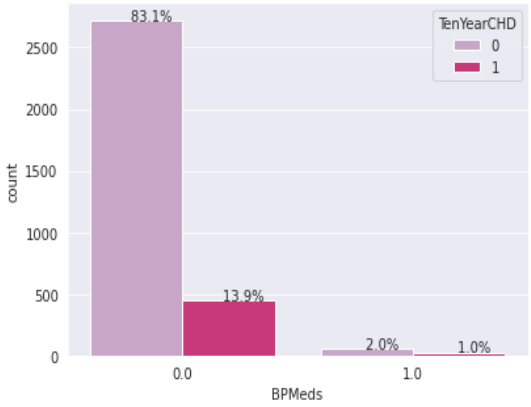


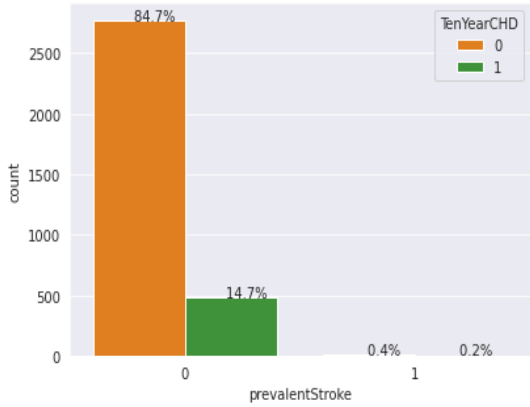
- We can see the box plot and say that for male heart disease patients has more Diastolic BP level as compared to female heart disease patients.
- Normal < 80 mmHg.

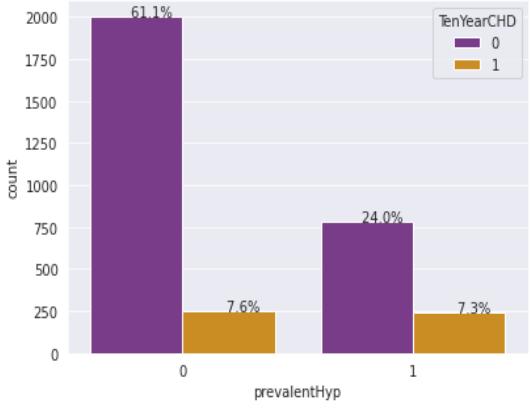


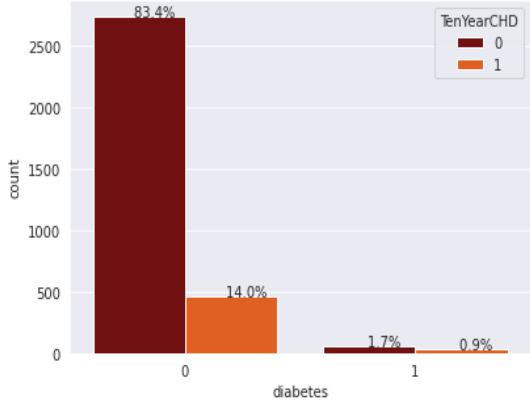
Analysis Of BP Meds | PrevalentStroke | PrevalentHyp | Diabetes vs Sex with Target class :

- BPMeds means whether or not the patient was on blood pressure medication i.e if the patients is take medication then it reduces the risk of heart disease, as compared to who won't take medication.



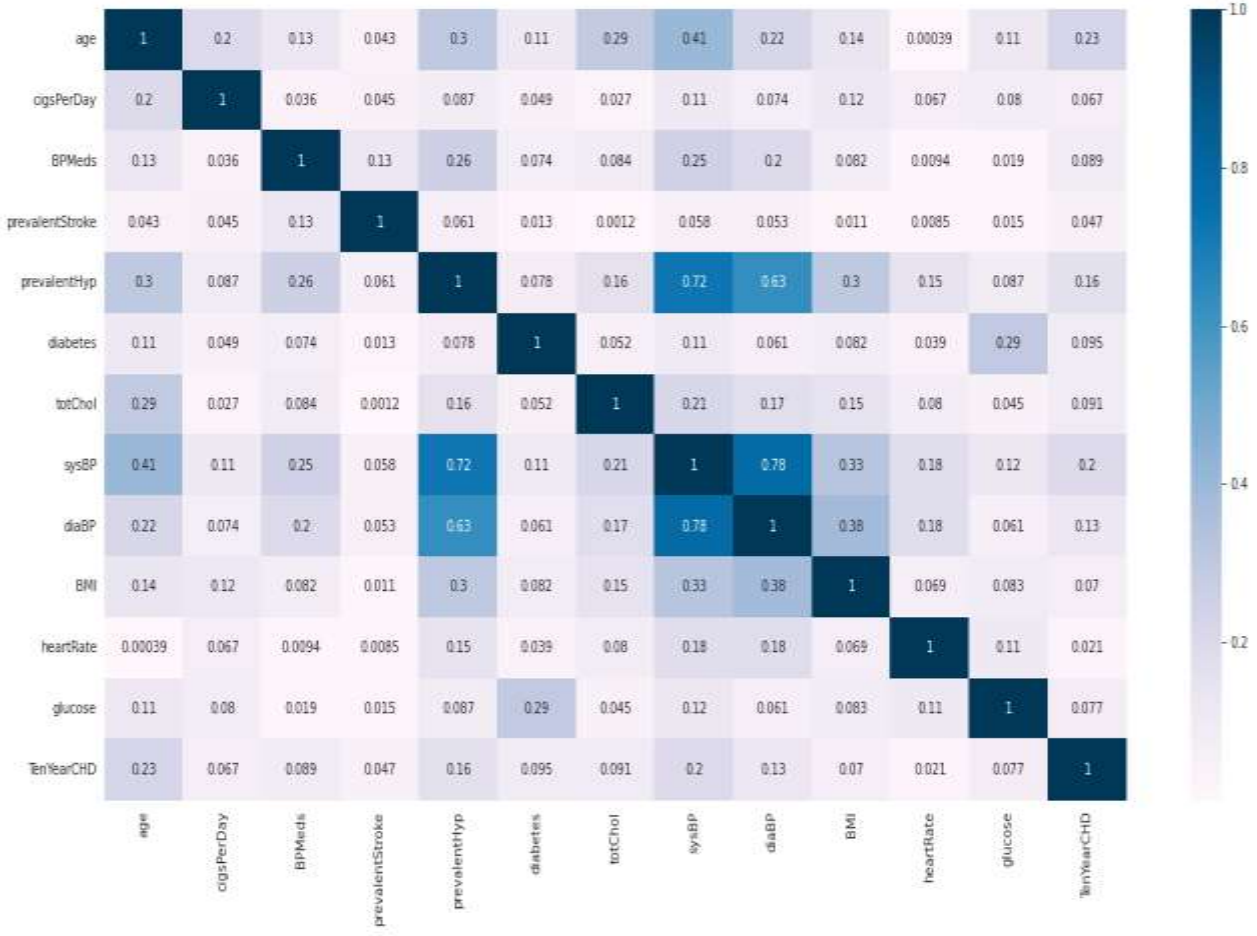






Correlation matrix:

- sysBP is moderately correlated with prevalenthyp, i.e. prevalent hypertension.
- diaBP and sysBP are somewhat moderately correlated.
- glucose level are also moderately correlated to whether patient is diabetic.



Label Encoding:

sex	is_smoking
M	NO
F	YES
M	YES
F	YES
F	NO

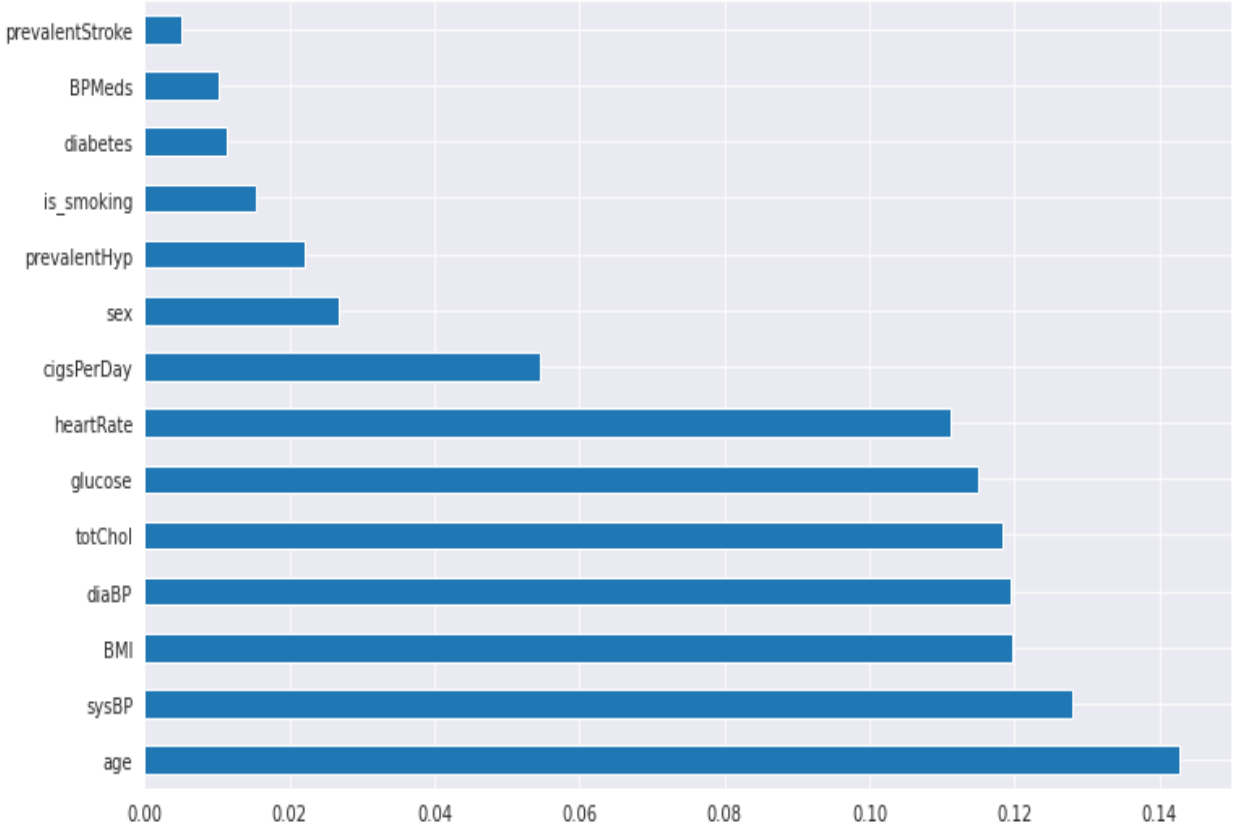
- We have two categorical columns i.e sex and is_smoking.

- After applying label encoding we converted into 0’s and 1’s.

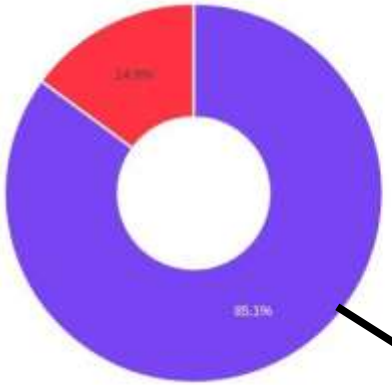
sex	is_smoking
1	0
0	1
1	1
0	1
0	0

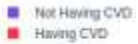
Feature Selection:

- For feature selection we used ExtraTreeClassifiers.
- We found that every feature is important.



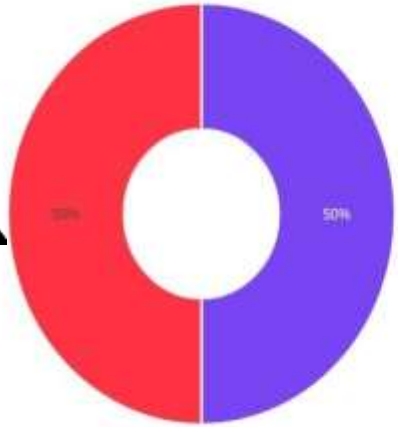
Handling Imbalanced Data:

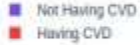





```
Class0: 2784
Class1: 488
Proportion: 5.7:1
```

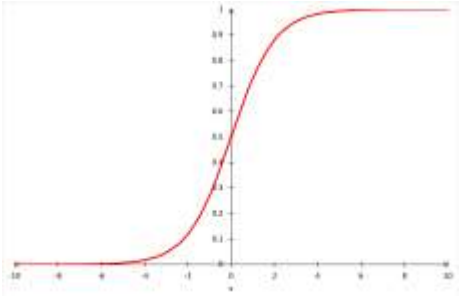
SMOTE



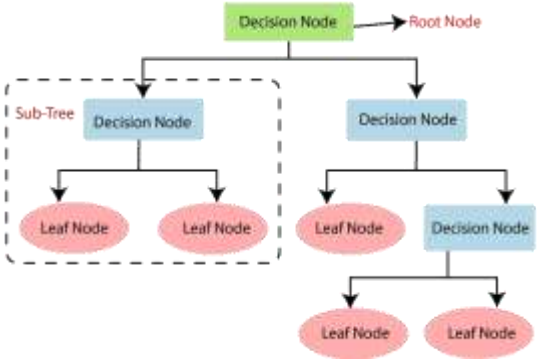


SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

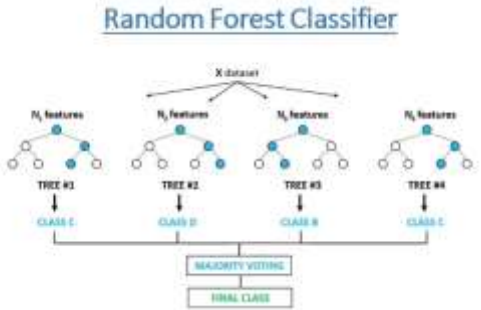
Model Building:



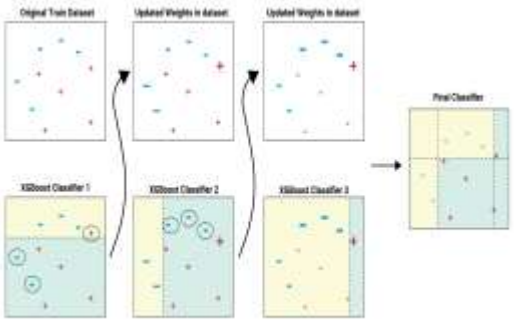
1] Logistic Regression



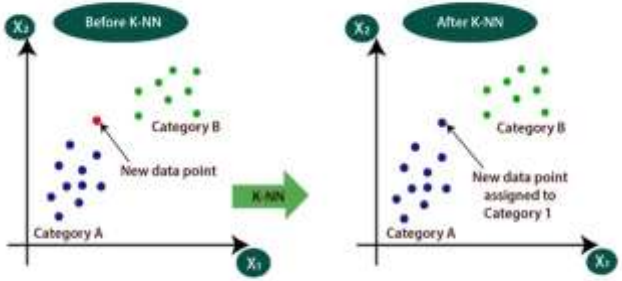
2] Decision Tree Classifier



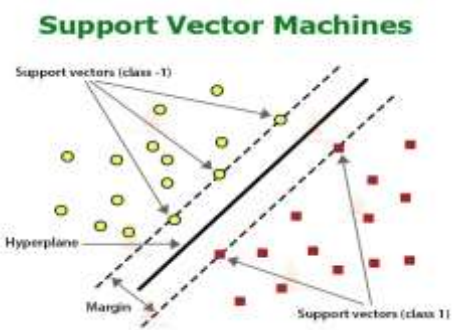
3] Random Forest Classifier



4] XGB Classifier



5] KNN Classifier

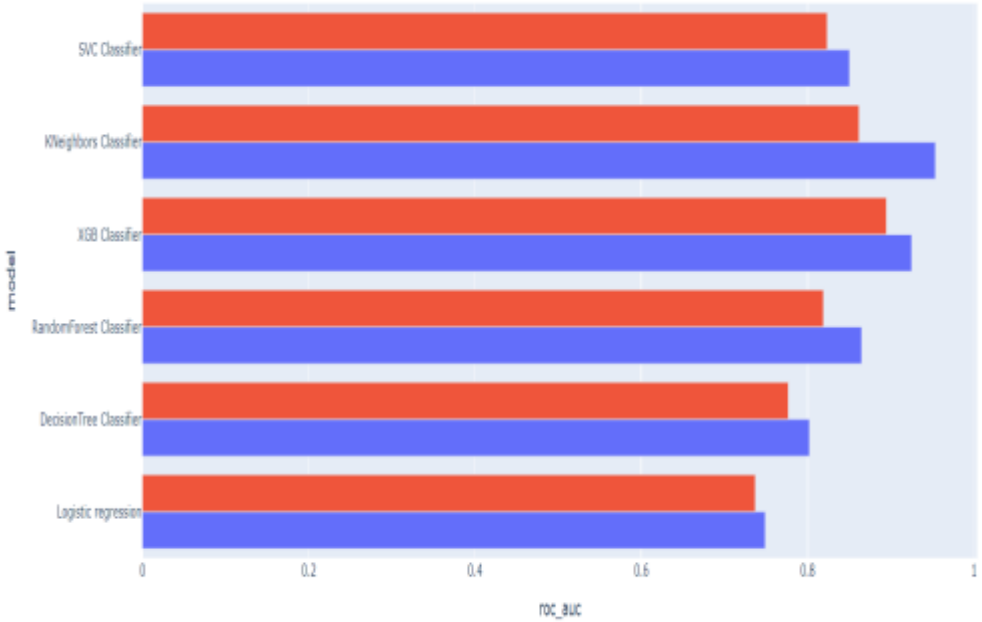


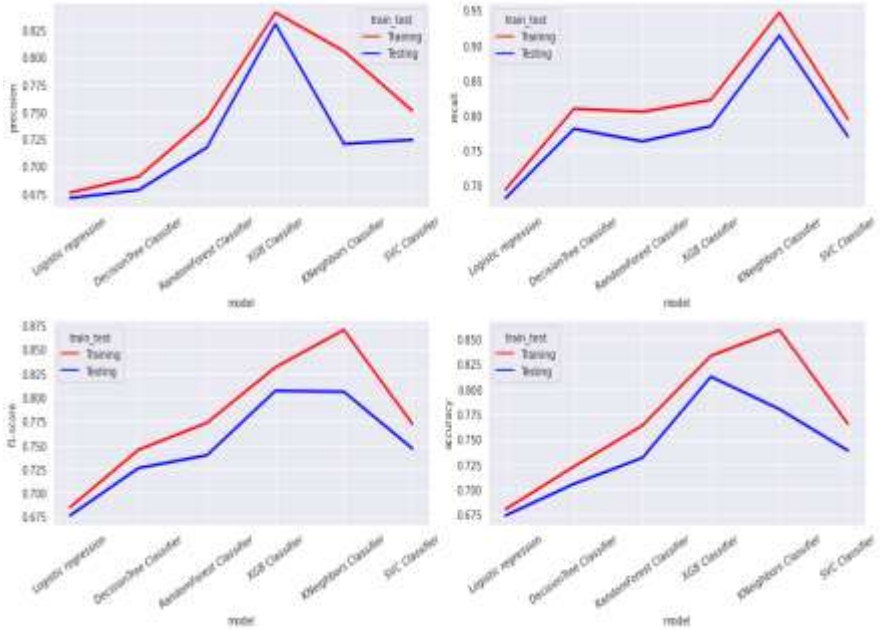
6] Support Vector Machines

Evaluating models:

		Model	Precision	Recall	F1-Score	Accuracy	ROC_AUC
Training set	**0**	Logistic regression	0.6757	0.6939	0.6847	0.6803	0.7488
	1	DecisionTree Classifier	0.6903	0.8092	0.7450	0.7229	0.8020
	2	RandomForest Classifier	0.7440	0.8048	0.7732	0.7638	0.8648
	3	XGB Classifier	0.8411	0.8218	0.8313	0.8332	0.9249
	4	KNeighbors Classifier	0.8056	0.9466	0.8704	0.8590	0.9535
	5	SVC Classifier	0.7510	0.7944	0.7721	0.7654	0.8502
Testing set	**0**	Logistic regression	0.6708	0.6817	0.6762	0.6741	0.7368
	1	DecisionTree Classifier	0.6781	0.7806	0.7258	0.7056	0.7766
	2	RandomForest Classifier	0.7174	0.7626	0.7393	0.7316	0.8188
	3	XGB Classifier	0.8305	0.7842	0.8067	0.8124	0.8946
	4	KNeighbors Classifier	0.7206	0.9137	0.8057	0.7801	0.8616
	5	SVC Classifier	0.7242	0.7698	0.7463	0.7388	0.8233

Comparing different ML Models:





- In the above Models Evaluation Table(Testing set) our auc-roc score is more than 0.80 except Logistic regression and Decision Tree.So we can say that our model predicted the classes in a good manner.
- XGB Classifier performed the best having the best Recall,Precision,F1-Score and Accuracy Score.

Challenges:

- Large Dataset to handle
- Needs to plot lot of Graphs to analyse
- Handling Null values
- Feature selection
- Optimising the model
- Carefully tuned Hyperparameters

Conclusion:

- In the given dataset we observe that Coronary heart disease increases from age 51 to 67 then decreases.
- We draw the countplot and observe that no. of male heart patients is more than female and also notice that male get early age heart diseases as compared to females.
- We observe no. of heart patients who smoke more than as compared to those who won't and also notice that those who smoke get early heart disease as compared to those who won't.
- We draw the barplot and observe that no. of cigsperday taken by male is more than female.So, male heart patients is more as compared to females.
- We draw the boxplot and observe that female BMI(The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2) is more than male BMI. that's leads to OVERWEIGHT and So,female CHD patients is more than male CHD patients.

- We draw the boxplot and observe that female Cholesterol is more than male Cholesterol. that's leads to OVERWEIGHT and So, in that case also female CHD patients is more than male CHD patients.

- We Observe that Female heart disease patients has more Heart Rate as compared to male heart disease patients.

- We also observe that male heart disease patients has more glucose level as compared to female heart disease patients.

- In the Models Evaluation Table(Testing set) our auc-roc score is more 0.80 except Logistic regression and Decision Tree.So we can say that our model predicted the classes in a good manner.

- XGBClassifier are performing well which has the best Recall,Precision,F1-Score and Accuracy Score.


thank you